

大数据审计实验系统

V3.0

教学实施方案

课程教学

项目一：初始人工智能

任务一 人工智能的定义

一、理论知识实训

1. 人工智能的主要研究领域包括以下哪一项？

- A: 机器视觉
- B: 机器听觉
- C: 自然语言理解
- D: 以上全部

标准答案：D

2. 以下哪种技术属于人工智能的基本技术？

- A: 感知能力
- B: 动作能力
- C: 推理能力
- D: 以上全部

标准答案：D

3. 关于图灵测试，下列说法正确的是？

- A: 仅反映了结果的比较
- B: 涉及思维过程
- C: 必须采用机器视觉进行测试
- D: 必须采用机器翻译进行测试

标准答案：A

4. 人工智能的新技术不包括以下哪一项？

- A: 神经计算
- B: 多 Agent 技术
- C: 机械自动化

D: 群体智能

标准答案: C

5. 下列哪一项不属于机器学习的类型?

A: 符号学习

B: 连接学习

C: 自然计算

D: 深度学习

标准答案: C

6. 人工智能的核心目标是使机器具有类似于人类的感知、____和 ____ 能力。

标准答案: 记忆, 思维

7. 知识阈值理论认为, 一个系统所具有的 _____ 越多, 其智能就会越高。

标准答案: 可运用知识

多选题

1. 关于人工智能的定义和研究目标, 下列哪些说法正确?

A: 人工智能是通过人工方法在机器上实现类似人类的智能

B: 图灵测试能够全面衡量机器的智能水平

C: 人工智能的研究目标包括感知、思维和自适应能力

D: 自然智能和人工智能在本质上是相同的

标准答案: AC

2. 以下哪些观点体现了对智能本质的不同理解?

A: 知识阈值理论强调智能来源于知识数量和可用性

B: 思维理论认为智能的核心是知识的存储和传递

C: 进化理论认为智能可通过适应复杂环境逐步进化

D: 思维理论和进化理论在层次结构上的解释完全一致

标准答案: AC

任务二 人工智能的发展

一、理论知识实训

1. 以下哪位科学家被称为“控制论创始人”?

- A: 图灵
- B: 维纳
- C: 莱布尼茨
- D: 亚里士多德

标准答案: B

2. 1956 年达特茅斯会议标志着人工智能的正式诞生, 这一会议的核心目标是?

- A: 研究计算机硬件的发展
- B: 探讨如何使机器模仿人类智能
- C: 制定人工智能的伦理规范
- D: 研发专家系统

标准答案: B

3. 下列哪一项是人工智能形成期 (1956-1970 年) 的重要研究成果?

- A: ENIAC 计算机
- B: 逻辑理论机
- C: DENDRAL 系统
- D: 深度学习框架

标准答案: B

4. 专家系统的出现是人工智能发展的重要突破, 其首次成功应用是在以下哪个领域?

- A: 地质勘探
- B: 机器翻译
- C: 医学诊断
- D: 棋类博弈

标准答案: C

5. 以下哪项问题导致人工智能在知识应用期 (1971-1980 年) 遭遇挫折?

- A: 神经网络模型失败
- B: 自然语言理解取得突破
- C: 计算机性能不足

D: 专家系统缺乏常识性知识

标准答案: D

6. 下列哪一技术被认为是人工智能“知识爆炸”问题的核心原因之一?

A: 模式识别

B: 归纳学习

C: 知识表示的复杂性

D: 数据存储效率不足

标准答案: C

7. 以下哪项技术在人工智能的形成期得到发展, 并被认为是专家系统的基础?

A: 深度学习

B: 规则推理

C: 神经网络

D: 贝叶斯网络

标准答案: B

8. 下列关于“弱人工智能”和“强人工智能”的描述正确的是?

A: 强人工智能旨在解决特定领域的实际问题

B: 弱人工智能具备人类的自我意识

C: 强人工智能的目标是完全模拟人类智能

D: 弱人工智能是基于概率的推理系统

标准答案: C

9. 深蓝 (Deep Blue) 击败世界象棋冠军的成功标志着哪一人工智能研究领域的重要突破?

A: 自然语言处理

B: 专家系统

C: 机器学习

D: 博弈问题求解

标准答案: D

10. 下列哪项是深度学习在人工智能发展中的主要优势?

- A: 对大规模非结构化数据的处理能力
- B: 推理链条清晰易解释
- C: 适用于稀疏数据的小样本学习
- D: 自动化知识表示

标准答案: A

11. 在人工神经网络的发展史上, 以下哪项被认为是反向传播算法的理论基础?

- A: 梯度下降法
- B: 图灵机模型
- C: 马尔可夫链
- D: 贝叶斯定理

标准答案: A

12. 人工智能领域的自然语言处理研究通常分为两个主要任务: 语言理解和_____。

标准答案 : 语言生成

13. 深度学习的关键组件包括输入层、隐藏层和输出层, 其中隐藏层通过_____激活函数实现非线性映射。

标准答案: 非线性

任务三 人工智能的分类与应用

一、理论知识实训

1. 以下哪一项最能体现机器感知技术的核心目标?

- A: 模拟人类的视觉、听觉和触觉能力
- B: 实现深度学习算法的优化
- C: 使用逻辑推理处理复杂问题
- D: 通过规则学习生成知识图谱

标准答案: A

2. 以下哪一项最能体现机器感知技术的核心目标?

- A: 模拟人类的视觉、听觉和触觉能力
- B: 实现深度学习算法的优化

C: 使用逻辑推理处理复杂问题

D: 通过规则学习生成知识图谱

标准答案: A

3. 智能机器人在工业领域的主要应用优势不包括以下哪一项?

A: 降低人工成本

B: 提高生产效率

C: 代替所有人类决策

D: 增强设备维护的精准性

标准答案: C

4. 在金融领域, 人工智能技术应用于量化交易的核心在于以下哪种能力?

A: 数据挖掘

B: 数据加密

C: 生成模型

D: 数据合成

标准答案: A

5. 智能交通管理的主要目标是优化以下哪一项?

A: 自动驾驶汽车的设计

B: 交通流量和拥堵状况

C: 汽车智能语音助手

D: 交通事故保险模型

标准答案: B

6. 自然语言处理中的关键任务包括以下哪一项?

A: 机器感知

B: 模式匹配

C: 文本生成

D: 动作控制

标准答案: C

7. 以下哪项是当前人工智能研究中重点关注的未来发展趋势?

A: 模式匹配优化

- B: 可解释性增强
- C: 数据存储硬件升级
- D: 模糊逻辑简化

标准答案: B

8. 深度学习通过_____层提取复杂特征, 从而实现精准预测。

标准答案: 隐藏

9. 联邦学习的核心目标是实现分布式模型训练并保护_____。

标准答案: 数据隐私

多选题:

1. 关于机器感知技术, 下列说法正确的是?

- A: 机器视觉是通过摄像头和算法让计算机“看见”并理解图像。
- B: 机器听觉的关键技术包括语音识别和声音分类。
- C: 机器感知是智能系统的核心决策部分。
- D: 自然语言处理属于机器感知的研究范围。

标准答案: AB

2. 人工智能在医疗领域的应用主要包括以下哪些?

- A: 使用机器学习进行疾病预测和诊断。
- B: 利用深度学习设计新药物的分子结构。
- C: 基于专家系统的智能手术规划。
- D: 开发增强现实技术用于医疗培训。

标准答案: ABCD

3. 以下关于人工智能的基础学科的说法正确的是?

- A: 数学为人工智能提供了理论支撑, 包括概率论和模糊数学。
- B: 思维科学的研究方向包括认知心理学和逻辑学。
- C: 生物学模型是智能行为模拟的重要理论来源。
- D: 自动化技术仅在工业领域应用, 与人工智能无关。

标准答案: ABC

4. 下列关于强化学习的说法正确的是?

- A: 强化学习通过与环境交互学习策略。

B: 强化学习不依赖奖励机制。

C: 强化学习广泛应用于游戏和机器人控制。

D: 强化学习的核心数学模型是马尔可夫决策过程 (MDP)。

标准答案: ACD

5. 关于人工智能未来发展趋势, 下列哪些说法正确?

A: 迁移学习可以将知识从一个任务迁移到另一个不同任务中。

B: 联邦学习能够在保护数据隐私的情况下实现分布式训练。

C: 可解释性 AI 提高了深度学习模型的透明度, 但降低了模型性能。

D: 深度学习已完全解决自然语言处理中的所有挑战。

标准答案: AB

项目二：人工智能支撑技术

任务一 支撑技术概述

一、理论知识实训

1. 以下哪项不属于监督学习算法？

- A: 线性回归
- B: 逻辑回归
- C: 支持向量机
- D: K-means

标准答案：D

2. 在深度学习中，专门用于处理序列数据的神经网络是：

- A: CNN
- B: FNN
- C: RNN
- D: Transformer

标准答案：C

3. 以下哪种技术不属于特征工程方法？

- A: 特征选择
- B: 特征提取
- C: 特征构造
- D: 独热编码

标准答案：D

4. 强化学习的典型代表应用是：

- A: 图像分类
- B: AlphaGo
- C: 客户分群
- D: 房价预测

标准答案：B

5. 深度学习中用于防止过拟合的技术不包括：

A: Dropout

B: L2 正则化

C: 早停法

D: 批归一化

标准答案: D

6. 以下哪项是大数据处理框架?

A: Hadoop

B: PyTorch

C: TensorFlow

D: Keras

标准答案: A

7. 在模型优化中, 超参数调优方法不包括:

A: 网格搜索

B: 随机搜索

C: 特征缩放

D: 贝叶斯优化

标准答案: C

8. 主成分分析(PCA)的主要作用是:

A: 分类预测

B: 数据降维

C: 异常检测

D: 模型加速

标准答案: B

9. 以下哪种部署方式适合低延迟要求的场景?

A: 云端部署

B: 数据湖架构

C: 边缘计算

D: 批处理

标准答案: C

10. 在数据预处理中，处理缺失值的错误方法是：

- A: 直接删除
- B: 均值插补
- C: 预测模型填充
- D: 用固定值-1 填充

标准答案：D

任务二 Python 与人工智能支撑技术的关系

一、理论知识实训

1. Python 在 AI 开发中的核心优势不包括以下哪项？

- A: 动态类型系统
- B: 丰富的标准库
- C: C 扩展接口
- D: 静态类型检查

标准答案：D

2. 以下哪个库专门用于高效的多维数组运算？

- A: NumPy
- B: SciPy
- C: Pandas
- D: Matplotlib

标准答案：A

3. 在 Python AI 开发流程中，数据采集与清洗通常使用的工具是：

- A: Pandas-profiling
- B: BeautifulSoup
- C: Flask
- D: Numba

标准答案：B

4. 以下哪项是 Jupyter 生态系统的功能？

- A: 模型压缩
- B: 数据插补

C: Notebook 交互开发

D: 异常值检测

标准答案: C

5. Python 中用于加速数值计算的库是:

A: Scrapy

B: OpenCV

C: Cython

D: Numba

标准答案: D

6 以下哪项不属于 Python 科学计算生态的核心库?

A: NumPy

B: SciPy

C: Pandas

D: Scrapy

标准答案: D

7. 在 Python 中, 实现跨语言调用 C/C++库的接口是:

A: JPyype

B: Cython

C: ctypes

D: Numba

标准答案: C

8. 以下哪项是 Python AI 开发中性能优化的常见方法?

A: 向量化运算

B: 增加循环次数

C: 动态类型检查

D: 减少库调用

标准答案: A

9. Python 中用于快速部署 AI 模型的轻量级框架是:

A: Scikit-learn

B: Flask

C: OpenCV

D: Pandas

标准答案: B

10. 机器学习理论的基础学科不包括以下哪项?

A: 概率论

B: 统计学

C: 凸分析

D: 量子力学

标准答案: D

二、智能制造传感器数据分析

(一)、任务背景

在现代智能制造体系中，设备状态监测是保障生产安全和提高生产效率的关键环节。工业设备通过部署的温度传感器和振动传感器，能够实时监测设备的运行状态，及时发现潜在的故障风险。

传感器数据分析在智能制造中发挥着重要作用，通过对设备温度和振动数据的持续监测和分析，可以实现设备的预测性维护，避免突发故障造成的停机损失。同时，数据驱动的决策能够帮助企业优化生产流程，提高设备利用率。

随着工业 4.0 的深入发展，传统的人工巡检方式已难以满足现代化制造业对精度、速度和成本控制的要求。基于数据分析的智能监测系统能够在生产线上实现设备的自动化监测和异常识别，有效提升运维效率和设备可靠性。

本任务模拟了一个典型的工业设备监测场景，通过传感器采集的温度和振动数据，运用数据预处理、异常检测和可视化分析等技术手段，构建完整的设备状态监测解决方案。

(二)、任务一：数据质量检查

1. 对原始传感器数据 `raw_sensor_data.csv` 进行全面的质量评估和统计分析
2. 检查数据的完整性，识别缺失值和异常值的分布情况
3. 分析不同设备的数据分布特征和时间序列的连续性

4. 生成详细的数据质量报告，为后续预处理提供依据

在程序编码补全中根据上述信息补全对应代码，其中数据以数据中心文件列表中的实际数据为准。

填写完毕后，可点击需求实现查看完整代码，确认无误后点击执行并显示结果，待代码运行完毕后即可查看结果。

【参考答案】：

(1) isnull()

(2) to_datetime

(3) pd.read_csv

```
输出：
成功加载 raw_sensor_data.csv
=== 数据质量报告 ===

[1] 总记录数: 100
[2] 时间范围: 2025-06-17 08:00:00 至 2025-06-18 00:30:00

[3] 设备分布:
device_id
M1001  50
M1002  50
Name: count, dtype: int64

[4] 缺失值统计:
timestamp  0
device_id  0
temperature 5
vibration  5
dtype: int64

[5] 数值统计:
      temperature vibration
count  95.000000  95.000000
mean   49.095789  33.975789
std     8.772817   5.206680
min    31.900000  20.300000
25%    43.000000  30.950000
50%    48.100000  33.500000
```

图 2-1 输出结果示意图

(三) 任务二：数据预处理

1. 实现完整的数据清洗和预处理流程
2. 处理数据中的缺失值，采用合适的填充策略保证数据完整性
3. 识别并过滤异常的温度读数，确保数据的可靠性

在程序编码补全中根据上述信息补全对应代码，其中数据以数据中心文件列表中的实际数据为准。

填写完毕后，可点击需求实现查看完整代码，确认无误后点击执行并显示结果，待代码运行完毕后即可查看结果。

【参考答案】：

(1) ffill

(2) transform

(3) transform

输出：

预处理完成: processed_sensor_data.csv

运行结束

图 2-2

	A	B	C	D	E	F
1	timestamp	device_id	temperature	vibration	temp_ma	
2	2025/6/17 8:00	M1001	47.5	31.3	47.5	
3	2025/6/17 8:10	M1002	49.3	37.8	49.3	
4	2025/6/17 8:20	M1001	43.8	25.9	45.65	
5	2025/6/17 8:30	M1002	48.8	32.7	49.05	
6	2025/6/17 8:40	M1001	82.4	32.6	57.9	
7	2025/6/17 8:50	M1002	53.8	34.3	50.63333	
8	2025/6/17 9:00	M1001	42.1	31.05	53.95	
9	2025/6/17 9:10	M1002	47.4	35.3	49.825	
10	2025/6/17 9:20	M1001	31.9	43.6	49.54	
11	2025/6/17 9:30	M1002	54.8	37.9	50.82	
12	2025/6/17 9:40	M1001	42.9	48.6	48.62	
13	2025/6/17 9:50	M1002	46.3	37.1	50.22	
14	2025/6/17 10:00	M1001	41.9	31.05	48.24	
15	2025/6/17 10:10	M1002	53	37.3	51.06	
16	2025/6/17 10:20	M1001	45.6	27.2	40.88	
17	2025/6/17 10:30	M1002	47.4	33.5	49.78	
18	2025/6/17 10:40	M1001	42	31.9	40.86	
19	2025/6/17 10:50	M1002	54.7	40.3	51.24	
20	2025/6/17 11:00	M1001	39.9	33.2	42.46	
21	2025/6/17 11:10	M1002	77.8	35.2	55.84	

图 2-3

(四)、任务三：异常检查分析

1. 针对单个设备实施多种异常检测算法
2. 基于统计学方法检测温度数据中的异常波动
3. 运用四分位距和 Z-score 等方法识别振动数据的异常模式
4. 生成每个设备的异常检测报告，量化异常事件的发生频率

在程序编码补全中根据上述信息补全对应代码，其中数据以数据中心文件列表中的实际数据为准。

填写完毕后，可点击需求实现查看完整代码，确认无误后点击执行并显示结果，待代码运行完毕后即可查看结果。

【参考答案】：

(1) mean



输出：
异常检测完成，结果已保存

运行结束

其他：

- (1) 文件: processed_sensor_data.csv
- (2) 文件: anomaly_report_M1001.csv
- (3) 文件: anomaly_report_M1002.csv
- (4) 文件: sensor_data_with_anomalies.csv

图 2- 4

	A	B	C	D	E	F	G	H
1	timestamp	device_id	temperature	vibration	temp_ma	temp_anomaly	vib_iqr_anomaly	is_anomaly
2	2025/6/17 8:00	M1001	47.5	31.3	47.5	FALSE	FALSE	FALSE
3	2025/6/17 8:20	M1001	43.8	25.9	45.65	FALSE	FALSE	FALSE
4	2025/6/17 8:40	M1001	82.4	32.6	57.9	TRUE	FALSE	TRUE
5	2025/6/17 9:00	M1001	42.1	31.05	53.95	FALSE	FALSE	FALSE
6	2025/6/17 9:20	M1001	31.9	43.6	49.54	FALSE	TRUE	TRUE
7	2025/6/17 9:40	M1001	42.9	48.6	48.62	FALSE	TRUE	TRUE
8	2025/6/17 10:00	M1001	41.9	31.05	48.24	FALSE	FALSE	FALSE
9	2025/6/17 10:20	M1001	45.6	27.2	40.88	FALSE	FALSE	FALSE
10	2025/6/17 10:40	M1001	42	31.9	40.86	FALSE	FALSE	FALSE
11	2025/6/17 11:00	M1001	39.9	33.2	42.46	FALSE	FALSE	FALSE
12	2025/6/17 11:20	M1001	49.1	27.1	43.7	FALSE	FALSE	FALSE
13	2025/6/17 11:40	M1001	38.4	30.9	43	FALSE	FALSE	FALSE
14	2025/6/17 12:00	M1001	48.7	34.3	43.62	FALSE	FALSE	FALSE
15	2025/6/17 12:20	M1001	48.7	30.7	44.96	FALSE	FALSE	FALSE
16	2025/6/17 12:40	M1001	42.7	45.2	45.52	FALSE	TRUE	TRUE

图 2- 5 anomaly_report_M001.csv

	A	B	C	D	E	F	G	H
1	timestamp	device_id	temperature	vibration	temp_ma	temp_anomaly	vib_iqr_anomaly	is_anomaly
2	2025/6/17 8:10	M1002	49.3	37.8	49.3	FALSE	FALSE	FALSE
3	2025/6/17 8:30	M1002	48.8	32.7	49.05	FALSE	FALSE	FALSE
4	2025/6/17 8:50	M1002	53.8	34.3	50.63333	FALSE	FALSE	FALSE
5	2025/6/17 9:10	M1002	47.4	35.3	49.825	FALSE	FALSE	FALSE
6	2025/6/17 9:30	M1002	54.8	37.9	50.82	FALSE	FALSE	FALSE
7	2025/6/17 9:50	M1002	46.3	37.1	50.22	FALSE	FALSE	FALSE
8	2025/6/17 10:10	M1002	53	37.3	51.06	FALSE	FALSE	FALSE
9	2025/6/17 10:30	M1002	47.4	33.5	49.78	FALSE	FALSE	FALSE
10	2025/6/17 10:50	M1002	54.7	40.3	51.24	FALSE	FALSE	FALSE
11	2025/6/17 11:10	M1002	77.8	35.2	55.84	TRUE	FALSE	TRUE
12	2025/6/17 11:30	M1002	77.8	39.1	62.14	TRUE	FALSE	TRUE

图 2-6 anomaly_report_M002.csv

	A	B	C
1	timestamp	device_id	is_anomaly
2	2025/6/17 8:00	M1001	FALSE
3	2025/6/17 8:20	M1001	FALSE
4	2025/6/17 8:40	M1001	TRUE
5	2025/6/17 9:00	M1001	FALSE
6	2025/6/17 9:20	M1001	TRUE
7	2025/6/17 9:40	M1001	TRUE
8	2025/6/17 10:00	M1001	FALSE
9	2025/6/17 10:20	M1001	FALSE
10	2025/6/17 10:40	M1001	FALSE
11	2025/6/17 11:00	M1001	FALSE
12	2025/6/17 11:20	M1001	FALSE
13	2025/6/17 11:40	M1001	FALSE
14	2025/6/17 12:00	M1001	FALSE

图 2-7 sensor_data_with_anomalies.csv

(五)、任务四：数据可视化

1. 创建直观的设备状态监测图表
2. 展示温度趋势变化和移动平均线，便于观察设备热状态
3. 绘制振动频率分析表，标记异常点
4. 生成专业的设备状态检测报告，支持多种语言显示和字体自适应

在程序编码补全中根据上述信息补全对应代码，其中数据以数据中心文件列表中的实际数据为准。

填写完毕后，可点击需求实现查看完整代码，确认无误后点击执行并显示结

果，待代码运行完毕后即可查看结果。

【参考答案】：

(1) subplots

输出：
中文字体设置成功：黑体

正在处理设备 M1001...
已生成分析图表: [UserFolderPath]/reports/device_analysis_M1001.png

正在处理设备 M1002...
已生成分析图表: [UserFolderPath]/reports/device_analysis_M1002.png

成功生成以下分析图表:
- [UserFolderPath]/reports/device_analysis_M1001.png
- [UserFolderPath]/reports/device_analysis_M1002.png

运行结束

图 2-8

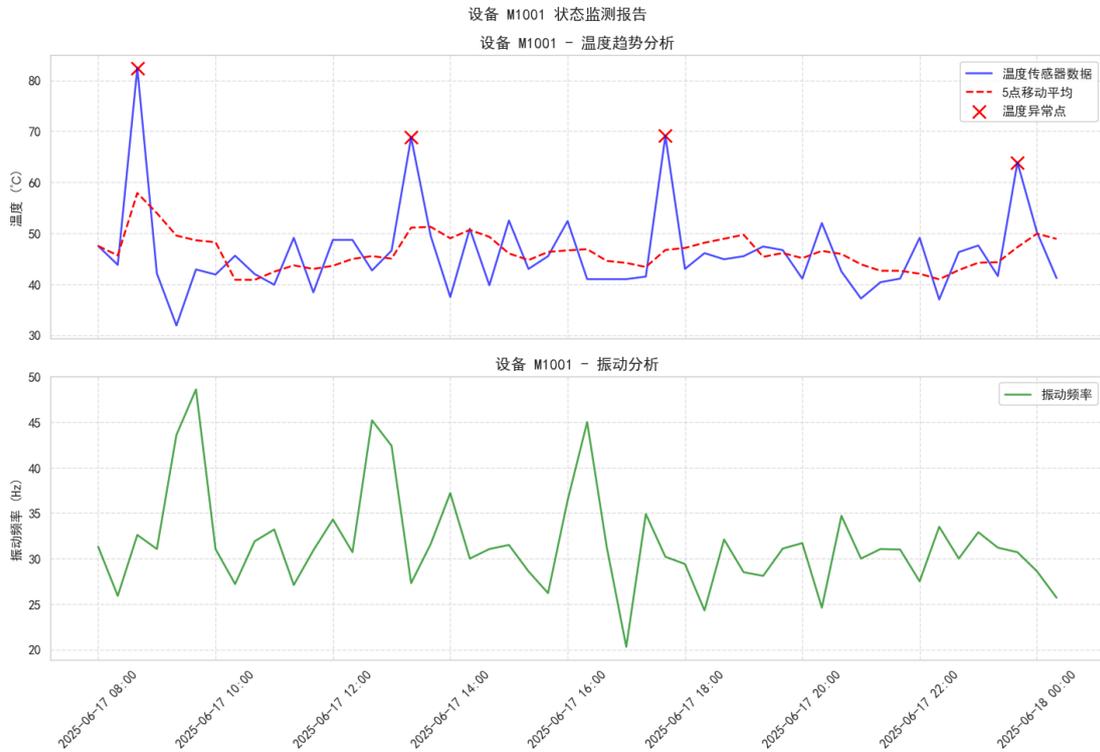


图 2-9 设备 M1001 状态监测报告

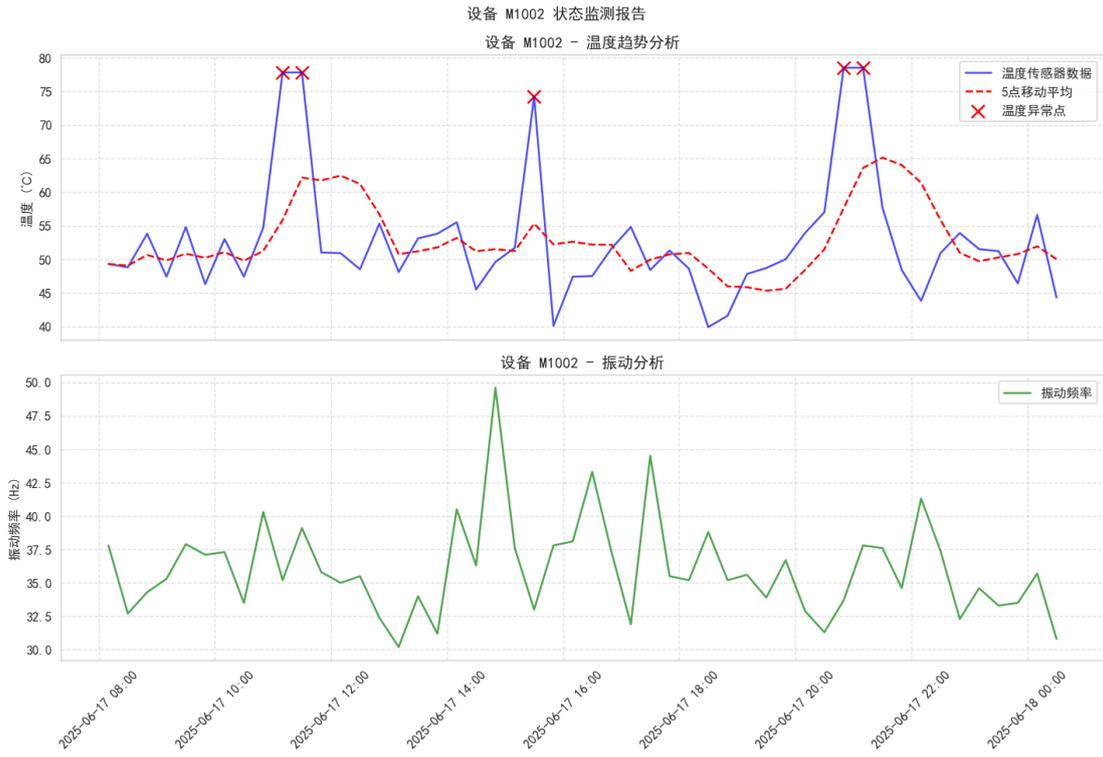


图 2- 10 设备 M1002 状态监测报告

项目三：人工智能在审计中的应用

任务一 审计文书撰写篇

任务二 审计访谈篇

任务三 专项审计篇

任务四 合同审计篇

任务五 审计程序篇

项目四：机器学习与应用

任务一 机器学习概述与应用场景

一、任务背景

在智能制造领域，生产模式选择与采购计划优化是提高生产效率与资源利用率的重要组成部分。针对生产模式选择，需要综合考虑历史订单数据、产品特性（如品种数、数量、规格化程度）、生产效率以及制造周期等因素，通过聚类分析将订单匹配到最合适的生产模式，从而实现生产资源的最优配置。

另一方面，在采购计划优化中，基于货品数量、预算金额、供应商距离、价格波动等多维数据，结合集中采购与分散采购的特点，通过决策树等方法选择最优采购模式，不仅能满足生产需求，还可实现成本节约与时间效率提升。这种结合数据分析与智能决策的方式，推动了智能制造领域从生产到采购全流程的高效化与科学化发展。

二、订单需求聚类

（一）、任务背景

智能制造领域各生产模式的特点如下所示：

1. 通用品流水生产：

①产品品种稳定，社会长期需要；②设计定型，产品标准化的（产品质量规格符合更广泛的标准，可使生产活动保持统一性与协调性，从而提高效率和销售面。）；③按时供应；机器设备常年完好，时效；实行计划预修制度；各生产环

节的工作质量稳定。产品检验能随生产在流水线上进行；④产品总数量大，约5000-20000件；⑤品种数少，约1-5种；⑥制造周期短，约1-10天；⑦成本低，约5000-15000元；⑧规格标准化；⑨生产效率高，约1.5-2。

2. 小批量多品种生产：

①在规定的生产期间内，产品种类（规格、型号、尺寸、形状、颜色等）较多，而每个种类产品生产数量较少；②效率低，成本高，不易实现自动化，生产计划与组织较复杂，但适合偏小众的、高级的、独特的和流行的商品；③产品总数量较多，约100-5000件；④品种数偏多，约10-50种；⑤制造周期较长，约10-30天；⑥成本较高，约15000-30000元；⑦规格不标准化；⑧生产效率较低，约0.5-1.0。

3. 定制（集单）生产：

①个性化、产品独特、生产方式灵活；②生产周期长，准备工作大，无法快速实施生产；③成本高，注重客户沟通；④产品数量最小，约10-500件；⑤品种数偏小，约1-10种；⑥制造周期长，约30-60天；⑦成本高，约30000-50000元；⑧规格不标准化；⑨生产效率低，约0.3-0.7。

4. 小单快返：

①先小批量生产不同款式的产品进行市场测试，再通过终端数据反馈，对其中的“爆款”进行快速返单，以此实现利润最大化，同时减少库存风险；②适用需求变化大的行业，需要做许多销售测试，依赖反馈情况；③产品数量较小，约100-2000件；④品种数偏多，约5-20种；⑤制造周期偏短，约5-15天；⑥成本稳定且较低，约10000-20000元；⑦生产效率较高，约1.0-1.5。

本任务是基于生产模式聚类.xlsx数据集的聚类分析，根据订单需求进行聚类，选择合适的生产模式进行生产制造。数据集中包含了多种订单需求，包括产品总数量、工艺复杂度、产品规格是否标准化、制造周期/天、销售周期/天、生产效率、生产成本、生产模式等，具体数据如下图所示：

	A	B	C	D	E	F	G	H	I
1	品种数	产品总数量	工艺复杂指数	产品规格是否标准化	制造周期/天	销售周期/天	生产效率	生产成本	生产模式
2	4	11710	0.68	1	1	34	1.88	13308.78	通用品流水生产
3	4	14914	0.69	1	6	31	1.88	8074.7	通用品流水生产
4	1	16298	0.92	1	2	29	1.86	10444.13	通用品流水生产
5	3	14596	0.76	1	7	30	1.79	6540.74	通用品流水生产
6	3	9143	0.96	1	5	10	1.78	5597.12	通用品流水生产
7	4	11658	1.21	1	9	18	1.89	9448.88	通用品流水生产
8	1	7878	0.89	1	6	34	1.98	10120.73	通用品流水生产

图 4-1

(二)、肘部法则确定聚类数

加载包含 8 个特征的生产模式数据并利用肘部法则分确定聚类数目。

具体要求：

1. 计算 $K=1$ 到 10 时的簇内平方和 (WSS)
2. 绘制 WSS 随 K 值变化的曲线图
3. 通过观察拐点确定最佳聚类数

在程序编码补全中根据上述信息补全对应代码，其中数据以数据中心文件列表中的实际数据为准。

填写完毕后，可点击需求实现查看完整代码，确认无误后点击执行并显示结果，待代码运行完毕后即可查看结果。

【参考答案】：

- (1) 1, 11
- (2) KMeans
- (3) ft

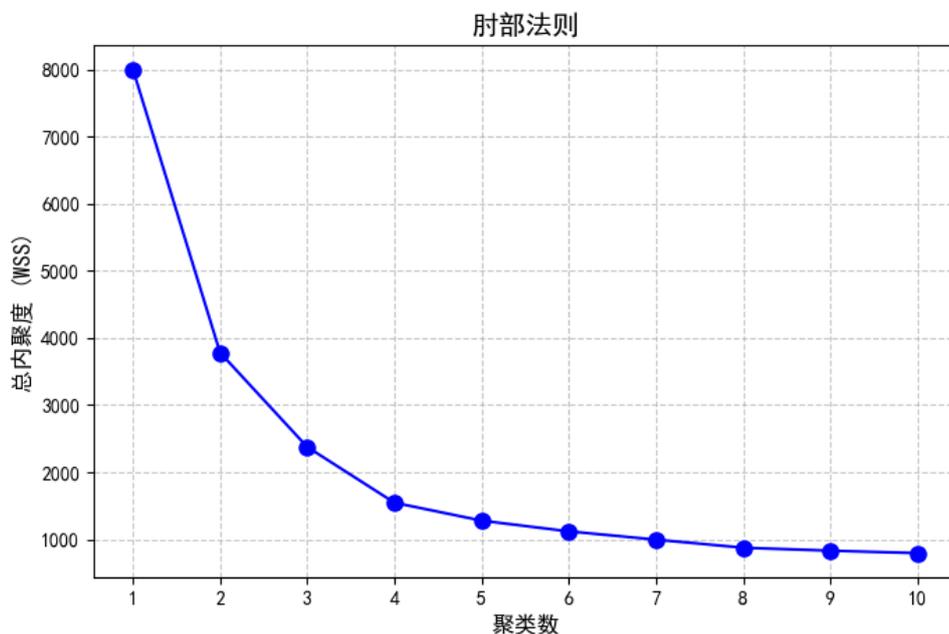


图 4-2

步骤分析

结果分析(填空题 4.00 分)

1. 通过观察拐点确定最佳聚类数为

(1)

正确答案：

(1)4

(三)、订单需求聚类

对订单需求进行聚类，要求如下：

1. 使用 K-Means 算法（利用肘部法则分析已确定最佳聚类数）对标准化后的数据进行聚类

2. 聚类结果可视化结果可视化

在程序编码补全中根据上述信息补全对应代码，其中数据以数据中心文件列表中的实际数据为准。

填写完毕后，可点击需求实现查看完整代码，确认无误后点击执行并显示结果，待代码运行完毕后即可查看结果。

【参考答案】：

(1)StandardScaler

(2)4

(3)kmeans.fit_predict

输出：

聚类结果已保存，前5条数据：

	品种数	产品总数量	工艺复杂指数	制造周期/天	销售周期/天	生产效率	生产成本	cluster
0	4	11710	0.68	1	34	1.88	13308.78	2
1	4	14914	0.69	6	31	1.88	8074.70	2
2	1	16298	0.92	2	29	1.86	10444.13	2
3	3	14596	0.76	7	30	1.79	6540.74	2
4	3	9143	0.96	5	10	1.78	5597.12	2

运行结束

图 4-3

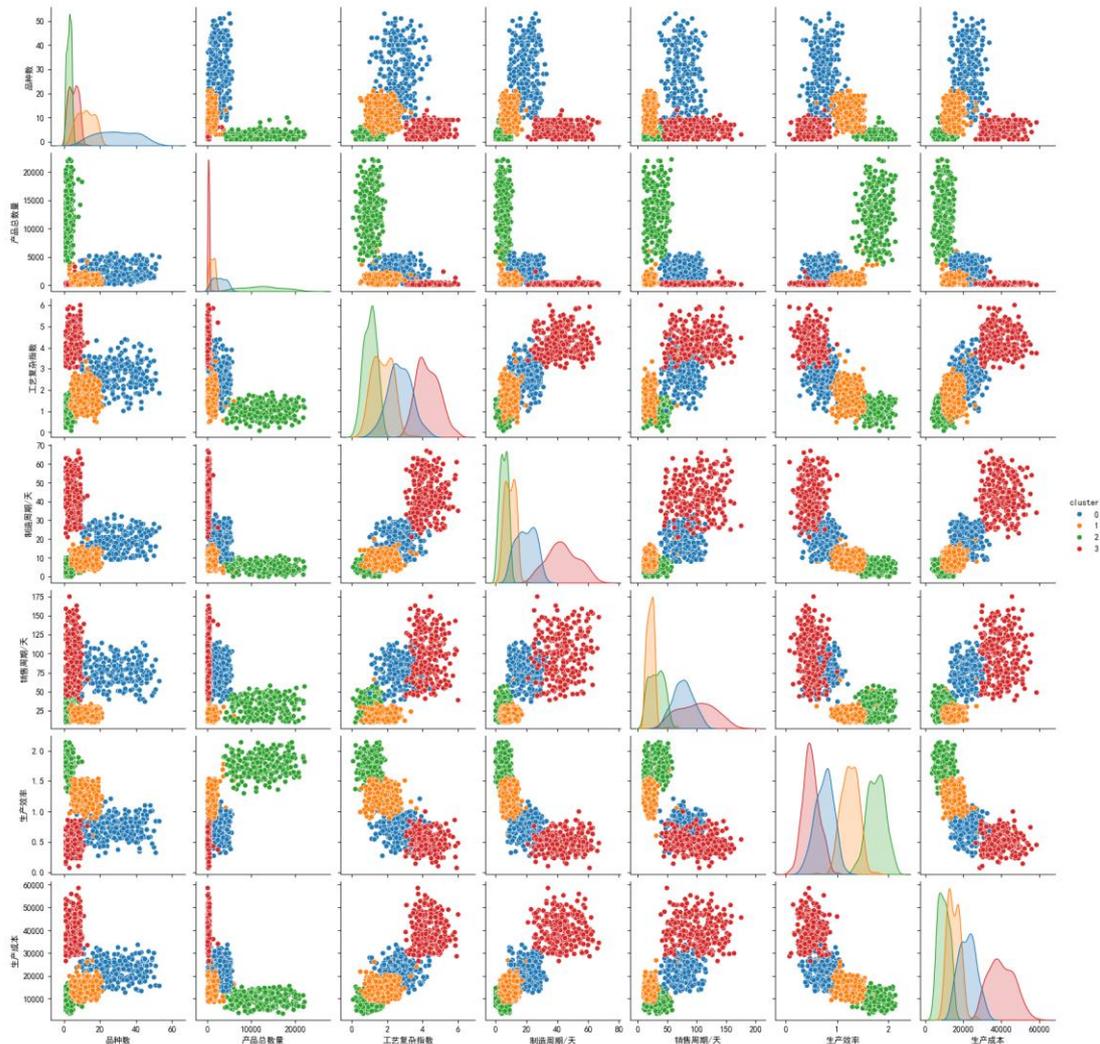


图 4- 4 聚类结果图

步骤分析

结果分析（填空题 4.00 分）

1. 根据生产模式聚类可视化(聚类结果), 蓝色的数据点代表的类别是: (1), 黄色的数据点代表的类别是: (2), 绿色的数据点代表的类别是: (3), 红色的数据点代表的类别是: (4)。

正确答案:

- (1) 小批量多品种生产
- (2) 小单快返
- (3) 通用品流水生产
- (4) 定制（集单）生产

(四)、生产模式匹配

将得到的聚簇与题目背景中的四种生产模式匹配。

在程序编码补全中根据上述信息补全对应代码，其中数据以数据中心文件列表中的实际数据为准。

填写完毕后，可点击需求实现查看完整代码，确认无误后点击执行并显示结果，待代码运行完毕后即可查看结果。

【参考答案】：

(1) groupby

(2) similarity_scores

(3) map

输出:

=== 聚类特征分析 ===

各簇的归一化特征均值:

	品种数	产品总数量	工艺复杂指数	制造周期/天	销售周期/天	生产效率	生产成本
cluster							
0	1.000000	0.198491	0.513519	0.374906	0.706374	0.201950	0.412017
1	0.345502	0.070275	0.212874	0.105217	0.000000	0.596479	0.174072
2	0.000000	1.000000	0.000000	0.000000	0.122924	1.000000	0.000000
3	0.090461	0.000000	1.000000	1.000000	1.000000	0.000000	1.000000

=== 相似度得分 ===

簇 0 与各生产模式的相似度(距离越小越相似):

小批量多品种生产: 0.354

小单快返: 0.964

定制生产: 1.151

通用品流水生产: 1.550

簇 1 与各生产模式的相似度(距离越小越相似):

小单快返: 0.338

通用品流水生产: 1.019

小批量多品种生产: 1.020

定制生产: 1.516

簇 2 与各生产模式的相似度(距离越小越相似):

通用品流水生产: 0.276

小单快返: 1.164

小批量多品种生产: 1.654

定制生产: 2.085

簇 3 与各生产模式的相似度(距离越小越相似):

定制生产: 0.418

小批量多品种生产: 1.337

小单快返: 1.651

通用品流水生产: 2.174

=== 最终映射关系 ===

簇 0 → 小批量多品种生产

簇 1 → 小单快返

簇 2 → 通用品流水生产

簇 3 → 定制生产

结果已保存到 [UserFolderPath]/生产模式分类结果.csv

运行结束

图 4-5

步骤分析

结果分析（填空题 4.00 分）

1. 根据匹配结果中各簇与各生产模式的相似度可以得到最终映射关系为：簇 0→(1)；簇 1→(2)；簇 2→(3)；簇 3→(4)

正确答案：

(1)小批量多品种生产

(2)小单快返

(3)通用品流水生产

(4)定制生产

（五）、聚类结果评估

将聚类匹配结果与原始结果的生产模式标签进行匹配计算正确率。

在程序编码补全中根据上述信息补全对应代码，其中数据以数据中心文件列表中的实际数据为准。

填写完毕后，可点击需求实现查看完整代码，确认无误后点击执行并显示结果，待代码运行完毕后即可查看结果。

【参考答案】：

(1)confusion_matrix

(2)heatmap

(3)classification_report

```
输出:
正在加载并合并数据...

合并后的数据前5行:
  品种数 产品总数量 工艺复杂指数 产品规格是否标准化 ... 生产成本 生产模式_真实 cluster
生产模式_预测
0  4  11710  0.68    1 ... 13308.78 通用品流水生产    2 通用品流水生产
1  4  14914  0.69    1 ...  8074.70 通用品流水生产    2 通用品流水生产
2  1  16298  0.92    1 ... 10444.13 通用品流水生产    2 通用品流水生产
3  3  14596  0.76    1 ...  6540.74 通用品流水生产    2 通用品流水生产
4  3   9143  0.96    1 ...  5597.12 通用品流水生产    2 通用品流水生产

[5 rows x 11 columns]

开始评估聚类匹配结果...

整体准确率: 100.00%

详细分类报告:
      precision  recall  f1-score  support

定制 (集单) 生产    1.00    1.00    1.00    250
  小单快返    1.00    1.00    1.00    250
小批量多品种生产    1.00    1.00    1.00    250
通用品流水生产    1.00    1.00    1.00    250

accuracy                1.00    1000
macro avg    1.00    1.00    1.00    1000
weighted avg    1.00    1.00    1.00    1000

各簇内部准确率:
  cluster 准确率
0     0  1.0
1     1  1.0
2     2  1.0
3     3  1.0

评估结果已保存到 聚类评估结果.csv

运行结束
```

图 4-6

步骤分析

结果分析(判断题 3.00 分)

- 1. 聚类匹配结果整体准确率为 100.00%.

正确答案：正确

三、决策树采购计划决策

(一)、任务描述

货品采购清单，含有：货品数量、预算金额、供应商距离、完成时限、价格波动、技术参数、采购方式。采购需遵循适价、适时、适质、适量、适地原则。

变量抉择：采购需求通常包括项目名称、数量、预算金额、完成时限、采购方式和技术参数等。在决定采用集中采购还是分散采购时，应考虑下面的因素或标准：

1. 采购需求的通用性
2. 地理位置
3. 供应市场结构
4. 潜在的节约
5. 所需的专门技术
6. 价格波动

基于采购数据选择采购模式：集中采购 o 分散采购。

本任务是根据决策树采购计划决策.xlsx 数据集构建一个基于决策树的分类模型，根据货品数量、预算金额、供应商距离、完成时限、价格波动、技术参数等特征，预测企业应该采用“分散采购”还是“集中采购”方式，具体数据如下图所示。

	A	B	C	D	E	F	G
1	货品数量	预算金额	供应商距离	完成时限	价格波动	技术参数	采购方式
2	70	13823	61	84	2	1	分散采购
3	357	91730	983	23	1	1	集中采购
4	77	37230	697	38	2	2	分散采购
5	338	8794	220	9	2	0	集中采购
6	749	75924	706	47	2	1	集中采购
7	606	71938	835	19	2	2	集中采购

图 4-7

(二)、操作说明

以下是任务所需使用的模块的含义及功能：

1. train_test_split()

功能：将数据集随机划分为训练集和测试集。

参数：test_size=0.3：测试集占总数据集的 30%；random_state=42：随机种子，保证每次运行时的分割结果一致。

2. DecisionTreeClassifier()

功能：创建一个决策树分类器模型。

参数：criterion='entropy'：使用信息增益作为划分标准；max_depth=5：限制树的深度为 5，避免过拟合；random_state=42：设置随机种子，保证结果一致性。

3. model.fit()

功能：训练模型。

4. model.predict()

功能：用训练好的模型对新数据进行预测。

5. metrics.accuracy_score()

功能：计算模型的准确率（accuracy）。

6. metrics.confusion_matrix()

功能：生成混淆矩阵，显示预测分类与实际分类的对比。

（三）、任务实施

请根据任务操作说明完成决策树计划采购决策。

在程序编码补全中根据上述信息补全对应代码，其中数据以数据中心文件列表中的实际数据为准。

填写完毕后，可点击需求实现查看完整代码，确认无误后点击执行并显示结果，待代码运行完毕后即可查看结果。

【参考答案】：

(1) train_test_split

(2) entropy

(3) fit

(4) predict

(5) confusion_matrix

输出:

模型准确率: 0.97

混淆矩阵:

[[12 1]

[1 46]]

运行结束

图 4-8

采购决策树模型

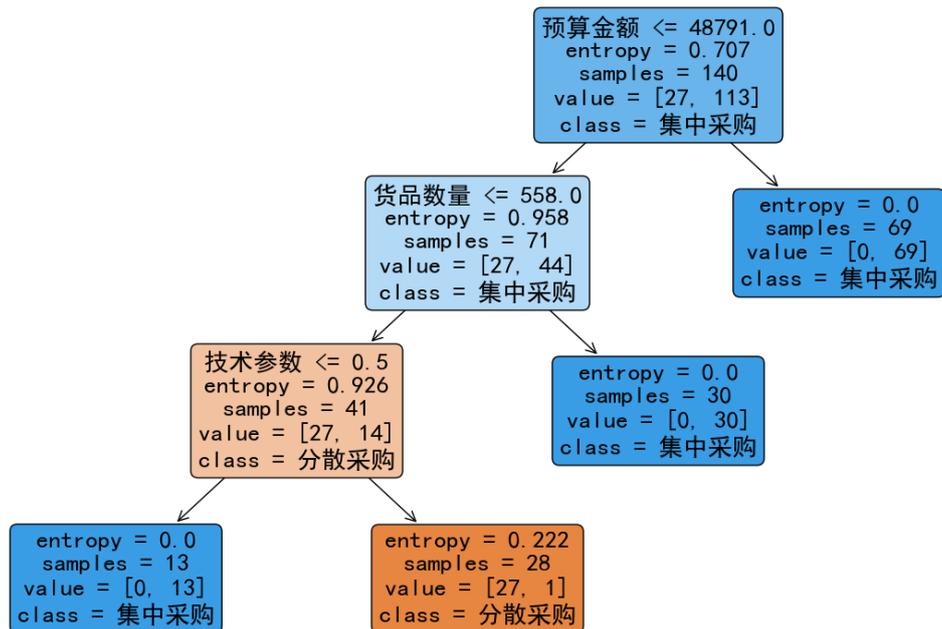


图 4-9

步骤分析:

结果分析(填空题 4.00 分)

1. 由模型输出结果可知, 模型准确率为: (1)

正确答案:

(1)0.97

结果分析(填空题 4.00 分)

2. 通过混淆矩阵结果, 我们可以得到: 在用于测试的 60 个样本中, 判断错误的样本有(1)个。

正确答案:

(1)2

任务二 机器学习模型与方法

一、任务背景

在智能制造领域，设备状态预测与生产质量分析是保障高效生产和产品质量的重要环节。设备状态预测基于设备的历史数据，包括设备购入时长、累计工作时长、品牌口碑、标定使用年限、上次维修时间、性能评分、设备效率（OEE）、能源消耗以及故障状态等信息，结合机器学习模型实现对故障状态的精确预测，帮助判断设备是否能直接投入使用或需进行维修，从而优化设备管理与维护计划。

此外，在生产质量分析中，通过构建基于随机森林的模型，将温度、压力、速度、湿度、加工时间等生产参数与产品缺陷状态进行关联分析，精准识别影响产品质量的关键因素。模型输出特征的重要性排序，为深入理解生产过程中的关键变量提供直观支持。

二、随机森林工业制造缺陷影响因素分析

（一）、任务说明

本任务是根据 随机森林工业制造缺陷.xlsx 数据集，构建一个随机森林模型，用于分析生产参数（如温度、压力、速度、湿度、加工时间等）对产品缺陷状态的影响。模型将输出每个特征的重要度，根据特征的重要性排序，识别出对缺陷状态影响最大的变量。通过可视化技术，展示特征的重要性排名，帮助深入理解生产过程中的关键影响因素。识别哪些生产参数对产品缺陷状态影响最大。

特征（生产参数）：温度、压力、速度、湿度、加工时间等。

目标变量：缺陷状态（0 为无缺陷，1 为有缺陷）。

具体数据如下图所示：

	A	B	C	D	E	F	
1	温度	压力	速度	湿度	加工时间	缺陷状态	
2	290.89	5252.82	1.49	48.63	125.64	1	
3	298.11	4851.79	1.51	55.28	134.12	0	
4	289.73	5155.17	1.33	42.17	106.34	1	
5	313.93	4464.28	1.68	51.94	118.47	1	
6	272.33	5114.49	1.35	37.09	113.14	1	
7	292.99	4880.59	1.33	61.04	118.64	1	
8	334.39	4714.3	1.21	55.65	113.37	1	

图 4- 10

(二)、任务操作

请根据说明说明建立随机森林模型进行分析。

在程序编码补全中根据上述信息补全对应代码，其中数据以数据中心文件列表中的实际数据为准。

填写完毕后，可点击需求实现查看完整代码，确认无误后点击执行并显示结果，待代码运行完毕后即可查看结果。

【参考答案】：

(1) df. drop

(2) fit

(3) predict

(4) accuracy_score

输出：

模型准确率: 0.99

分类报告:

	precision	recall	f1-score	support
0	1.00	0.96	0.98	50
1	0.99	1.00	1.00	250
accuracy		0.99	300	
macro avg	1.00	0.98	0.99	300
weighted avg	0.99	0.99	0.99	300

特征重要性:

特征	重要性
0 温度	0.355866
1 压力	0.353954
3 湿度	0.222075
2 速度	0.036490
4 加工时间	0.031615

运行结束

图 4- 11

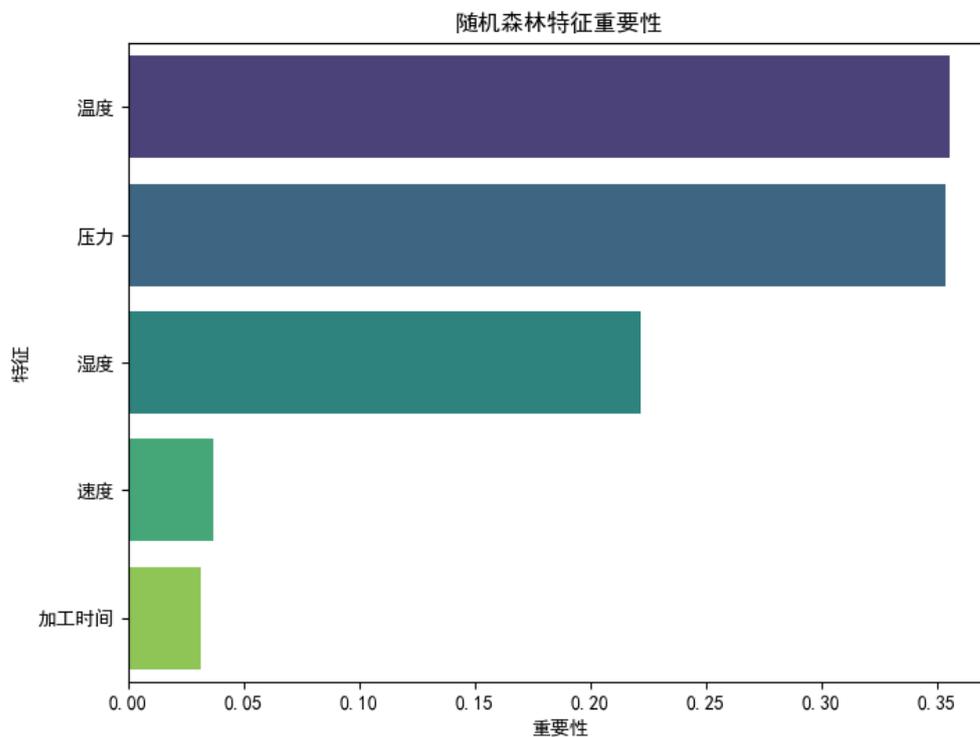


图 4- 12

三、SVM 实现生产设备故障状态预测

(一)、数据计算操作说明

根据生产设备的历史数据（设备购入时长，具体工作时长，品牌（制造商）口碑，标定使用年限，距离上次维修时间，性能评分，设备效率 OEE, 能源消耗，故障状态)通过对机器故障状态的预测，从而判断机器是否能直接投入使用。

故障：需要维修；无故障：可直接使用

公式：OEE=稼动率×产能效率×良率

其中：

稼动率=实际工作时间/计划工作时间

产能效率=实际产能/标准产能

良率=良品数/实际生产数

请根据公式计算表格。

(二)、数据计算

请根据数据计算操作说明计算相关变量并上传文件。

首先下载答题卡，根据上述公式说明，在 Excel 的第 M2 单元格内输入公式 =E2/D2, 在第 N2 单元格内输入公式 =F2/E2, 在第 O2 单元格内输入公式 =L2*M2*N2, 分别按 Enter 键得到计算结果并应用于该列，即可得到如下的计算结果。

J	K	L	M	N	O	P
PerformanceScore	EnergyConsumption	Availability	PerformanceEfficiency	QualityRate	OEE	FaultCondition
77.85	941.19	0.976845328	0.828884402	0.919990541	0.744908849	0
92.62	312.43	0.810769039	0.944353364	0.990572722	0.758434451	0
97.54	996.66	0.957047036	0.997303687	0.987084379	0.942139009	1
73.52	383.88	0.978982907	0.850088722	0.913387643	0.76014159	0
67.21	450.85	0.82138376	0.999689002	0.897199478	0.736715893	0
81.22	553.19	0.94504628	0.855244045	0.891469349	0.720525825	1
75.86	492.86	0.96712599	0.867516906	0.966856577	0.811190876	1
90.2	968.14	0.882672478	0.886711549	0.892428566	0.698482313	1
63.55	684.48	0.819547648	0.825008539	0.90517436	0.612018986	1
89.48	738.11	0.820778878	0.857219427	0.90128081	0.634130002	1
60.02	800.47	0.928956261	0.888969489	0.997839233	0.824029382	1
95.65	632.3	0.91971798	0.866626608	0.909725776	0.725098816	1
70.89	235.92	0.995278689	0.881578996	0.862637091	0.756892265	0
65.51	932.09	0.942054106	0.811574308	0.880838998	0.673442734	1

图 4-13

(三)、SVM 分析说明

本任务是构建一个基于支持向量机(SM)的分类模型，用于预测生产设备的故障状态（正常或故障），以下是代码关键函数说明：

1. StandardScaler: 标准化处理，将特征值转换为均值为 0、标准差为 1 的分布。
2. 使用 SVC(支持向量机)进行训练，kernel='rbf':SVM 使用径向基核函数。
3. accuracy_score: 计算预测值与真实值的准确率。
4. classification_report: 生成包括精确率、召回率、F1 分数等指标的详细报告。
5. pd. crosstab: 计算实际值与预测值之间的交叉表，用于生成混淆矩阵。
6. sns. heatmap: 绘制热力图，显示混淆矩阵内容。

(四)、SVM 分析

请根据 SVM 分析说明完成下面任务。

在程序编码补全中根据上述信息补全对应代码，其中数据以数据中心文件列表中的实际数据为准。

填写完毕后，可点击需求实现查看完整代码，确认无误后点击执行并显示结果，待代码运行完毕后即可查看结果。

【参考答案】：

(1)StandardScaler

(2) fit

(3) predict

(4) accuracy_score

输出:

Accuracy: 0.92

	precision	recall	f1-score	support
0	0.98	0.84	0.90	95
1	0.87	0.98	0.92	105
accuracy			0.92	200
macro avg	0.92	0.91	0.91	200
weighted avg	0.92	0.92	0.91	200

运行结束

图 4- 14

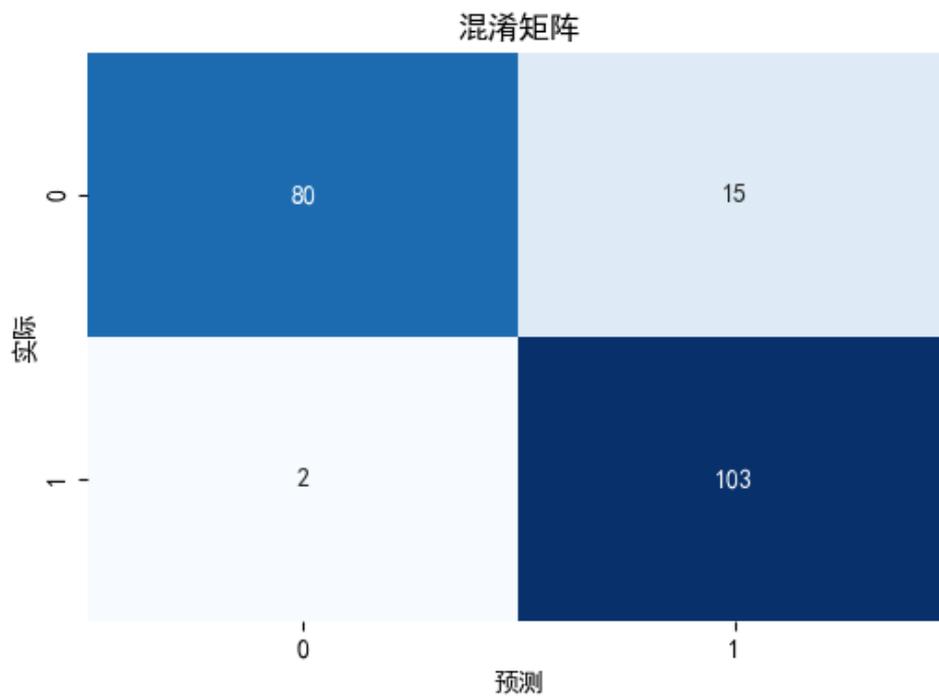


图 4- 15

步骤分析:

结果分析(填空题 4.00 分)

1. 类别 0: 召回率为(1), 说明实际为 0 的样本中, 有(2)%被正确预测为 0。

类别 1: 召回率为(3), 说明实际为 1 的样本中, 有(4)%被正确预测为 1。

正确答案:

(1)0.84

(2)84

(3)0.98

(4)98

结果分析(填空题 4.00 分)

2. 类别 0: F1-score 为(1), 说明对类别 0 的综合预测性能较好。类别 1: F1-score 为(2), 说明对类别 1 的综合预测性能更强。

正确答案:

(1)0.90

(2)0.92

任务三 机器学习的评价指标与可视化工具

一、任务背景

在智能交通与车辆性能分析领域, 燃油效率预测是提升汽车设计优化与能源利用率的重要环节。基于 UCI 机器学习存储库中的 Auto MPG 数据集, 通过回归分析方法预测汽车的燃油效率 (MPG, 即每加仑行驶的英里数)。该数据集包含多个影响燃油效率的关键特征, 如发动机排量(displacement)、重量(weight)、马力(horsepower)、加速度(acceleration)等。

通过分析这些特征与燃油效率之间的关系, 构建高精度的回归预测模型, 识别对燃油效率影响最大的变量。模型结果结合可视化技术, 揭示特征的重要性排序与影响规律, 为汽车设计优化提供数据支持。

二、回归分析汽车燃油效率

(一)、任务说明

数据有关汽车的技术规格, 来源于 UCI 机器学习存储库。

基于回归分析汽车燃油效率.xs 数据集的回归分析, 用来预测汽车的燃油效率(MPG, 即每加仑行驶的英里数)。数据集中包含了影响汽车燃油效率的多个特征, 包括发动机排量(displacement)、重量(weight)、马力(horsepower)、加速度(acceleration)等, 具体数据如下图所示:

mpg	cylinders	displacement	horsepower	weight	acceleration	model year	origin	car name
18	8	307	130	3504	12	70	1	chevrolet chevelle malibu
15	8	350	165	3693	11.5	70	1	buick skylark 320
18	8	318	150	3436	11	70	1	plymouth satellite
16	8	304	150	3433	12	70	1	amc rebel sst
17	8	302	140	3449	10.5	70	1	ford torino
15	8	429	198	4341	10	70	1	ford galaxie 500
14	8	454	220	4354	9	70	1	chevrolet impala
14	8	440	215	4312	8.5	70	1	plymouth fury iii

图 4-16

通过分析和建模，探索各个特征对燃油效率的影响，并构建回归预测模型。

Y	X1	X2	X3	X4	X5	X6	X7	X8	X9
mpg	气缸	排量	马力	重量	加速 度	车型 年份	地区 1	地区 2	地区 3

(二)、线性回归分析

请根据任务说明完成线性回归分析。

在程序编码补全中根据上述信息补全对应代码，其中数据以数据中心文件列表中的实际数据为准。

填写完毕后，可点击需求实现查看完整代码，确认无误后点击执行并显示结果，待代码运行完毕后即可查看结果。

【参考答案】：

(1) heatmap

(2) train_test_split

(3) LinearRegression

输出:

X_train.shape: (278, 9)

X_test.shape: (120, 9)

y_train.shape: (278,)

y_test.shape: (120,)

均方误差(MSE): 9.199829719898041

R-squared: 0.8397001030031428

Column: cylinders

Weight: 0.01350067895789724

Column: displacement

Weight: 0.012386431307854818

Column: horsepower

Weight: -0.00581627239996938

Column: weight

Weight: -0.006787764366854158

Column: acceleration

Weight: 0.12348869856247928

Column: model year

Weight: 0.8160123684104331

Column: origin_1

Weight: -2.0063789258946714

Column: origin_2

Weight: 1.0449959691275343

Column: origin_3

Weight: 0.9613829567671367

运行结束

图 4- 17

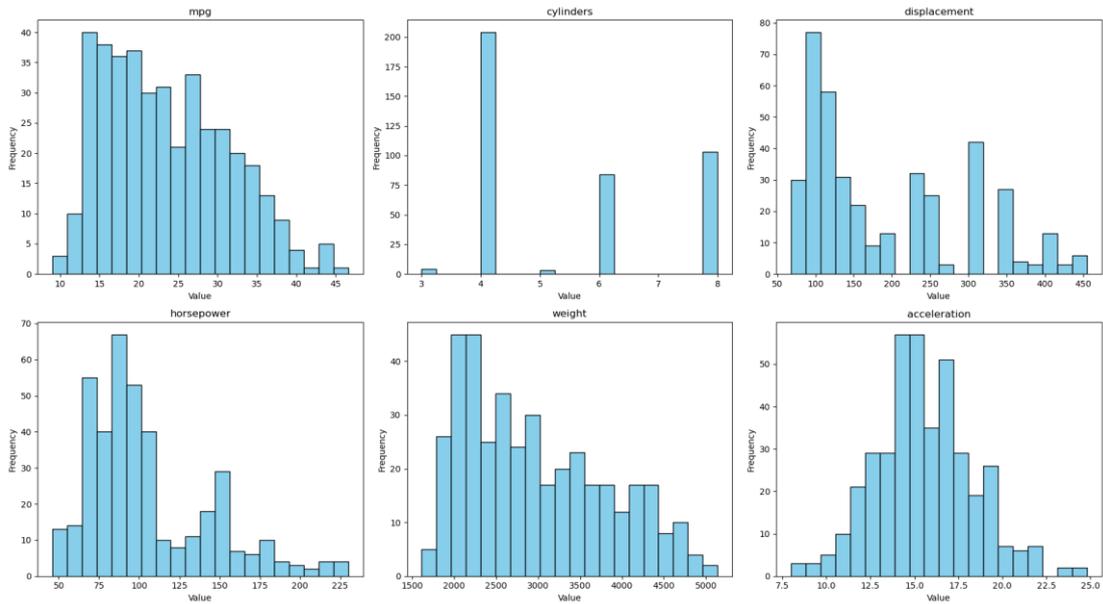


图 4-18 直方图

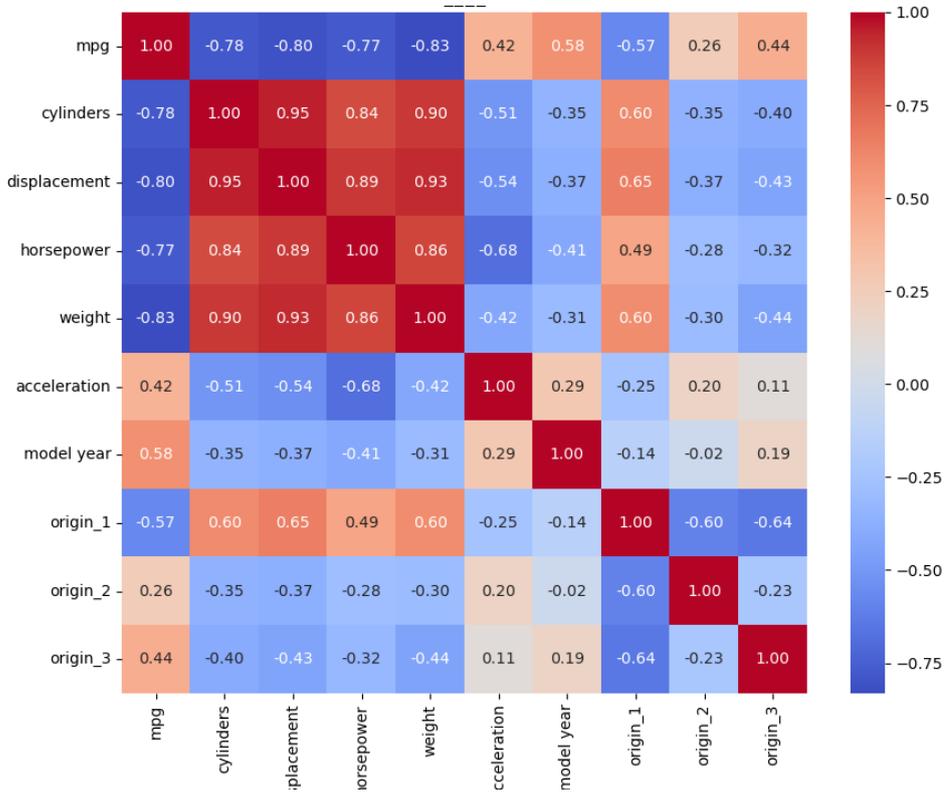


图 4-19 热力图

这张汽车变量的相关性热力图显示：油耗（mpg）与气缸数、排量、马力、车重呈强负相关，意味着动力越强、车身越重，油耗通常越高；而气缸数、排量、马力、车重这些动力相关参数之间呈强正相关，体现了它们随车型规格提升的同步性；此外，代表产地的 origin_1、origin_2、origin_3 之间呈强负相关，这是因为单辆车仅对应一个产地，不同产地变量具有互斥性。

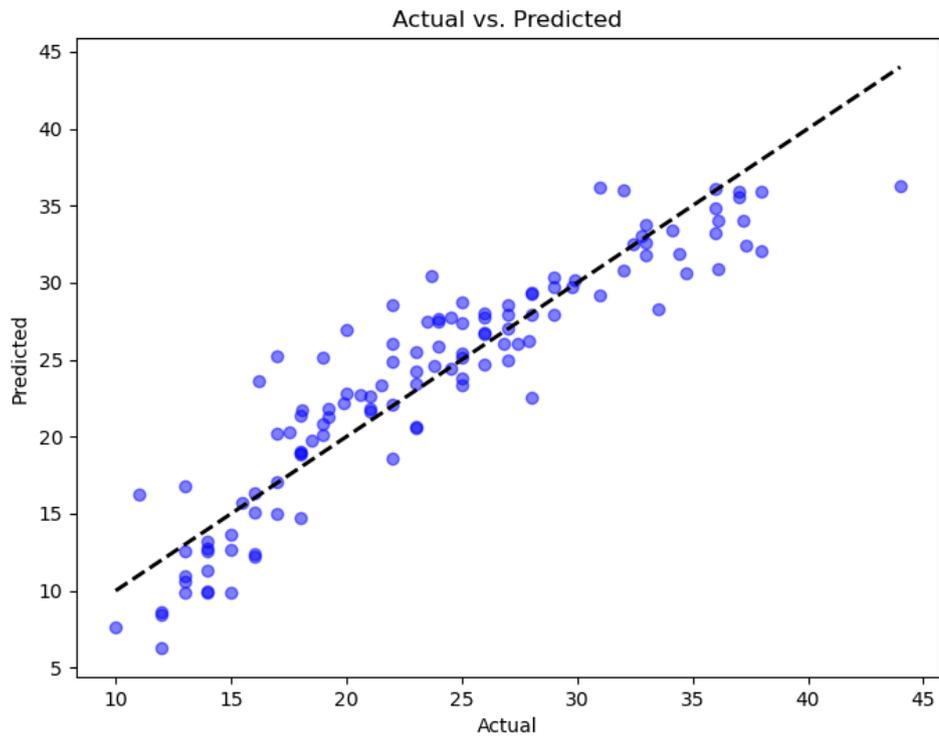


图 4- 20 线性回归图

步骤分析:

结果分析(填空题 4.00 分)

1. 由热力图分析可知, 个变量之间相关性较(1) (高/低)

正确答案:

(1) 高

结果分析 (填空题 4.00 分)

2. 根据线性回归模型拟合结果, 给出回归方程: (保留三位小数)

$Y = (1) * X_1 + (2) * X_2 - (3) * X_3 - (4) * X_4 + (5) * X_5 + (6) * X_6 - (7) * X_7 + (8) * X_8 + (9) * X_9$

正确答案:

(1) 0.014

(2) 0.012

(3) 0.006

(4) 0.007

(5) 0.123

(6) 0.816

(7) 2.006

(8) 1.045

(9) 0.691

结果分析（多选题 3.00 分）

3. 汽车燃油功率起到副作用的变量有：

A. 气缸

B. 马力

C. 重量

D. 加速度

正确答案：BC

结果分析（填空题 4.00 分）

4. 线性回归模型的 MSE(保留三位小数)：(1)；R2: (2)

正确答案：

(1) 9.200

(2) 0.840

（三）、Lasso 和 ridge 回归分析

由上一节分析我们可以得出，该数据集变量之间有较强的相关性。说明变量之间可能存在多重共线性，会导致普通的线性模型出现过拟合、预测准确度下降等问题。

为了解决上述问题，我们尝试运用 Lasso 和 ridge 回归分析进行模型预测，并将他们的预测结果进行比较，得出结论。

在程序编码补全中根据上述信息补全对应代码，其中数据以数据中心文件列表中的实际数据为准。

填写完毕后，可点击需求实现查看完整代码，确认无误后点击执行并显示结果，待代码运行完毕后即可查看结果。

【参考答案】：

(1) X.pop

(2) Lasso

(3) Ridge

```
输出:
Best alpha for Lasso: 0.1
(Lasso) Mean Squared Error (MSE): 9.089727949237052
(Lasso) R-squared: 0.8416185409561796

Column: cylinders
Lasso Weight: 0.0

Column: displacement
Lasso Weight: 0.007641092976791166

Column: horsepower
Lasso Weight: -0.004206257398867426

Column: weight
Lasso Weight: -0.0066303331498859305

Column: acceleration
Lasso Weight: 0.09141545079310151

Column: model year
Lasso Weight: 0.799724833149304

Column: origin_1
Lasso Weight: -2.170856285351352

Column: origin_2
Lasso Weight: 0.0

Column: origin_3
Lasso Weight: 0.0
Best alpha for Ridge: 10.0
(Ridge) Mean Squared Error (MSE): 9.154084484290687
(Ridge) R-squared: 0.8404971782511874

Column: cylinders
Ridge Weight: 0.04734418002701921

Column: displacement
Ridge Weight: 0.009270015286182433

Column: horsepower
Ridge Weight: -0.003435570096814272

Column: weight
Ridge Weight: -0.006766212924646262

Column: acceleration
Ridge Weight: 0.1215486067524941

Column: model year
Ridge Weight: 0.8100699380332453

Column: origin_1
Ridge Weight: -1.670727396685844

Column: origin_2
Ridge Weight: 0.8610765855411878

Column: origin_3
Ridge Weight: 0.8096508111446229

运行结束
```

图 4- 21

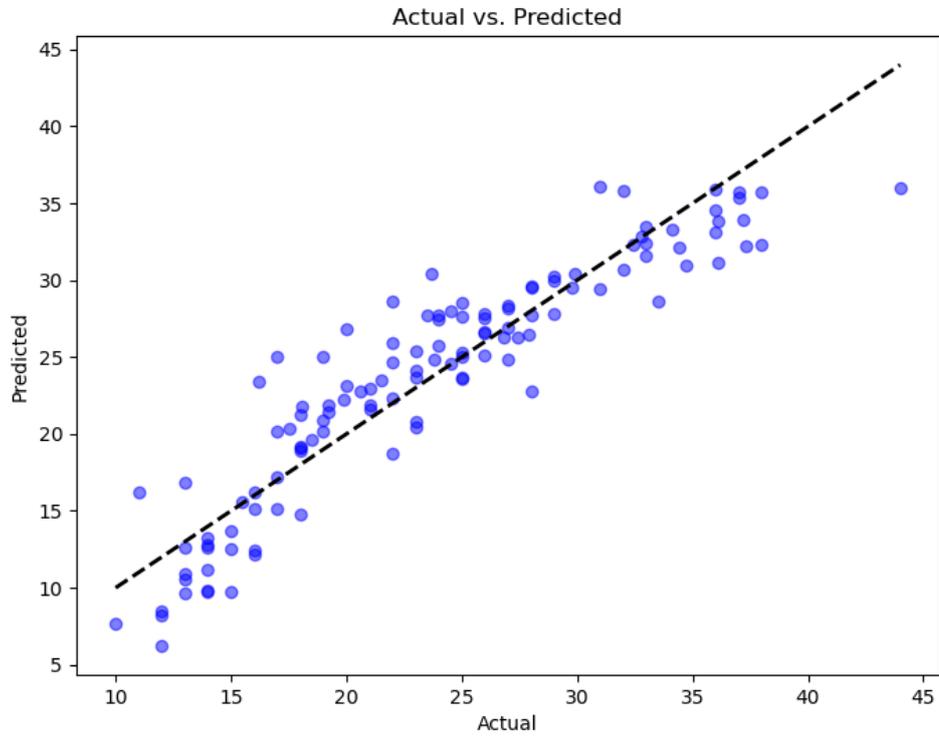


图 4- 22 Ridge 线性回归图

观察上图可知，散点整体沿对角线（理想预测线）分布，说明模型预测值与实际值的趋势一致性较好；但部分散点（尤其是实际值偏低或偏高的区域）偏离对角线，体现 Ridge 在处理极端值时仍有一定误差，不过整体拟合程度处于合理范围。

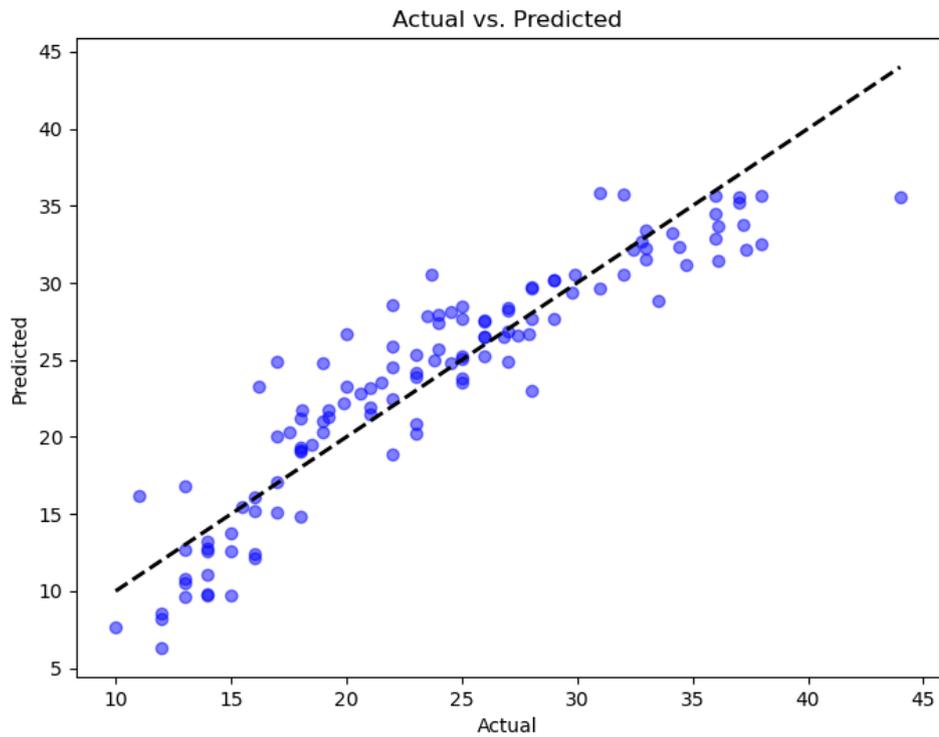


图 4-23 Lasso 线性回归图

散点的分布规律与 Ridge 相似，同样围绕对角线聚集，说明 Lasso 的预测趋势与实际值匹配度较高；对比来看，两张图的散点密集程度、偏离幅度差异不大，意味着在该数据场景下，Ridge 与 Lasso 的预测效果较为接近。

步骤分析：

结果分析（单选题 2.00 分）

1. 从 MSE 的角度来看，哪种模型的拟合效果更好？

- A. Lasso 回归
- B. Ridge 回归

正确答案：A

项目五：机器学习在审计科研中的应用

一、学习目标

任务	学习目标	课时
任务一：分类与预测模型介绍	1.了解常量模型、K 最邻近、随机森林、线性回归模型的定义、适用范围、函数、特点、优缺点、算法步骤等内容 2.了解逻辑回归、朴素贝叶斯、决策树、支持向量机等模型的定义,适用范围、优缺点等内容 3.了解规则归纳、自适应梯度、算法、随机梯度下降、神经网络等模型的定义、适用范围、特点和优缺点等内容	2
任务二：聚类分析模型介绍	1.了解 K 均值聚类模型的定义、适用范围、函数、特点、优缺点 2.了解层次聚类模型的定义、适用范围、函数、特点、优缺点 3.了解 DBSCAN(密度聚类)的定义、适用范围、函数、特点、优缺点 4.了解社区发现算法的定义、适用范围、函数、特点、优缺点	2
任务三：数据降维模型介绍	1.了解 K 均值聚类模型的定义、适用范围、函数、特点、优缺点 2.了解层次聚类模型的定义、适用范围、函数、特点、优缺点 3.了解 DBSCAN(密度聚类)的定义、适用范围、函数、特点、优缺点 4.了解社区发现算法的定义、适用范围、函数、特点、优缺点	2
任务四：商务数据智能分析挖掘工具介绍	1.掌握商务分析与挖掘工具的数据上传、数据处理和数据分析的基本操作 2.能够运用商务分析与挖掘中的 ARMIA 模型、K 均值聚类模型进行预测和分类	2

二、知识准备

通过平台的 PPT 完成知识点的学习，见【知识准备】。

三、随堂练习

完成知识点的学习后，可【随堂练习】进行学习巩固。

序号	知识点	习题数量
1	常量模型	5

2	K 最邻近	5
3	随机森林	5
4	线性回归模型	5
5	逻辑回归	5
6	朴素贝叶斯	5
7	决策树	5
8	支持向量机	5
9	神经网络	5
10	K 均值聚类	5
11	层次聚类	5
12	DBSCAN(密度聚类)	5
13	社区发现算法	5
14	主成分分析	5
15	MDS 算法	5
16	流形学习	5
17	t-SNE	5
	合计	85

四、课堂实训

实训 1：营业收入预测

（一）数据准备

1. 导入数据

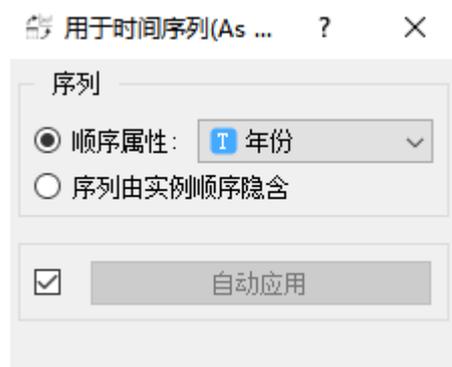
利用商业智能数据分析与挖掘平台软件进行营业收入预测，在【数据】菜单中，点击【文件】组件添加到右边区域的画布中，在安装目录下，找到“朗姿股份”（Excel）文件，导入数据。

本案例需要对数据进行时间序列分析，因此“年份”的类型为时间数据，并作为元数据；“营业收入”的类型为数值数据，作为目标变量；“序号”不具有预测特征，忽略该变量。



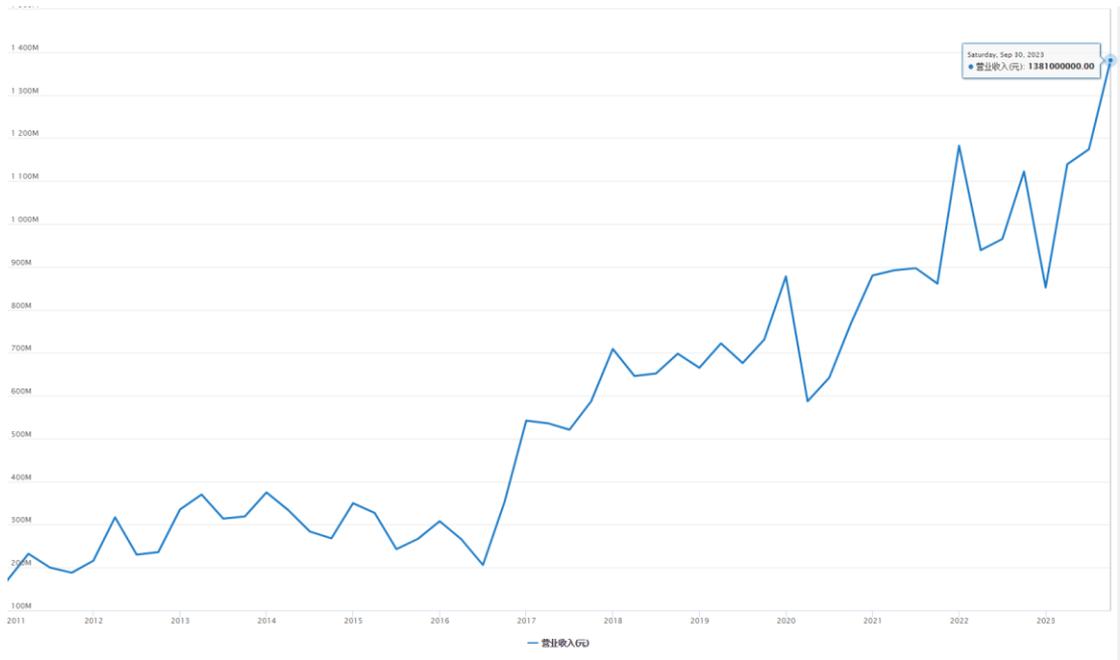
2.时间序列

选择组件【用于时间序列】，并设置顺序属性为“年份”。该操作将原数据表重新解释为 Timeseries 对象，后续才能进行时间序列分析。



3.折线图

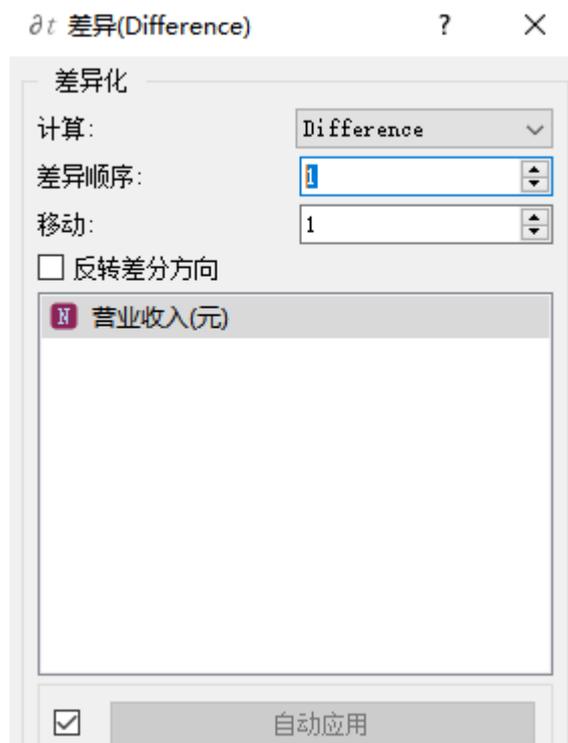
选择组件【折线图 (Line Chart)】查看源数据，判断是否为平稳数据。

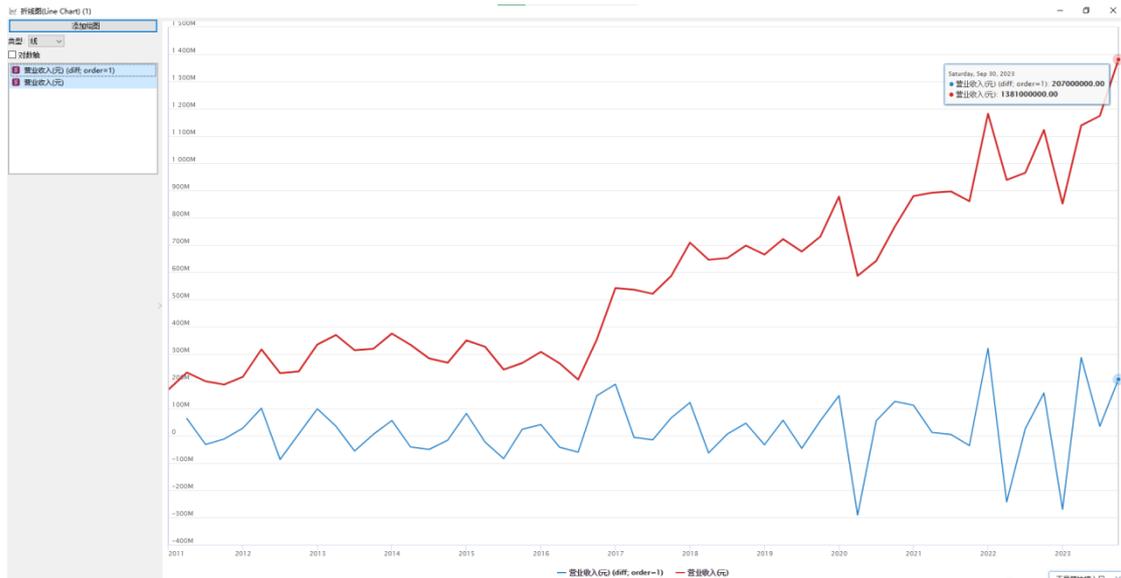


观察折线图可以得到以下结论：营业收入指标有明显的随着时间上升的趋势，序列可能是非平稳的。而 ARIMA 模型要求对平稳数据进行建模，因此，对于不平稳的序列，我们要进行差分变换消除其随机影响。

4. 差分变换

选择组件【差异】进行差分变换，并通过【折线图 (Line Chart)】组件观察进行差分变换后的数据是否平稳。



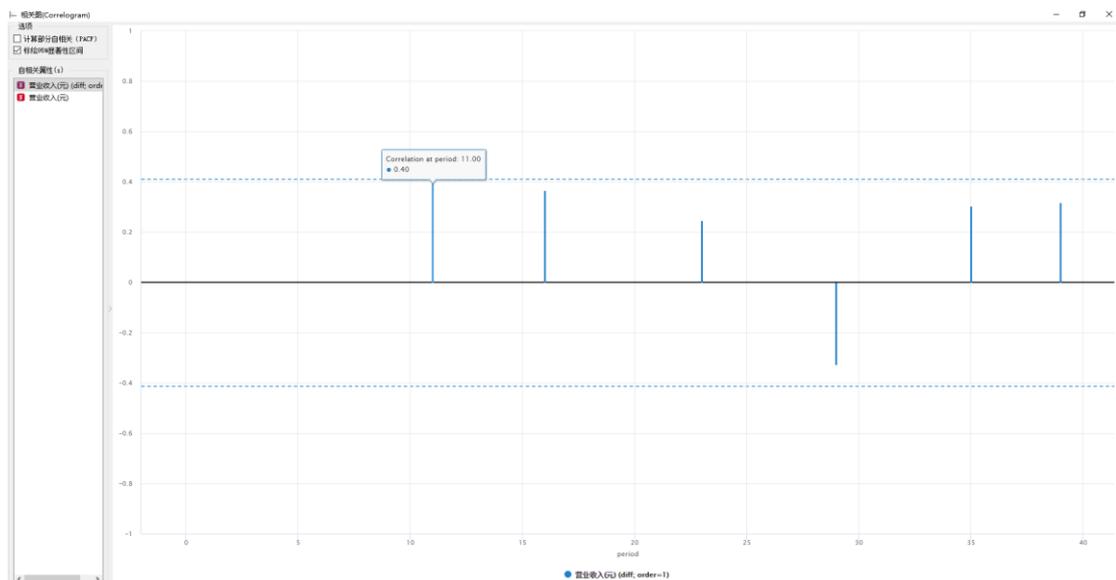


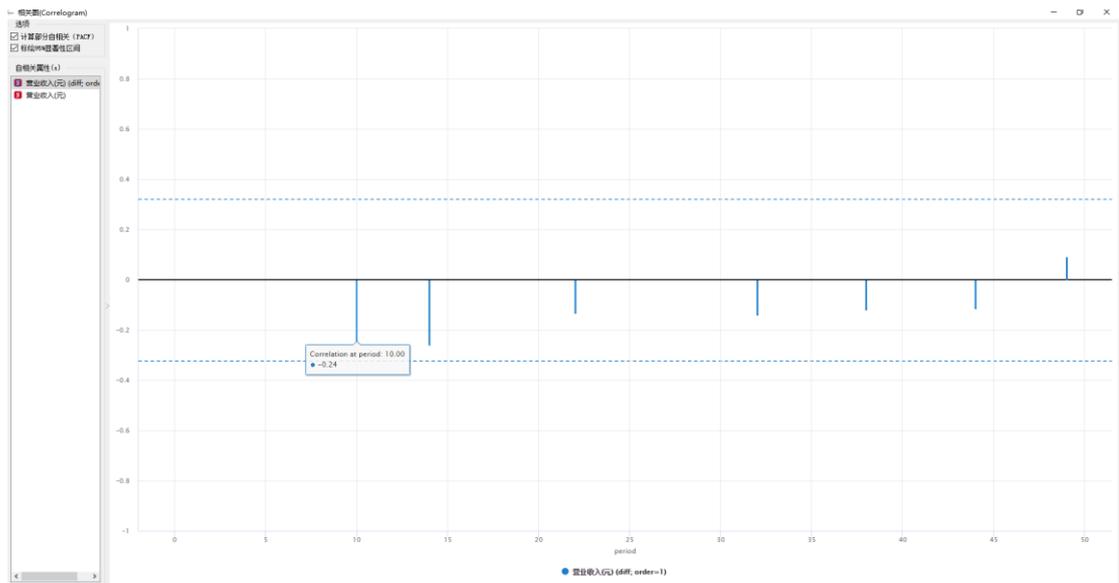
由上图可以得到以下结论：1.本次差分为的差异顺序为 1，即为一阶差分，ARIMA 模型中 $d=1$ ；2.观察折线图，原始数据与经过一阶差分的数据相比，经过一阶差分的数据已无随时间变化的明显趋势，该序列已为平稳序列。

(二) ARIMA 模型

1.相关图

选择【相关图】组件，获得经过一阶差分的营业收入 ACF 和 PACF 图，并以此得到 ARIMA 模型的最优 p 、 q 值。 p 是 PACF 图中自相关性截尾的滞后数， q 是 ACF 图中偏自相关性截尾的滞后数。





由上图可以得到以下结论：经过差分的营业收入 ACF 和 PACF 值均落在 95%的置信区间内，可以认为该序列是平稳的，故 $p=0$ ， $q=0$ 。

2.ARIMA 模型

选择组件【ARIMA 模型】对时间序列建模。

该组件支持以下的参数选择：

- (1) 名称：默认情况下，该名称派生自模型及参数；
- (2) p 、 d 、 q 值由上述步骤中获取；
- (3) 使用外部（独立）变量：若需要连接外生数据作为附加系列，则勾选该选项；
- (4) 预测未来的步骤：指希望模型预测未来多少时间点；
- (5) 置信区间：表示预测的不确定范围，常见的置信区间为 95%。

本案例选择的参数如下图所示，用户可以根据具体需求修改参数。

ARIMA模型
?
×

名称

Parameters

自回归顺序 (p):

差异度 (d):

移动平均阶数 (q):

使用外部 (独立) 变量 (ARMAX)

预测

预测未来的步骤:

置信区间:

3.折线图 (Line Chart)

选择组件【折线图 (Line Chart)】查看 ARIMA 模型的预测结果，从图中虚线部分可查看预测值，如下图所示。



4.模型评估

选择组件【模型评估】查看 ARIMA 模型的评估效果。

该组件支持以下参数选择：

(1) 折数：指在时间序列的交叉验证中，将数据集分成多少个连续的子集，每个子集轮流作为测试集，而其他子集作为训练集；

(2) 预测步骤：指模型预测未来时间单位的数量。

该组件提供以下统计量来评估模型的效果：

(1) 均方根误差 (RMSE)：提供了与原始数据相同单位的误差度量，RMSE 越小，模型的性能越好；

(2) 平均绝对误差 (MAE)：计算所有预测值与实际值之差的绝对值的平均值，MAE 对异常值的敏感度较低，是评估模型稳健性的指标；

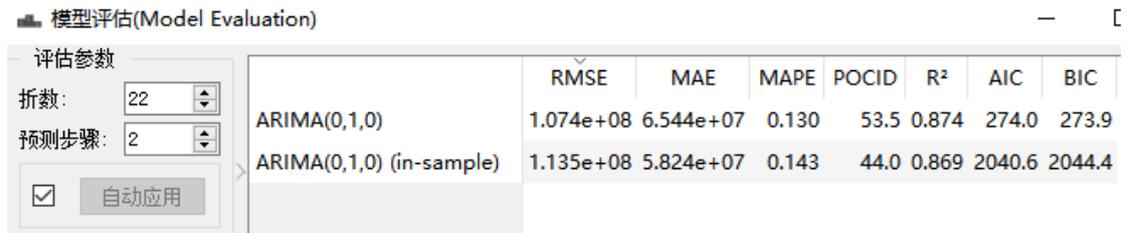
(3) 平均绝对百分比误差 (MAPE)：通过比较预测值与实际值之间的差异来评估模型的性能，MAPE 的值越小，模型的预测性能越好；

(4) 方向变化预测 (POCID)：预测目标对象在未来可能发生的方向变化；

(5) 决定系数 (R²)：表示模型捕捉到的数据可变性的比例，越接近 1，模型的解释能力越强；

(6) 赤池信息准则 (AIC)：基于似然函数的一种准则，用于衡量模型的拟合优度，并惩罚模型复杂度，AIC 越小，表示模型在保持简洁的同时拟合效果越好；

(7) 贝叶斯信息准则 (BIC)：与 AIC 类似，但对模型中参数的数量惩罚更大，BIC 越小，模型的性能越好。



模型评估(Model Evaluation)

评估参数

折数: 22

预测步骤: 2

自动应用

	RMSE	MAE	MAPE	POCID	R ²	AIC	BIC
ARIMA(0,1,0)	1.074e+08	6.544e+07	0.130	53.5	0.874	274.0	273.9
ARIMA(0,1,0) (in-sample)	1.135e+08	5.824e+07	0.143	44.0	0.869	2040.6	2044.4

(三) VAR 模型

1.VAR 模型

选择组件【VAR 模型】对时间序列建模。

该组件支持以下的参数选择：

(1) 名称：默认情况下，该名称派生自模型及参数；

(2) 参量：指模型中需要估计的参数个数，参数数量较少的模型更容易被解释；

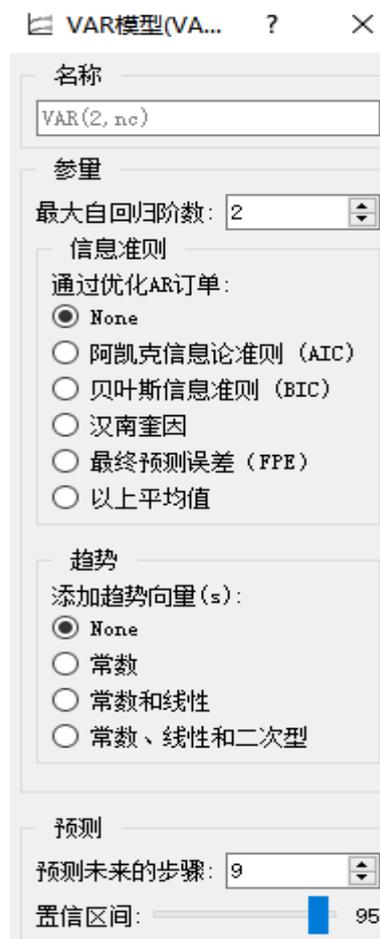
(3) 优化 AR 订单：可选择 AIC、BIC、HQIC、FPE 或以上平均值来优化模型参数的数量；

(4) 趋势：①常数：在数据集中添加一个常数列，该列的所有值都是 1，用于添加截距项；②常数和线性：添加一个常数列和从 1 开始递增的线性趋势列，用于捕捉数据随时间或顺序线性增长或减少的趋势；③常数、线性和二次型：添加一个常数列、线性列和一个二次方列，如 1, 4, 9... 等，用于捕捉数据中的非线性趋势；

(5) 预测未来的步骤：指希望模型预测未来多少时间点；

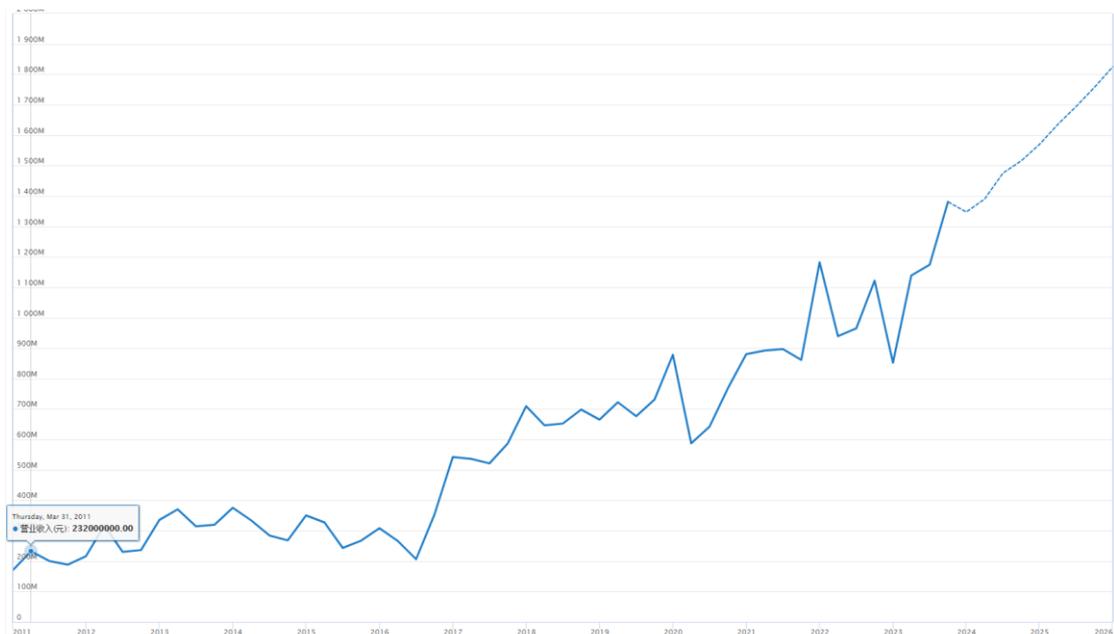
(6) 置信区间：表示预测的不确定范围，常见的置信区间为 95%。

本案例选择的参数如下图所示，用户可以根据具体需求修改参数。



2.折线图（Line Chart）

选择组件【折线图（Line Chart）】查看 VAR 模型的预测结果，从图中虚线部分可查看预测值，如下图所示。



3.模型评估

选择组件【模型评估】查看 VAR 模型的评估效果。

评估参数		RMSE	MAE	MAPE	POCID	\hat{R}^2	AIC	BIC
折数: 22	VAR(2,nc)	1.141e+08	5.411e+07	0.129	53.5	0.858	34.2	33.9
预测步骤: 2	VAR(2,nc) (in-sample)	9.987e+07	5.162e+07	0.127	44.9	0.898	39.0	39.3

(四) 模型结果分析

由 ARIMA 模型和 VAR 模型的评估结果可知，两个模型中 RMSE、MAE、MAPE、POCID、 R^2 值差异并不大，而 VAR 模型的 AIC 和 BIC 显著小于 ARIMA 模型的 AIC 和 BIC 值。由 AIC 越小，表示模型在保持简洁的同时拟合效果越好；BIC 越小，模型的性能越好可知，VAR 模型优于 ARIMA 模型。

(五) 整体流程图



实训 2：门店抽样审计

(一) 案例背景

某服装公司是一家拥有多个店铺的连锁企业，主要销售女装、男装和童装等产品。由于公司规模逐步扩大，业务范围和复杂度也不断增加，因此需要对各个店铺进行审计，以确保公司的经营活动符合法律法规和公司内部规定。但是，由于店铺数量众多，审计人员需要运用只能选择部分店铺进行审计，那么如何选择店铺的问题，为审计资源的安排提供参考意见呢？

在内部审计部门进行业务规划过程中，在确定重点审计门店时，首先要根据门店的销售额、订单数量、销售数量、店铺面积、年租金、商品数量、客户数量等维度，对门店进行全方位的分级分类，然后基于风险导向原则确定审计对象。

(二) 相关知识

1、基于多维特征值分类

由于描述门店的特征值维度较多，每个维度指标仅能刻画审计对象的某个方面，指标之

间相互配合共同刻画了审计对象的特征。指标变量选择越准确、测量越可靠，得到的分类结果就越能描述审计对象间的本质区别。但当审计对象的特征维度很多，特征的量纲差异较大时，简单的筛选和分类已无法满足全面、完整的处理需求，而机器学习算法中的聚类分析可以根据多维空间距离对审计对象进行分组。

2、聚类分析

聚类分析是把数据分类成子集或簇的过程，每一个簇中的对象相似而与其他簇中的对象相异，这些子集或簇的集合叫聚类。

聚类分析是一种无监督学习，又称为自动分类。它具有鲜明的数据挖掘功能，也就是说事先并不需要知道聚类分析后会形成的簇群个数或者层次结构，而是在分析过程中根据不同维度特征值之间的距离，通过无监督的学习和自动收敛，发现事先不能预见的聚类，从而较大程度避免了主观判断对分类的影响。

基本的聚类算法包括划分方法、层次方法、基于密度的方法和基于网络的方法。

（三）分析框架

1、审计对象画像指标选择

对服装门店审计对象的选择不仅要考虑门店的经营成果、成本及效率，还要考虑到库存及商品管理、客户数量。综合这些因素，本案例中选取了表 3-1 所示的 4 个维度 7 类指标。

表 3-1 指标清单

指标维度	指标名称	指标解释
经营成果类指标	销售额	指门店通过销售商品所获得的总收入，反映门店的销售能力和市场占有率
	订单数量	指顾客下单的次数，反映顾客购买行为的频率
	销售数量	指实际出售的商品数量，反映门店的销售情况
成本及效率类指标	店铺面积	指门店的营业面积，其大小会影响商品的展示效果和顾客的购物体验
	年租金	指门店每年支付的租金，是门店运营成本的重要组成部分
库存及商品类指标	商品数量	指门店库存中商品的总数量，体现门店库存管理水平
客户类指标	客户数量	指在一定时期内光顾门店的顾客总数，体现门店的吸引力和市场渗透率

2、分组目标

基于服装门店的多维度指标，使用数据挖掘工具将全国的机构自动分为 N 个群组，实现如下目标。

- 各群组内部各门店具有明显的相似性和群组间的相异性。
- 同一群组内的任一家门店都能作为该组的代表抽样。
- 从不同群组中各选择一家门店作为审计检查对象，能全面反映全集团的经营管理情况。
- 可以较好地平衡审计对象抽样数量和审计覆盖有效性，降低审计风险。

(四) 实训步骤

1、K 均值聚类法

(1) 导入数据

利用商业智能数据分析与挖掘平台软件进行门店聚类分析，在【数据】菜单中，点击【文件】组件添加到右边区域的画布中，在安装目录下，找到“服装门店数据”文件，选择“门店统计数据”子表。



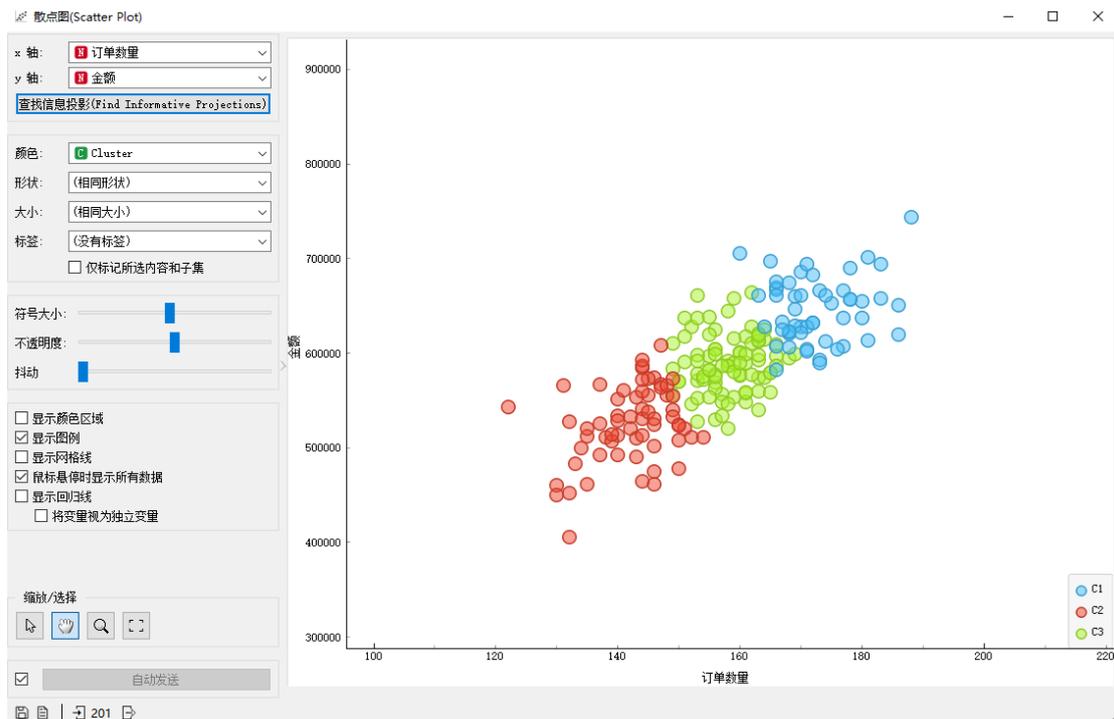
(2) k-均值聚类

选择组件【K 均值】，根据本案例将门店聚类为 3 类，设置“聚类数量”参数为 3，并设置“归一化”预处理。用户可以根据具体需求设置不同的聚类数量。



(3) 散点图

通过【散点图】，查看各聚类的分布情况，输出结果如下所示。



可从结果中提取到以下有效信息：

设置 x 轴为订单数量，y 轴为金额，从散点图得到结果：销售额 $C1 > C3 > C2$ ，订单数量 $C1 > C3 > C2$ ；且 C1 的销售额集中在 620000–680000 元，C2 的销售额集中在 510000–560000 元，C3 的销售额集中在 560000–620000 元，与 Python 操作中以“销售额”聚类得到的结果基本一致。同理，在 x 轴、y 轴中选取不同的特征变量，可查看 C1、C2、C3 的分布情况。

(4) 整体流程图



2、层次聚类法

(1) 导入数据

与“K均值聚类法”中导入数据操作相同。

(2) 距离矩阵

分类聚类要求在输入上使用距离矩阵，通过【距离】组件来计算，其中使用欧几里得距离度量。



(3) 层次聚类

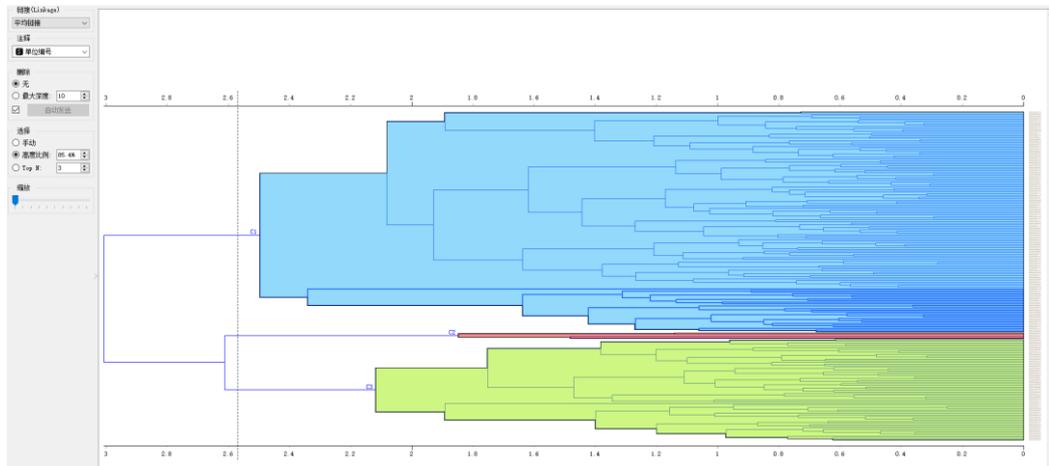
通过【层次聚类】组件从距离矩阵中计算任意类型对象的分层聚类，并显示相应的树状图。

该组件支持以下测量聚类之间距离的方法：（1）单联动计算两个聚类中最近元素之间的距离；（2）平均链接计算两个聚类的元素之间的平均距离；（3）加权联动使用WPGMA方法；（4）完全链接计算聚类最远元素之间的距离；（5）Ward联动计算误差平方和的增加，Ward的最小方差准则最小化了总的聚类内方差。本案例选取“平均

链接”进行计算，用户可以根据具体需求选取不同的聚类距离测量方法。

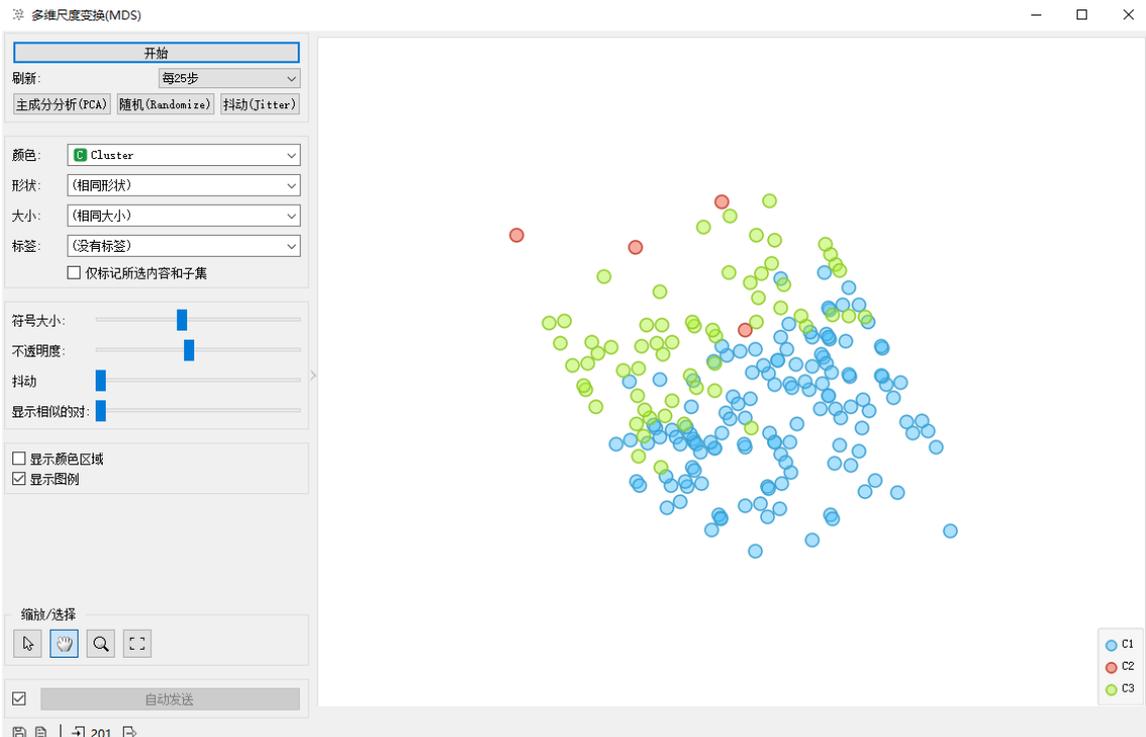
树状图中节点的标签可以在“注释”框中选择；改变树状图的最大深度，可以对树状图进行修剪，该操作只会影响显示，不会影响实际的聚类。本案例在“注释”框中选取单位编号，没有对树状图进行修剪，用户可以根据具体需求选取不同的选项。

该组件还提供了三种不同的选择方法：（1）手动：在树状图内单击将选择一个聚类，通过按住 **Ctrl/Cmd** 来选择多个集群；每个选定的聚类都以不同的颜色显示，并在输出中被视为单独的聚类；（2）高度比率：单击树状图的底部或顶部标尺可在图形中放置一条截止线，行右侧的项目属于选中状态；（3）**Top N**：选择排名靠前的节点数。本案例将门店分为三个聚类，故高度比例为 **85.4%**。用户可以根据具体需求，选取不同的高度比率得到不同的聚类数量。



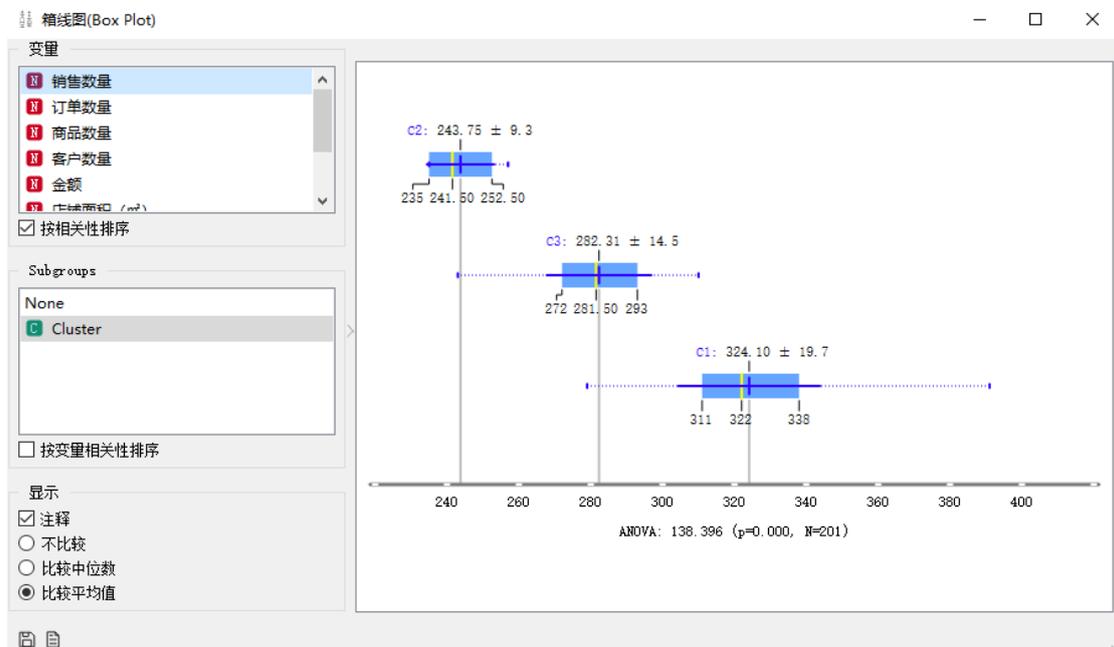
(4) 多维尺度变换（MDS）

通过【多维尺度变换】组件，将步骤 3 获得的聚类传递给 MDS，MDS 显示本案例数据的 2D 投影，并由聚类标签着色。本案例选取“主成分分析”，用户可以根据具体需求选取不同的分析方法。



(5) 查看集群特征

在【箱线图】组件中，将变量选择为本案例的各特征，可查看每个聚类的特征；使用“按相关性排序”选项，将根据变量在子组之间的区分程度对变量进行排序。



可从结果中提取到以下有效信息：

- ① “销售数量”在本案例的七个特征中，对门店聚类的区分程度最高；
- ②在变量中选取“销售数量”，可得到如下结果：各聚类中销售数量 $C1 > C3 > C2$ ；在

C1 中，销售数量的平均值为：324.10，中位数：322，标准偏差区间为 324.10 ± 19.7 ，C1 销售数量最大值接近 390，使用同样的方法对 C2 和 C3 的销售数量进行判断。选取不同的变量，能够对不同特征进行分析。

(6) 整体流程图



项目六：深度学习与应用

任务一 深度学习基础理论及发展脉络

一、任务背景

在智能制造领域，生产线的能耗（即实际耗电量预测是指充分考虑历史生产能耗、生产计划、设备运行状态、环境因素（如温湿度）以及突发性事件（如设备故障或订单调整）等因素的影响，）对未来一段时间生产线能耗进行精确预测。能耗预测是智能制造中优化资源配置与生产计划的重要组成部分

深度学习在工厂电力负荷预测中，通过分析历史用电数据，实现对未来负荷变化的高精度预测。这种技术帮助工厂优化能源调度、降低成本并提高设备和电网的运行效率。

二、工厂电力负荷案例实操

（一）、建模说明

1. 数据读取与预处理

从 Excel 文件 "工厂电力负荷.xlsx" 中读取工厂电力负荷数据。

提取数据的第 2 到第 10 行、第 2 列到最后一列作为训练和测试数据。

将数据划分为训练集和测试集：训练集的输入和目标分别是数据的前七行，测试集的输入和目标是数据的最后两行。

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	
1	分钟天数	0:00		0:15	0:30	0:45	1:00	1:15	1:30	1:45	2:00	2:15	2:30	2:45	3:00
2	第一天	2440.337682	2327.206	2316.857	2275.813	2334.824	2244.648	2238.813	2215.227	2281.386	2200.475	2205.737	2198.968	2241.21	
3	第二天	2076.55199	2006.738	1991.797	1992.045	2135.305	1925.326	1888.833	1880.229	1945.131	1872.117	1866.76	1839.701	1921.47	
4	第三天	2485.03481	2364.407	2339.75	2324.74	2353.086	2310.16	2309.768	2264.225	2337.033	2272.152	2266.434	2227.968	2309.	
5	第四天	2533.101506	2445.452	2397.743	2420.007	2451.328	2360.833	2360.252	2338.126	2423.924	2295.11	2387.999	2310.044	2394.75	
6	第五天	2517.491014	2431.264	2397.512	2380.861	2422.319	2345.824	2345.132	2317.913	2360.81	2273.923	2264.147	2285.619	2334.25	
7	第六天	2553.621895	2449.748	2409.626	2431.009	2462.321	2379.157	2354.436	2403.264	2460.24	2337.523	2333.814	2279.519	2353.21	
8	第七天	2495.445677	2425.727	2414.864	2402.415	2396.228	2346.714	2459.279	2349.681	2325.524	2317.271	2427.044	2284.43	2337.75	
9	第八天	2481.47546	2424.064	2392.248	2359.961	2432.506	2355.877	2370.181	2312.817	2278.285	2265.508	2270.102	2258.195	2310.	
10	第九天	2500.111916	2506.456	2437.924	2424.644	2485.669	2379.102	2395.006	2372.413	2417.141	2337.027	2328.824	2334.784	2318.86	
11															

图 6-1

2. 构建神经网络模型

我们希望使用前一天的电力负荷数据来预测后一天的电力负荷数据。考虑创建一个 两层神经网络模型：

第一层：包含 7 个神经元；若神经元太少，模型无法捕捉数据中的复杂模式，导致欠拟合，若神经元太多，模型可能记住噪声，从而导致过拟合现象发生，且计算成本增加。由于我们选择前七天的数据作为训练集，考虑选择第一层神经元的个数为 7 个。

使用 ReLU 激活函数，并应用 L2 正则化（权重衰减）来防止过拟合；

第二层：输出层，包含 96 个神经元，用来预测后一天的每个时间点的电力负荷，输出层的神经元个数与预测的时间步长一致（96 个时间点）。使用 线性激活函数，同样运用 L2 正则化，确保网络不会过度依赖特定的权重。线性激活函数适用于回归问题，保证输出值可以是连续的实数，而不是限制在某个范围内。

使用 Adam 优化器和均方误差（MSE）作为损失函数来编译模型，适合回归需求。

3. 模型训练与预测

使用训练集的输入和目标数据对模型进行训练，训练周期（epochs）选择范围在 250~300 之间，并且设置 verbose=0，不输出训练过程中的详细日志。

例如，设置 300 个训练周期，意味着模型将迭代 300 次所有训练数据。这个值基于问题的复杂性和训练数据量，确保模型充分学习而不过拟合。

使用训练好的模型对测试数据进行预测，并将结果展平为一维数组以便后续处理。

4. 结果可视化与误差评估

使用 `matplotlib` 绘制 预测值与真实值的对比图：横轴表示时间，纵轴表示用电量；绘制预测值和真实值的曲线，并通过图例区分两者；添加网格，便于观察曲线变化。

计算模型预测结果的误差：

RMSE（均方根误差）：衡量预测值与真实值之间的整体偏差，单位与原数据一致；MAE（平均绝对误差）：计算预测值与真实值之间的绝对误差的均值；MAPE（平均绝对百分比误差）：表示预测误差相对于真实值的百分比，通常用于直观评价模型的误差。

（二）、案例实操

根据建模说明，填写 Python 代码空缺的参数，包括第一层的激活函数、使用的优化器和想要进行的训练周期。

在技术需求转换中根据关键词填写对应参数，其中数据以数据中心下的实际数据为准。

填写完毕后，可点击需求实现查看完整代码，确认无误后点击执行并显示结果，待代码运行完毕后即可查看结果。如下所示：

```
输出：  
电力负荷回归模型评估如下所示：  
1/1 [=====] - 0s 62ms/step  
RMSE: 47.16618728637695, MAE: 33.068172454833984, MAPE: 1.254890114068985%
```

图 6-2

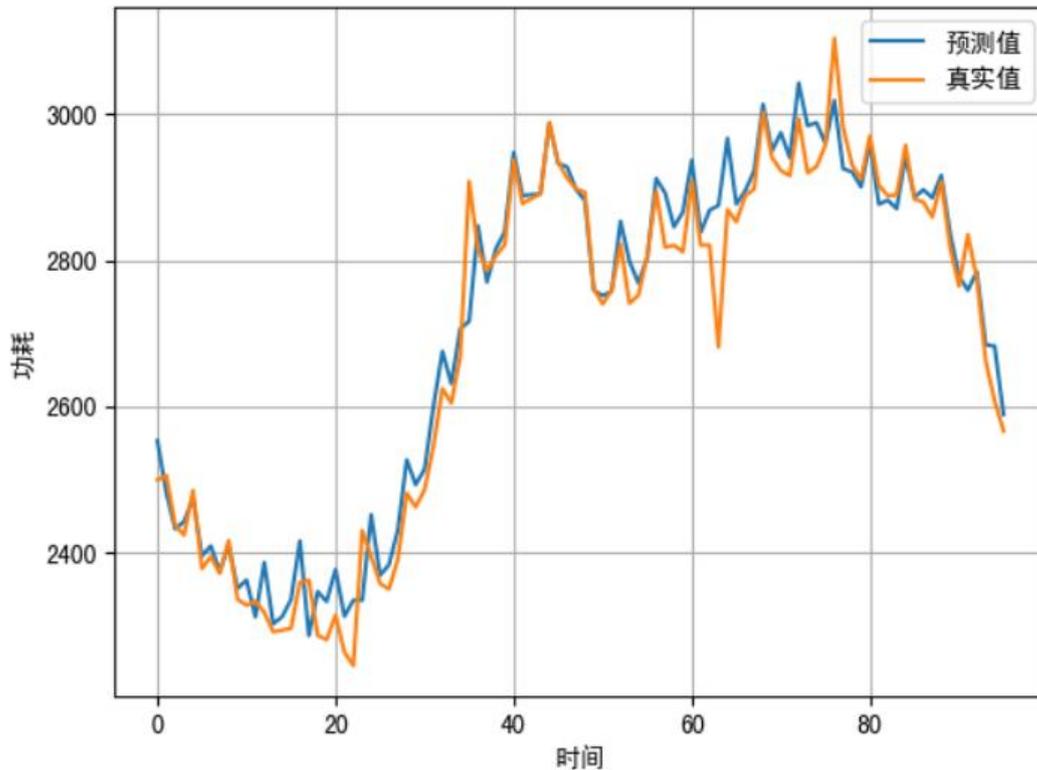


图 6-3 任务一回归结果

参考答案：

序号	关键词	参数
1	第一层激活函数	relu
2	优化器类型	adam
3	训练周期	250~300（即填写 250~300 之间的任意整数）

步骤分析：

步骤分析(单选题 3.00 分)

1. 代码中使用的深度学习框架是什么

- A. PyTorch
- B. TensorFlow
- C. Scikit-learn
- D. MXNet

正确答案： B

步骤分析（单选题 3.00 分）

2. 神经网络模型中第一层 Dense 层的激活函数是什么？

A. sigmoid

B. tanh

C. relu

D. linear

正确答案：C

步骤分析（单选题 3.00 分）

3. 可视化部分显示的是什么内容的对比？

A. 训练集和测试集的对比

B. 预测值和真实值的对比

C. 不同模型的性能对比

D. 不同时间段的功耗对比

正确答案：B

任务二 深度学习框架与工具

一、任务背景

在审计工作中，数据准确性与流程效率是核心要求，手写数字识别技术可在审计证据采集与过程管控中发挥关键作用。审计人员开展现场核查时，常会遇到被审计单位以手写表格形式记录的财务数据、业务台账等资料比如采购明细中的金额、库存台账的数量、费用报销单的数值等关键审计信息。通过手写数字识别技术，审计系统能自动提取这些手写记录的核心数据，减少人工录入带来的差错，同时缩短数据采集周期，提升审计工作效率

在审计业务流程中，手写识别结合 CNN 技术可显著优化审计数据处理的精准度与效率。借助 TensorFlow、PyTorch 等深度学习框架的支持，能够快速搭建并优化适配审计场景的识别模型。深度学习技术可利用审计领域的手写数据集进行训练，通过识别手写财务数字、业务凭证编号等字符或特定标识，验证模型在审计复杂数据识别任务中的性能，为审计工作提供可靠的技术支撑进一步强化审计过程的规范性与结果的准确性

二、手写数字识别案例实操

（一）、数据准备说明

1. 核心审计场景

1) 审计场景中存在大量含手写 / 印刷数字的凭证(如发票金额、合同编号、报销单数据)，需人工核验数字准确性，此数据集的数字识别能力可迁移至此。

2) 数据集的 28x28 灰度图像识别逻辑，可作为审计数字识别系统的基础模型训练素材，解决凭证中数字提取与校验的核心需求。

3) 训练集的“标签 - 像素”对应关系，能帮助模型学习数字的视觉特征，为审计中数字真实性校验(如篡改数字识别)提供技术支撑。

2. 具体审计应用场景

- 凭证数字自动核验：对发票、收据等凭证上的数字(金额、编号)进行图像识别，与系统记录比对，替代人工核对，提升效率。

- 篡改数字检测：基于正常数字的像素特征模型，识别凭证中被篡改的数字(如篡改金额后的像素异常)，降低审计风险。

- 批量数据校验：对海量含数字的审计凭证(如银行流水、出库单)进行批量识别与数据提取，快速匹配财务系统数据，定位差异项。

3. 技术落地路径

1) 基于现有数据集训练数字识别模型，优化模型对不同手写风格、印刷字体数字的识别准确率。

2) 扩展训练数据，加入审计场景常见的凭证数字图像(如不同字体、模糊、倾斜的数字)，提升模型适配性。

3.) 对接审计系统，实现凭证图像上传、数字识别、自动比对、异常标记的全流程自动化。

点击文件列表下载数据集。

数据包括三个文件：train.csv, test.csv, sample_submission.csv, train.csv 的部分数据如下所示：

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	pix
1	label	pixel0	pixel1	pixel2	pixel3	pixel4	pixel5	pixel6	pixel7	pixel8	pixel9	pixel10	pixel11	pixel12	pixel13	pixel14	pixel15	pixel16	pixel17	pixel18	pixel19	pixel20	pix
2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
4	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
5	4	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
8	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
9	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
10	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
11	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
12	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
13	9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
14	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
15	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
16	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
17	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
18	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
19	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
20	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
21	5	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
22	8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
23	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n	n

图 6-4

其中，`train.csv` 和 `test.csv` 包含从 0 到 9 的手绘数字的灰度图像。每张图像的高度为 28 像素，宽度为 28 像素，总共为 784 像素。每个像素都有一个与之关联的像素值，指示该像素的亮度或暗度，数字越大表示越暗。此像素值是介于 0 和 255 之间的整数（包括 0 和 255）。

训练集 `train.csv` 有 785 列。第一列是标签（`label`），表示用户绘制的数字；其余列是关联图像的像素值。训练集中的每个像素列都有一个类似于 `pixelx` 的名称，其中 `x` 是介于 0 和 783 之间的整数（包括 0 和 783）。为了在图像上找到这个像素，假设我们将 `x` 分解为 $x = i * 28 + j$ ，其中 `i` 和 `j` 是 0 和 27 之间的整数，包括 0 和 27。则 `pixelx` 位于 28 x 28 矩阵的行 `i` 和列 `j` 上（按零索引）。例如，`pixel31` 表示位于第四列第二行的像素。

从视觉上看，如果我们省略“像素”前缀，像素将组成图像，如下所示：

```
000 001 002 003 ... 026 027
028 029 030 031 ... 054 055
056 057 058 059 ... 082 083
.....
728 729 730 731 ... 754 755
756 757 758 759 ... 782 783
```

测试集 `test.csv` 与训练集相同，只是它不包含“标签”列。

（二）、数据准备实操

请根据数字准备说明完成数据准备实操。

1. 数据加载

从文件列表中读取训练集（`train.csv`）和测试集（`test.csv`），数据格式为 CSV。打印数据集的形状（`shape`）以确认样本量和特征数：

训练集包含标签列（`label`）和 784 列像素特征（对应 28×28 图像）。

测试集只有像素特征（无标签）。

2. 统计标签分布：

使用 `sns.countplot()` 绘制条形图，直观展示训练集中每个数字（0-9）的样本数量。

通过 `Y_train.value_counts()` 输出具体频数统计，使条形图按频数排序。

在技术需求转换中根据关键词填写对应参数，其中数据以数据中心下的实际数据为准。

填写完毕后，可点击需求实现查看完整代码，确认无误后点击执行并显示结果，待代码运行完毕后即可查看结果。如下所示：

```

输出：
训练集大小为：(42000, 785)
测试集大小为：(28000, 784)
    
```

图 6-5

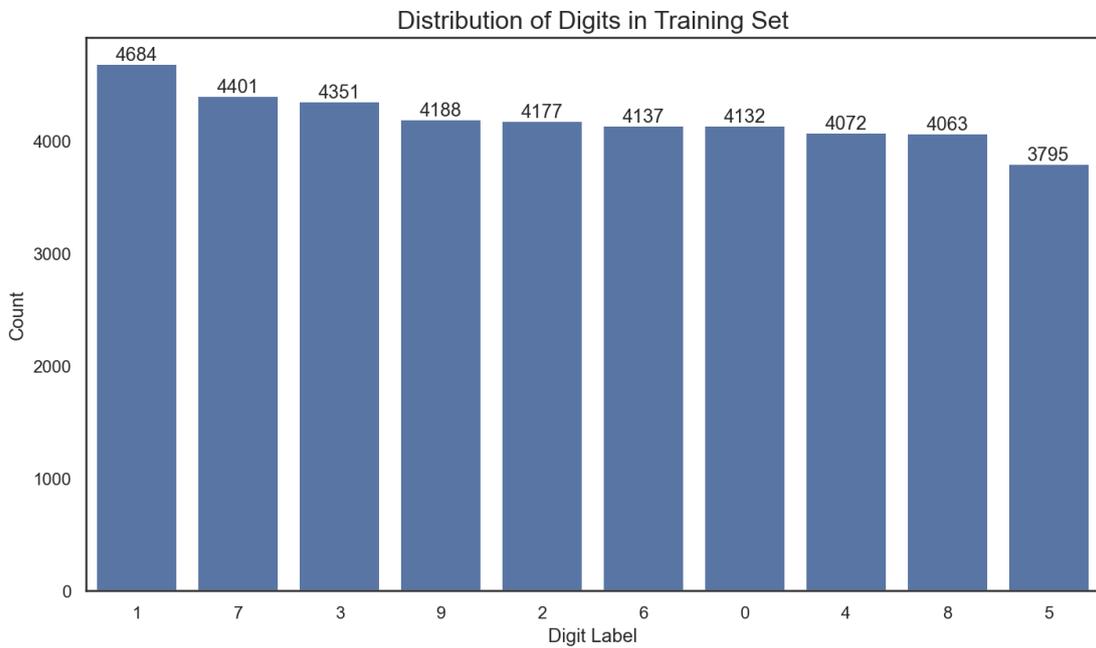


图 6-6 任务二数据准备条形图

参考答案：

序号	关键词	参数
1	打印数据形状的函数	shape
2	条形图的函数	countplot

步骤分析：

结果分析(填空题 2.00 分)

1. 训练集大小为：(1)行，(2)列（特征数+标签列），测试集大小为：(3)行，(4)列（特征数）

正确答案：

(1) 42000

(2) 785

(3) 28000

(4) 784

(三)、数据预处理说明

1. 检查空值和缺失值

`isnull()`: 检测数据中的缺失值。

对于每个元素，返回 True 表示该值为缺失值 (NaN)，否则返回 False。

`any()`: 用于检查布尔值 DataFrame 中，是否至少有一个 True 值。

True 表示该列中至少有一个缺失值，False 表示该列没有缺失值。

`describe()`: 对布尔值 Series 进行描述性统计。

主要统计如下指标：

`count`: 列的总数量。

`unique`: 唯一值的个数（通常是 True 和 False）。

`top`: 出现最多的值（如 False 表示缺失值最少）。

`freq`: top 值的出现频数。

2. 归一化

我们执行灰度归一化以减少照明差异的影响。

数据归一化将像素值从原来的范围 $[0, 255]$ 映射到 $[0, 1]$ ：`X_train=X_train/255.0` 和 `test=test/255.0`。输入数据是图像，每个像素的灰度值在 0-255 之间，归一化后所有像素值都在 $[0, 1]$ 范围内，适合神经网络的输入。此外，CNN 在 $[0, 1]$ 数据上的收敛速度比在 $[0, 255]$ 上的收敛速度更快。

3. 重塑图像

注意到训练和测试图像 ($28\text{px} \times 28\text{px}$) 是作为包含 784 个值的一维向量存入数据框的，结合上述介绍的背景知识，我们以 3 个维度重塑图像（高度=28px, 宽度=28px, 通道=1），将所有数据重塑为 $28 \times 28 \times 1$ 3D 矩阵。

`X_train.values.reshape(-1, 28, 28, 1)` 将平坦的一维数据（通常是一个 28×28 的矩阵被展平成 784 维向量）重新转换为三维形状 `[height, width, channels]`；-1 表示自动计算样本数，`[28, 28, 1]` 表示高度为 28

像素，宽度为 28 像素，通道数为 1（灰度图像）。

图像维度适配卷积神经网络(CNN)的输入要求，CNN 通常接受 3D 张量作为输入，其中通道数是一个关键维度。

这里多了个通道维度，是因为后续建模会用到 Keras。Keras 最终需要一个额外的维度来对应通道。MNIST 图像是灰度的，它只使用一个通道。对于 RGB 图像，有 3 个通道，我们会将 784px 向量重新整形为 28x28x33D 矩阵。

4. 独热编码

独热编码是一种将分类数据(categorical data)转换为机器学习或深度学习模型能够处理的数值格式的方法。它的核心思想是将每个类别表示为一个长度为 n 的向量，其中 n 是类别的总数。这个向量中只有一个位置为 1（表示当前类别），其余位置为 0。

神经网络的输出层通常是一个大小为类别数量的向量（在这里是 10），表示每个类别的预测概率。独热编码可以使标签形式与模型输出一致。

`Y_train=to_categorical(Y_train,num_classes=10)` 将标签转为独热编码，`num_classes:=10` 指定类别数（数字 0-9）。

5. 划分训练集和验证集

这里将训练集分成两部分：一小部分(10%)成为评估模型的验证集，其余部分(90%)用于训练模型。由于我们 42000 张训练图像的标签分布较为平衡，所以训练集的随机拆分不会导致某些标签在验证集中过度表示。那么，对于一些不平衡的数据集，简单的随机拆分可能会导致验证期间的评估不准确。为避免这种情况，可以在 `train test split` 函数中使用 `stratify=Y_train` 选项，可以保证验证集和训练集中的数据分类比例与 `Y_train` 一致。

6. 数据集可视化

通过可视化图像并查看标签，我们可以更好地理解这些示例之一。

`X_train[k][:,:,0]` 提取第 k 个样本的第一个通道（灰度图），参数 `cmap='gray'` 指定灰度颜色映射。

可视化图像能够直观检查数据是否正确加载和预处理，如图像是否清晰、标签是否对应等。

（四）、数据预处理实操

请根据数据预处理说明完成数据预处理实操。

步骤：数据加载→分离特征/标签→检查缺失值→归一化处理→重塑维度→独热编码标签→划分验证集→可视化验证。

首先读取训练集(train.csv)和测试集(test.csv),分离特征和标签,将像素值除以 255,归一化到[0,1]区间,将二维像素数据(784 列)转换为三维张量(28,28,1);使用 to_categorical 将数字标签(0-9)转换为独热编码(One-Hot Encoding),便于多分类任务处理;使用 train test splite 函数从训练集中随机抽取 10%作为验证集(Xval,Yval);最后随机选取 6 张训练集图像,使用 subplotst 函数以灰度图形式展示,直观检查数据是否正确加载和预处理。

输出结果：显示 6 张手写数字的灰度图像,验证数据是否正确。

在技术需求转换中根据关键词填写对应参数,其中数据以数据中心下的实际数据为准。

填写完毕后,可点击需求实现查看完整代码,确认无误后点击执行并显示结果,待代码运行完毕后即可查看结果。如下所示:

输出:

已完成对数据的独热编码以及验证集的划分。部分训练集图像可视化如下:

图 6-7

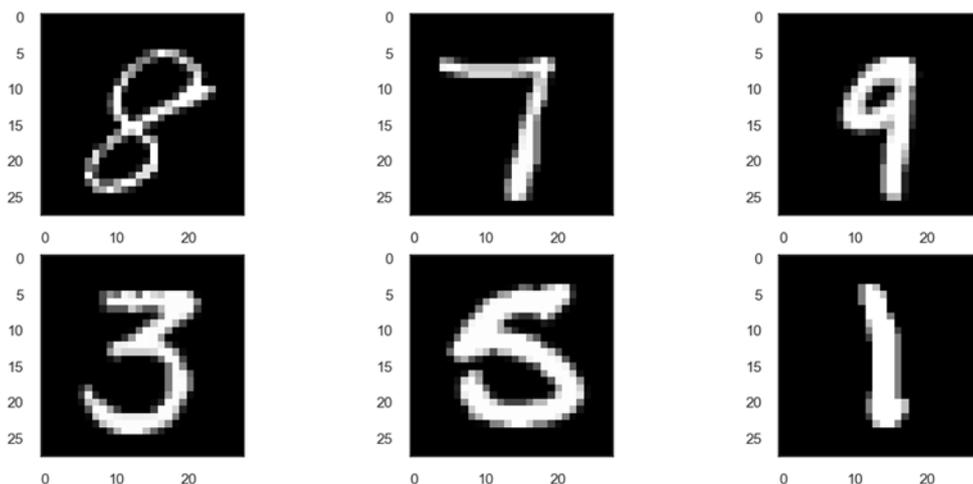


图 6-8 任务二数据预处理结果

参考答案:

序号	关键词	参数
1	打印数据形状的函数	to_categorical
2	条形图的函数	train_test_split
3	创建子图的函数	subplots

步骤分析：

步骤分析（填空题 2.00 分）

1. 从训练集中选取的验证集有(1)条数据。

正确答案：

(1)4200

（五）、模型搭建说明

1. 构建模型

根据案例背景提供信息，我们选择深度学习模型中的卷积神经网络来进行模型搭建。

卷积神经网络（Convolutional Neural Networks, CNN）是深度学习中非常常见的模型，其在图像处理中应用广泛。对于其原理和相关概念在这里就不做详细介绍了，下面主要讲介绍模型的架构。

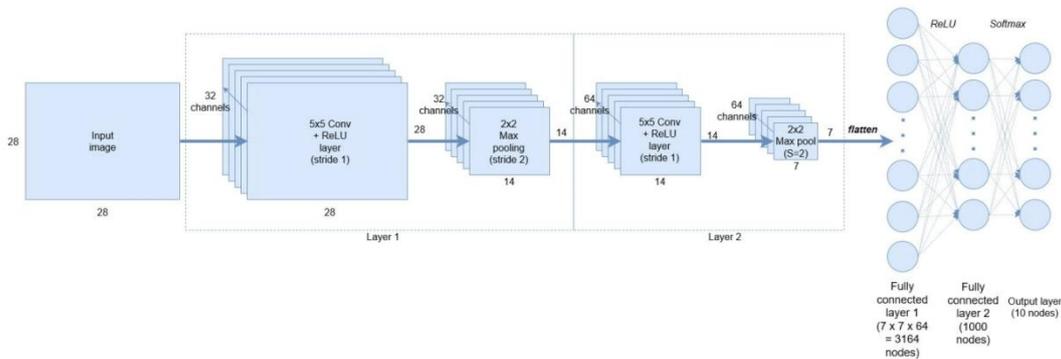


图 6-9

首先，我们输入 28*28 像素的单通道图像。

其次，通过两层卷积和池化操作，使得 CNN 能够结合局部特征并学习图像的更多全局特征。

第一层：第一个图是卷积层，使用 5*5 的 32 个卷积核（Conv2D），它好比一组可学习的过滤器，将图像进行线性变换，并通过激活函数 ReLU，将这个卷积层的线性计算系统引入非线性特征。第二个图是池化层，进行 2*2 的最大

值池化 (MaxPool2D)，它充当一个下采样过滤器，查看 2 个相邻像素并选择最大值，将原始 28*28 像素的图像缩小为 14*14 像素的图像，这能降低计算成本，并在一定程度上减少过拟合。

第二层：同样第一个图使用 5*5 的 64 个卷积核，并通过激活函数 ReLU 进行非线性变换。第二个图进行 2*2 的最大值池化 (MaxPool2D)，将 14*14 像素的图像进一步缩小为 7*7 像素的图像。

然后，通过 Flatten (展平) 将最终的特征图转换为一个一维向量，也就是第一个完全连接层中的 7764=3164 个节点，那么它结合了之前卷积层的所有找到的局部特征 3164 个。

最后，第二个完全连接层(也就相当于人工神经网络 ANN 的隐藏层)包含 1000 个节点，第一层到第二层通过 ReLU 激活函数；最后我们通过 Softmax 激活函数输出 10 个节点，也就是 10 种类别数字标签的概率分布。

注：训练模型之后，神经网络的权重与训练样本太过匹配，往往会导致在处理新样本的时候表现平平。Dropout 是一种正则化方法，对于每个训练样本，它会将层中的一部分节点随机丢弃掉(即将它们的权重 weight 设置为零)。这会迫使神经网络以分布式的方法去学习特征，从而提高模型的泛化能力并减少过拟合。

2. 设置优化器和退火器

定义好模型之后，我们需要设置优化器、损失函数和评价函数以及学习率退火器。

首先设置优化器，优化器能迭代地改进参数(过滤核值、权重和神经元的偏差等等)以最小化损失。常见的优化器有随机梯度下降(sgd)、Adam 和 RMSprop，这里我们选择 RMSprop。RMSprop 是一个十分有效的优化器，并且比随机梯度下降要快，它是一种 AdaGrad 调整方法，通过把累计梯度加权平均，试图降低其激进、单调递减的学习率。

其次，选取损失函数。损失函数用来衡量模型在具有已知标签的图像上的表现有多差，具体指的是观察到的标签和预测的标签之间的错误率。我们使用“categorical_crossentropy”的多分类损失函数，而“binary_crossentropy”一般用于二分类。

然后，选取评价函数“准确度”用于评估我们模型的性能。即题目要求的分类准确性，用 accuracy 表示。

最后，为了使优化器更快收敛并最接近损失函数的全局最小值，还需要设置学习率退火器。学习率越高，步数越大，收敛越快。然而，高学习率的采样非常差，优化器可能会陷入局部最小值。最好是在训练期间降低学习率，以有效地达到损失函数的全局最小值。为了保持高学习率的快速计算时间的优势，每 X 步（时期）动态降低学习率，这取决于是否有必要（当精度没有提高时）。为了更好的调节学习率，可以使用 keras.callbacks 中的 ReduceLROnPlateau 回调函数，它会随着训练而自动调节学习率，如果 3 个 epoch 后精度没有提高，选择将学习率减少一半。

3. 数据增强

为了避免过拟合问题，需要人为地扩展我们的手写数字数据集。以改变数组表示同时保持标签相同的方式改变训练数据的方法被称为数据增强技术。我们可以通过灰度、水平翻转、垂直翻转、随机裁剪、颜色抖动、平移、旋转等操作来进行数据的增强，使得现有的数据集更大，同时达到防止过拟合，提高泛化能力的作用。

对于数据增强，这里选择：

- (1) 将训练图像随机旋转 10 度
- (2) 将训练图像随机缩放 10%
- (3) 将图像水平随机移动 10% 的宽度
- (4) 将图像垂直随机移动 10% 的高度

注：这里没有应用 vertical_flip 或 horizontal_flip，因为水平翻转或垂直翻转会导致错误分类对称数字，例如 6 和 9。

4. 模型拟合

之后通过 fit 函数进行模型训练，使用数据增强来提升模型的泛化能力，并且在每个 epoch 后用验证集进行评估。训练过程中，使用回调函数来动态调整学习率，确保模型能够高效训练和收敛。

(六)、模型搭建实操

请根据模型搭建说明完成模型搭建实操，输出训练历史的准确率。

在技术需求转换中根据关键词填写对应参数，其中数据以数据中心下的实际数据为准。

填写完毕后，可点击需求实现查看完整代码，确认无误后点击执行并显示结果，待代码运行完毕后即可查看结果。如下所示：

```
输出：  
数据分类已完成。验证集准确率为98.7%
```

图 6-10

参考答案：

序号	关键词	参数
1	优化器类型	RMSprop
2	损失函数类型	categorical_crossentropy
3	生成随机变换图像的函数	ImageDataGenerator

步骤分析：

结果分析(判断题 2.00 分)

1. 在经过 30 个周期(epochs)拟合模型后，准确率达到 98.5%以上。

正确答案：正确

(七)、模型评估说明

1. 模型训练过程曲线

绘制损失曲线（训练集 Vs 验证集）和准确率曲线，观察模型收敛情况。

history.history 是 Keras 模型训练后返回的字典，包含每个 epoch 的指标记录；使用 plot 函数绘制折线图，横坐标是进行的 epoch 个数，纵坐标分别是损失值和准确率。

ax[0]: 显示训练损失（蓝色）和验证损失（红色），用于监控模型是否过拟合（如验证损失持续上升）。

ax[1]: 显示训练准确率（蓝色）和验证准确率（红色），用于观察模型的学习效果。

2. 混淆矩阵

混淆矩阵有助于发现模型的缺陷，自定义函数 plot_confusion_matrixix: 显示分类结果的混淆矩阵，分别以绝对值和百分比形式展示；normalize=True

时，将矩阵每行归一化为百分比，便于分析类别间的混淆比例。

混淆矩阵中对角线表示正确分类数，非对角线为误分类数。使用 `model.predict()` 函数返回每个类别的概率，`np.argmax()` 转换为具体类别，使用 `confusion_matrix()` 统计真实标签与预测标签的匹配情况。

3. 错误样本分析

提取错误预测：

找出验证集中被误分类的样本 (`Y_pred != Y_true`)，并提取它们的预测值、真实值和输入图像。计算预测概率与真实概率的差异 (`delta_pred_true_errors`)。

可视化错误样本：

显示 6 个最典型的错误案例（预测概率与真实概率差异最大）；使用变量 `delta_pred_true_errors` 衡量模型对错误类别的“过度自信”程度，`np.argsort(-delta_pred_true_errors)` 实现按差值降序排序。每张图标注预测标签和真实标签，直观分析模型弱点（如数字“4”被误判“9”）。

（八）、模型评估实操

根据模型评估说明，绘制训练和验证的损失和准确度曲线、混淆矩阵，并展示 6 个最典型的错误案例图片。

在技术需求转换中根据关键词填写对应参数，其中数据以数据中心下的实际数据为准。

填写完毕后，可点击需求实现查看完整代码，确认无误后点击执行并显示结果，待代码运行完毕后即可查看结果。如下所示：

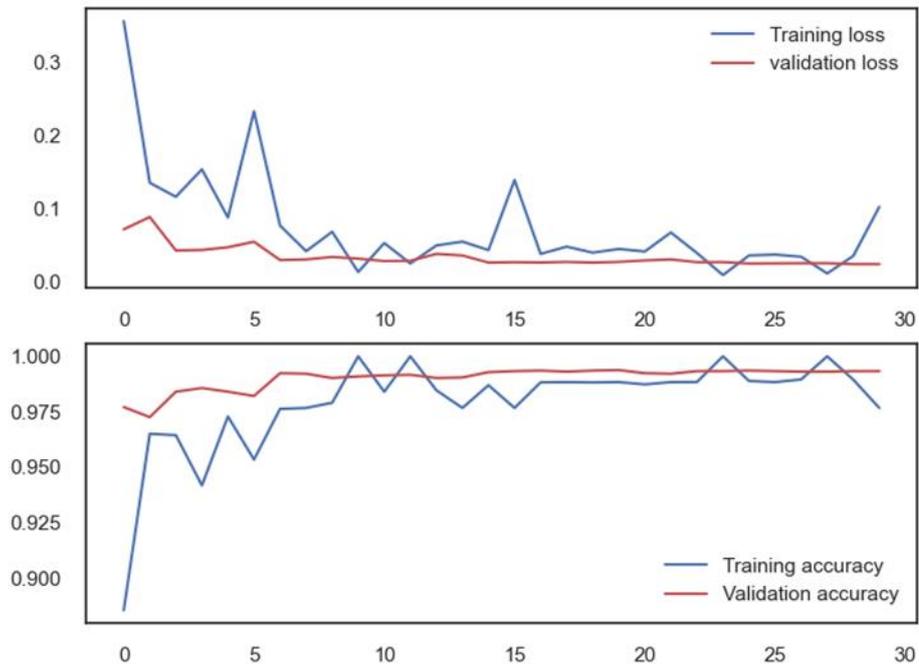


图 6-11 任务二模型评估曲线

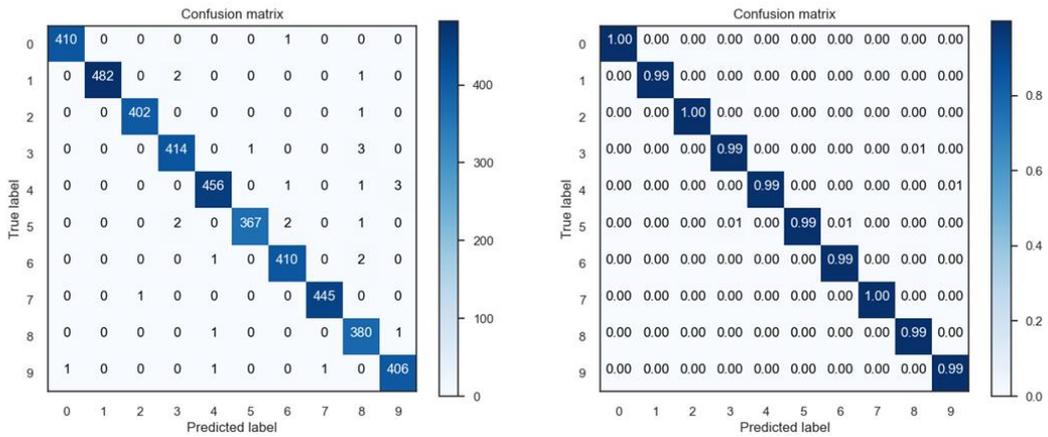


图 6-12 任务二模型评估混淆矩阵

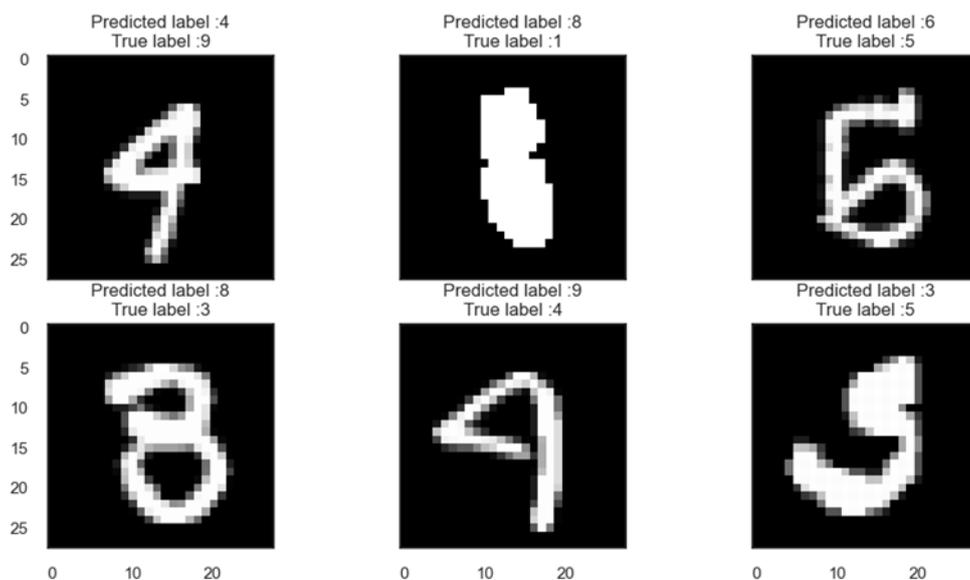


图 6-13 任务二模型评估失败图片

参考答案:

序号	关键词	参数
1	Keras 模型训练后返回的字典	history.history
2	混淆矩阵的函数	confusion_matrix

步骤分析:

结果分析 (填空题 2.00 分)

1. 通过混淆矩阵图结果, 可知各类别的准确率均在(1)%以上。其中, 数字三被误判成什么数字的概率最大: (2)。

正确答案:

(1)99

(2)8

结果分析

(填空题 2.00 分)

2. 在 2 个 epoch 后, 该模型在验证数据集上的准确率几乎达到(1)+%。在训练过程中, 发现验证准确度几乎每次都(2)(高于/低于)训练准确度, 也就是说不管怎样, 模型都不会过度拟合训练集。

正确答案:

(1)99

(2) 高于

任务三 深度学习模型架构

一、任务背景

手写数字识别的技术可以直接迁移到零件检测质量监测、识别等场景中。CNN 的特性使其特别适合处理工业图像数据，例如零件外观图像、表面纹理图像和复杂的几何形状。

在汽车制造中，零件的质量直接影响到整车的性能和安全性。随着智能制造和工业 4.0 的发展，传统的人工检测方法已难以满足现代汽车工业对精度、速度和成本控制的要求。基于卷积神经网络 (CNN) 的智能检测系统可以在生产线上实现零件的自动化检测和分类，有效提高效率 and 准确性。

二、零件检验案例实操

(一)、案例导入数据集数据库操作说明

1. 导入 Path 类 python

Path 类用于简化路径操作，帮助检查并管理存储零件检测数据（如训练数据、测试数据）的文件夹路径。

2. 定义数据存储路径

现有 test 和 train 两个文件夹，每个文件夹里有 def_front 和 ok_front 两个子文件夹，里面存放着用来训练模型的图像数据。部分图像数据如下所示：

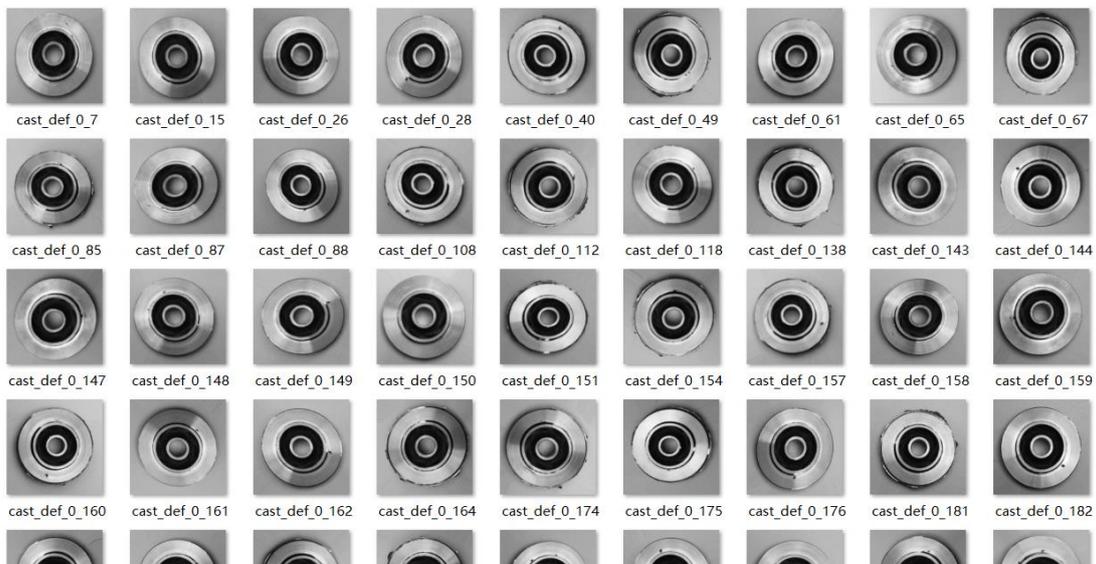


图 6-14

3. 验证路径是否存在

使用 `exists()` 函数检查数据目录是否存在。对于零件检测模型，数据的正确路径非常重要，因为模型训练和验证需要加载存储在该路径下的图像。

4. 列出目录内容

在零件质量检测任务中，这一步可以确认数据目录的组织结构是否正确。例如，查看是否存在 `train` 和 `test` 子目录，以及这些子目录中是否有图像分类文件夹（如 `ok_front` 和 `def_front`）。

（二）、案例导入数据集数据库操作

从 `pathlib` 模块导入 `Path` 类，用于处理文件系统路径。

使用 `exists` 函数检查路径是否存在：若打印 `True`，表示路径存在。

使用函数 `iterdir` 遍历路径内容：如果路径存在且是目录，`iterdir()` 会返回该目录下所有文件和子目录的生成器，逐个打印它们的 `Path` 对象。

在技术需求转换中根据关键词填写对应参数，其中数据以数据中心下的实际数据为准。

填写完毕后，可点击需求实现查看完整代码，确认无误后点击执行并显示结果，待代码运行完毕后即可查看结果。如下所示：

```
输出：  
True  
[DataFolderPath]/casting_data/casting_data
```

图 6-15

参考答案：

序号	关键词	参数
1	检查路径是否存在的函数	<code>exist</code>
2	遍历路径内容的函数	<code>iterdir</code>

步骤分析：

结果分析(填空题 2.00 分)

1. 输出结果为：(1)；由结果可知路径(2) (存在不存在)。

正确答案：

(1)True

(2) 存在

(三)、案例数据预处理操作说明

1. 数据增强与数据分割

导入可视化库 (matplotlib, seaborn) 和图像数据生成器 (ImageDataGenerator); ImageDataGenerator: 用于实时数据增强 (旋转、平移、翻转等) 和批量加载图像。从数据集 casting_data 中提取一批图像和对应标签, 通过旋转、翻转、亮度调整等操作扩充训练数据。

定义增强型生成器 (train_generator) 和基础生成器 (plain_generator), 基础生成器仅对图像进行归一化, 不进行增强 (用于对比可视化)。增强型生成器的参数如下所示:

```
rotation_range = 360:
```

随机旋转图像最多 360 度, 能够让模型在训练过程中接触到各种角度的图像, 增强模型的鲁棒性, 尤其是在没有明确方向性的任务中 (如检测是否是正面或反面图像)。360 度旋转可以有效模拟不同的图像取向, 避免模型仅对某个方向的特定样本进行过拟合。

```
width_shift_range = 0.05 和 height_shift_range = 0.05:
```

这两个参数控制水平和垂直的随机平移, 平移范围设置为图像宽度和高度的 5% (0.05)。平移变换模拟了图像的拍摄位置变化, 可以帮助模型适应不同位置的对象, 增加对输入图像的变动性理解。

```
shear_range = 0.05:
```

剪切变换 (shearing) 对图像的形状进行拉伸或压缩, 使得图像的形态发生变形, 模拟了拍摄角度的变化。这个变化增强了模型对形态变异的适应能力。

```
zoom_range = 0.05:
```

随机缩放图像, 范围为 5% (0.05), 通过放大或缩小图像来模拟不同的物体尺度。这能增加模型对目标大小变化的适应性。

```
horizontal_flip = True 和 vertical_flip = True:
```

水平和垂直翻转图像。这能有效地增加样本的多样性, 尤其是当物体或图像内容不依赖于特定方向时 (例如, 正面或反面的图像), 模型不容易过拟合于特定的方向。

```
brightness_range = [0.75, 1.25]:
```

随机调整图像的亮度范围，范围为 75%到 125%。这样可以模拟不同光照条件下拍摄的图像，增强模型在不同亮度条件下的表现。

```
rescale = 1./255:
```

归一化处理，将像素值从 0-255 缩放到 0-1，使模型的输入范围一致，防止由于大幅度的数值差异导致梯度计算问题。归一化通常是深度学习中预处理数据的标准步骤。

```
validation_split = 0.2:
```

将数据的 20%用作验证集，20%是常见的验证集比例。这样做的目的是在训练过程中评估模型的性能，防止模型过拟合并可以及时调整参数。

数据配置生成器(gen_args)参数如下所示：

```
target_size = IMAGE_SIZE: 图像大小调整为 (300, 300);
```

```
color_mode = "grayscale":
```

选择灰度图像模式，说明输入图像是单通道的（即黑白图像），因为图像数据本身就是灰度图。

```
batch_size = BATCH_SIZE:
```

批量大小设置为 64。批量大小控制每次梯度更新时使用的样本数。64 是一个常见的选择，既可以保证计算效率，又不会占用过多的内存。较大的批量可能会加速训练，但也可能导致训练不稳定；较小的批量则可以提供更多的噪声，帮助模型在训练过程中更好地泛化。

```
class_mode = "binary":
```

任务是二分类问题（例如，ok_front 和 def_front），因此 class_mode 设置为 binary，表示每个样本属于两个类别之一。

```
classes = {"ok_front": 0, "def_front": 1}:
```

指定类标签和它们对应的类别编号。这是对二分类任务的明确规定，将“ok_front”映射到 0，“def_front”映射到 1。这样可以确保模型的输出与实际类别一致。

```
seed = SEED_NUMBER:
```

设置随机种子（SEED_NUMBER = 123），用于保证数据处理中的随机操作（如

数据增强和批量分配)的可重复性。这有助于确保每次运行训练时使用相同的训练数据分配和增强配置，便于调试和实验结果的可复现性。

2. 创建可视化函数

创建用于可视化的图像网格：创建一个 8×8 的网格，用于显示当前批次的 64 张图像，每个子图对应一个零件图像。分别展示增强后和未增强的训练样本，每张图像下方标注类别（ok 或 defect）。

绘制每张图像及其对应标签：遍历网格子图，为每个子图显示图像内容和对应类别标签。

优化网格布局并添加标题：调整图像的网格布局，使子图之间不重叠。为整个可视化添加一个标题（由 `title` 参数指定），例如 "训练数据 - 合格与缺陷批次"，显示生成的图像网格。

（四）、案例数据预处理操作

根据案例数据预处理操作说明，完成代码的编写。

在技术需求转换中根据关键词填写对应参数，其中数据以数据中心下的实际数据为准。

填写完毕后，可点击需求实现查看完整代码，确认无误后点击执行并显示结果，待代码运行完毕后即可查看结果。如下所示：

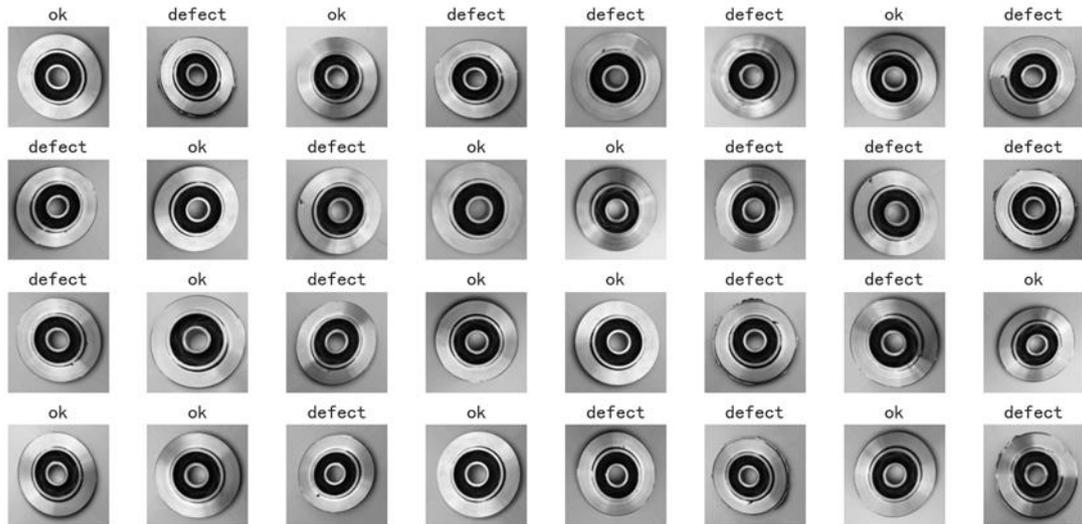


图 6-16 任务三数据预处理未增强的结果

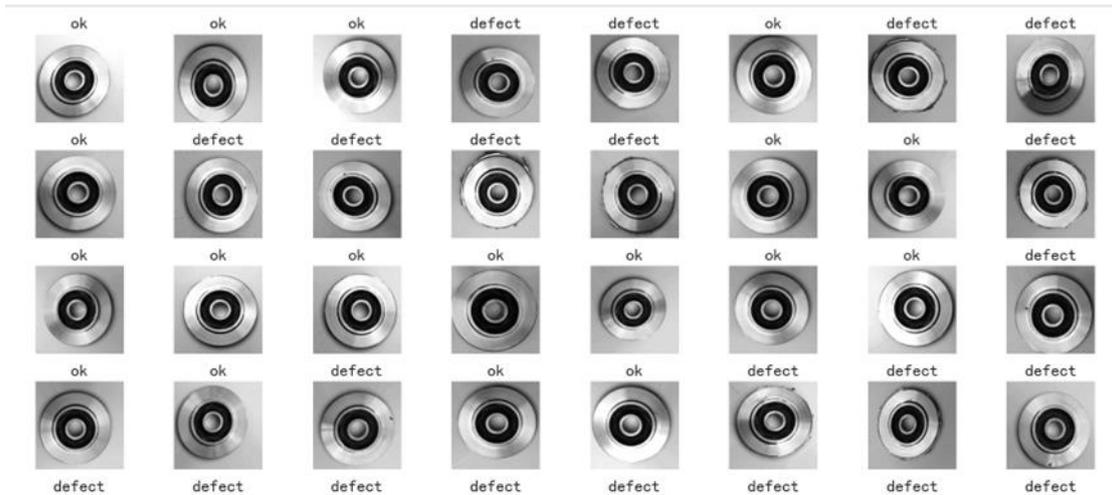


图 6-17 任务三数据预处理增强后的结果

参考答案：

序号	关键词	参数
1	图像数据生成器	ImageDataGenerator
2	0 映射的字符串	ok
3	1 映射的字符串	defect

步骤分析：

步骤分析(判断题 2.00 分)

1. 归一化处理中，像素值从 0~255 缩放到 0~1。

正确答案：正确

(五)、案例模型设计与训练操作说明

设计、训练并保存一个卷积神经网络（CNN）模型，用于二分类任务（如铸造缺陷检测）。

1. 模型定义（Sequential）

使用函数 `model_cnn = Sequential([...])`，构建一个顺序堆叠的 CNN 模型，包含卷积层、池化层、全连接层和正则化层。

模型架构细节：

1) 第一卷积层

卷积操作：提取图像的局部特征，如边缘、纹理等。

参数：

`filters = 32`：使用 32 个卷积核，学习 32 个不同的特征。

`kernel_size = 3`: 卷积核大小为 3×3 。

`strides = 2`: 步幅为 2, 特征图大小将减半。

`activation = "relu"`: 使用 ReLU 激活函数, 增强非线性建模能力。

`input_shape = IMAGE_SIZE + (1,)`: 输入图像的大小为 `IMAGE_SIZE`, (1,) 表示单通道灰度图。

2) 第一池化层

最大池化: 从 2×2 的窗口中选取最大值, 减小特征图的大小。

参数:

`pool_size = 2`: 池化窗口大小为 2×2 。

`strides = 2`: 步幅为 2, 特征图大小减半。

3) 第二卷积层

在第一卷积层的基础上, 进一步学习更高级的特征。

参数:

`filters = 16`: 使用 16 个卷积核, 学习更复杂的模式。

`kernel_size = 3`: 卷积核大小为 3×3 。

`strides = 2`: 步幅为 2, 特征图大小继续减半。

`activation = "relu"`: 使用 ReLU 激活函数。

4) 第二池化层

再次减小特征图大小, 提取更抽象的特征。与第一池化层类似, 进一步压缩数据表示。

5) 扁平化层

将多维的特征图展平为一维向量, 便于输入到全连接层。

6) 全连接层 1

全连接操作: 将展平的特征输入到 128 个神经元中, 学习更复杂的模式。

Dropout: 以 20% 的概率随机“丢弃”部分神经元, 防止过拟合。值可以取 $0 \sim 1$, 一般在 $0.2 \sim 0.5$; 小于 0.2 可能效果提升有限, 需要尝试微调, 大于 0.5 可能导致欠拟合。Dropout(`rate = 0.2`), 表示每个输入单元有 20% 的概率被随机置零, 剩余 80% 的单元会保留并乘以 $1/(1-\text{rate})=1.25$ 以保持总激活值的期望不变, 称为逆缩放。

7) 全连接层 2

全连接操作：再加入一个隐藏层，包含 64 个神经元。

Dropout：进一步正则化，以降低过拟合风险。

8) 输出层

全连接操作：输出单个神经元，用于二分类（0 或 1）。

激活函数：使用 sigmoid，将输出限制在 $[0, 1]$ 范围内，表示某类别的概率。

2. 模型编译

定义优化器：

`optimizer='adam'`：使用 Adam 优化器，结合了动量梯度下降和自适应学习率，适合复杂神经网络训练。

定义损失函数：

`loss='binary_crossentropy'`：使用二元交叉熵损失，适用于二分类任务。

定义评价指标：

`metrics=['accuracy']`：使用准确率作为模型性能的评价指标。

3. 模型训练

设置检查点

定义检查点为在训练过程中保存模型的最佳权重文件。

参数：

`'model/cnn_casting_inspection_model.hdf5'`：保存模型的路径和文件名。

`save_best_only=True`：仅保存验证集损失最小的模型。

`monitor='val_loss'`：监控验证集损失值。

`mode='min'`：以损失值的最小化为目标。

`verbose=1`：显示保存操作的日志。

训练模型

利用训练数据集和验证数据集，逐轮优化模型参数。

参数：

`train_dataset`：训练数据集，用于更新模型参数。

`validation_dataset`：验证数据集，用于评估模型的性能。

batch_size=16: 每次训练使用 16 个样本的小批量数据, 平衡训练速度与内存占用。

epochs=15: 训练 15 个周期, 每个周期完整遍历一次训练数据集。

callbacks=[checkpoint]: 指定回调函数, 保存最佳模型权重。

verbose=1: 显示训练过程中的详细信息。

(六)、案例模型设计与训练操作

请根据案例模型设计与训练操作说明完成模型设计与训练操作。

在技术需求转换中根据关键词填写对应参数, 其中数据以数据中心下的实际数据为准。

填写完毕后, 可点击需求实现查看完整代码, 确认无误后点击执行并显示结果, 待代码运行完毕后即可查看结果。如下所示:

```
输出:  
CNN模型已构建。损失率和准确率文件已保存: model/cnn_casting_inspection_model.keras
```

图 6-18

参考答案:

序号	关键词	参数
1	卷积层的激活函数	relu
2	输出层的激活函数	sigmoid
3	Dropout 取值	0.2~0.5

步骤分析:

结果分析 (填空题 2.00 分)

1. Conv2D 层: conv2d10(Conv2D) Output Shape: (None, 149, 149, 32) 输入的图片数据通过卷积操作得到输出。这里的 (None, 149, 149, 32) 表示: 图片经过卷积核处理后, 尺寸缩小为 (1) × (2)。输出的通道数 (即卷积核的数量), 说明这一层有 (3) 个卷积核, 提取了 (4) 种不同的特征。

正确答案:

(1) 149

(2) 149

(3) 32

(4) 32

结果分析 (填空题 2.00 分)

2. MaxPooling2D 层: `max_pooling_2d_10` (MaxPooling2D)

Output Shape: (None, 74, 74, 32) 最大池化层的作用是减小图像的空间尺寸 (宽度和高度), 同时保留最显著的特征。这里将 149x149 的图像尺寸池化到 (1) × (2)。32 表示池化操作后, 卷积操作的 (3) 个特征通道保持不变。

正确答案:

(1) 74

(2) 32

(3) 32

结果分析 (填空题 2.00 分)

3. Flatten 层: `flatten5` (Flatten) Output Shape: (None, 5184) Flatten 层将多维的输入展平成一维数组。这里将形状为 (18, 18, 16) 的输出展平为一个长度为 (1) 的一维向量。

正确答案:

(1) 5184

结果分析 (填空题 2.00 分)

4. Dense 层: `dense235` (Dense) Output Shape: (None, 1) 最后一层全连接层, 输出 (1) 个神经元, 通常用于二分类问题 (如输出概率值或类别标签)。

正确答案:

(1) 1

(七)、案例模型检验操作说明

1. 创建图表

用 `plt.subplots` 来创建一个图表对象并定义大小 (宽度 8 英寸, 高度 6 英寸)。

2. 导入历史数据

用 `model_cnn.history.history` 记录训练过程中的指标数据, 包含训练和验证的损失、准确率四个键值对。`model_cnn.history.epoch` 是训练的轮次列表, 长度等于训练的 epoch 数量。使用 Pandas DataFrame 将数据格式化, 行索引

设为 1, 2, ... (从第 1 个 epoch 开始)。Seaborn 的 lineplot 自动为 DataFrame 中的每列绘制训练和验证的损失、准确率四条曲线

3. 设置图标格式

设置图表的标题为 "TRAINING EVALUATION", 并且设置标题的字体加粗 (fontweight = "bold") 和字体大小为 20 (fontsize = 20)。x 轴标签为 "Epochs", y 轴标签为 "Metrics"。

4. 生成图表并分析

设置图例标题为 "Metrics"。每一条曲线 (例如损失曲线、准确率曲线) 会被自动标记, 并且通过图例进行区分。根据图表分析的得出结论。

(八)、案例模型检验操作

请根据案例模型检验操作说明完成模型检验操作。

在技术需求转换中根据关键词填写对应参数, 其中数据以数据中心下的实际数据为准。

填写完毕后, 可点击需求实现查看完整代码, 确认无误后点击执行并显示结果, 待代码运行完毕后即可查看结果。如下所示:

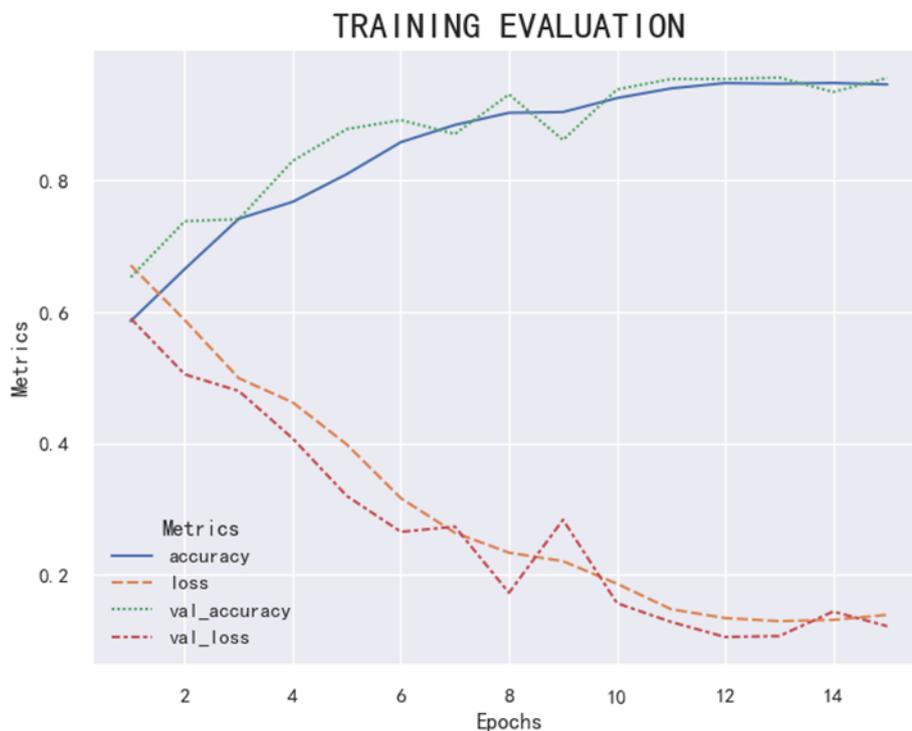


图 6-19 任务三模型检验结果

参考答案:

序号	关键词	参数
1	Seaborn 中折线图的函数	lineplot
2	x 轴标签	Epochs

步骤分析：

结果分析(填空题 2.00 分)

1. 通过观察曲线的训练结果我们可以得到：随着训练轮次的增加，损失逐渐(1)(减小/增大)，随着训练进行，模型的准确率逐步(2)(上升/下降)。验证损失与训练损失，两者趋势是(3)(相同/不同)的。验证准确率与训练准确率，两者趋势是(4)(相同/不同)的。

正确答案：

(1)减小

(2)上升

(3)相同

(4)相同

(九)、案例模型测试与评估操作说明

1. 加载训练好的模型

从指定路径加载已训练好的模型（文件格式为 `.keras`），其中包含模型的架构和参数权重。

确保测试时使用的模型与训练时保存的模型一致。

2. 使用模型进行预测

利用加载的模型对测试数据集（`test_dataset`）进行预测，获取每个样本属于目标类别的概率值。

3. 将预测概率转换为分类结果

定义分类阈值 `THRESHOLD = 0.5`。

将预测的概率值（`y_pred_prob`）转换为类别标签（0 或 1）。若概率值大于或等于阈值，则预测为正类（1）；否则为负类（0）。调用 `.reshape(-1,)` 确保结果是一维数组。

4. 获取真实类别的标签

提取测试数据集中每个样本的真实类别标签。

`test_dataset.classes` 是数据集中每个样本的真实标签数组，`test_dataset.index_array` 是当前样本的索引，用于获取与预测结果对应的真实标签。

5. 生成混淆矩阵

调用 `confusion_matrix` 生成混淆矩阵，表示预测结果与真实标签的对应情况。

使用 Pandas DataFrame 格式化混淆矩阵，添加多级标签（Actual 和 Predicted，以及类别名称 ok 和 defect），使表格更易读。

6. 输出分类结果

使用 `classification_report` 生成分类性能的详细报告，包含：

Precision（精确率）：预测为正类的样本中实际正确的比例。

Recall（召回率）：实际为正类的样本中被正确预测的比例。

F1-Score：精确率和召回率的综合评估指标。

Support：每个类别中样本的数量。

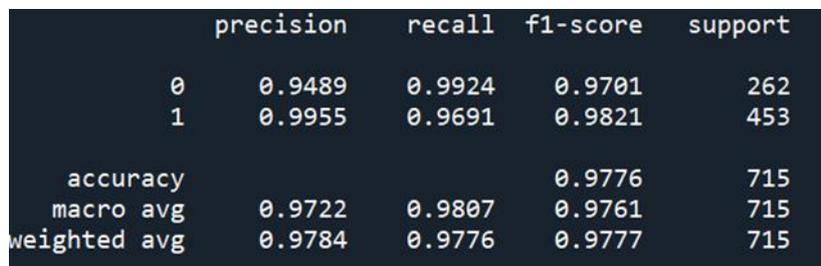
参数 `digits=4` 指定报告中的数值保留 4 位小数。

（十）、案例模型测试与评估

请根据案例模型测试与评估操作说明完成模型测试与评估操作。

在技术需求转换中根据关键词填写对应参数，其中数据以数据中心下的实际数据为准。

填写完毕后，可点击需求实现查看完整代码，确认无误后点击执行并显示结果，待代码运行完毕后即可查看结果。如下所示：



```
              precision    recall  f1-score   support

 0         0.9489      0.9924      0.9701      262
 1         0.9955      0.9691      0.9821      453

 accuracy          0.9776      715
 macro avg         0.9722      0.9807      0.9761      715
 weighted avg      0.9784      0.9776      0.9777      715
```

图 6-20 任务三模型测试与评估结果

参考答案：

序号	关键词	参数
1	分类阈值	0.5

2	生成分类性能的详细报告函数	classification_report
---	---------------	-----------------------

步骤分析:

结果分析(判断题 2.00 分)

1. 从图中可知, 所有预测正确的样本占总样本的比例为 97.76%, 说明模型在整体上表现良好。

正确答案: 正确

项目七：深度学习技术在审计研究中的应用

实训 1：BP 神经网络-审计风险预测模型

任务 1：采集医药财务数据——步骤 2 采集医药行业数据

本案例主要从平台仿真网站上采集数据，相关代码如下：

```
import pandas as pd
import re
import requests
```

- (1) pandas 用于数据处理和 Excel 文件的读写。
- (2) re 用于正则表达式匹配，这里用于从网页 HTML 中提取数据。
- (3) requests 用于发送 HTTP 请求，获取网页内容。

```
#读取文件
filename = '[DataFolderPath]/医药采集需求.xlsx'
zb_dc = {'主要指标': 'zyzb', '资产负债表': 'zcfzb', '利润表': 'lrb', '现金流量表': 'xjllb'}
df_dc = {k: {'行业': [], '股票': []} for k in zb_dc.keys()}
df_nor = pd.read_excel(filename, sheet_name='未舞弊', dtype=str,
usecols=['股票代码', '证券名称', '所属行业'])
df_wrong = pd.read_excel(filename, sheet_name='舞弊', dtype=str,
usecols=['股票代码', '证券名称', '所属行业'])
df_all = pd.concat([df_nor, df_wrong], axis=0, ignore_index=True)
df_all.drop_duplicates(inplace=True)
```

- (1) 从指定的 Excel 文件中读取两个表（未舞弊和舞弊），只选取股票代码、证券名称和所属行业三列。
- (2) 将两个 DataFrame 合并，并去除重复行。

```
#网址拼接并请求数据
for n, row in df_all.iterrows():
    url = start_url.split('index')[0] + f'details/detail_{row["股票代码"].split(".")[0]}.html'
    print(url)
    html = requests.get(url)
    charset = html.headers.get("Content-Type", '').split(';')[-1]
    html.encoding = charset.split('=')[-1] if 'charset' in charset else 'utf-8'
```

- (1) 遍历 df_all 中的每一行，根据股票代码拼接出详细的网页 URL。
- (2) 使用 requests.get 发送 HTTP 请求获取网页内容。

(3) 根据响应头中的 Content-Type 设置正确的编码，确保能够正确读取网页内容。

```
#正则表达查找网页元素，返回财务数据
for k, v in zb_dc.items():
    tbody = re.search(f'<tbody id="{v}Table">(.*?)</tbody>', html.text,
re.S)
    if not tbody:
        continue
    dtr = re.search(f'<tr class="trtit" id="{v}Title">(.*?)</tr>',
tbody.group(), re.S).group()
    dates = re.findall('<td.*?>(.*?)</td>', dtr, re.S)
    das = re.findall('<td.*?>(.*?)</td>', tbody.group(), re.S)
    for i in range(0, len(das), len(dates)):
        da = das[i: i+len(dates)]
        if da[0] not in df_dc[k].keys():
            df_dc[k][da[0]] = da[1:]
        else:
            df_dc[k][da[0]] += da[1:]
    df_dc[k]['行业'] += [row['所属行业'] for p in range(len(dates)-1)]
    df_dc[k]['股票'] += [row['证券名称'] for p in range(len(dates)-1)]
```

(1) 使用正则表达式从网页内容中提取所需的数据表格。

(2) 提取表头（日期）和数据。

(3) 将数据按日期分组，并存储到 df_dc 字典中，同时添加对应的行业和股票信息。

```
#保存数据
with pd.ExcelWriter('[UserFolderPath]/数据采集结果.xlsx') as writer:
    for k, v in df_dc.items():
        for dk, dv in v.items():
            if len(dv) != len(v['行业']):
                v[dk] = dv + [' ' for x in range(len(v['行业']) - len(dv))]
            df_result = pd.DataFrame(v)
            df_result.to_excel(writer, sheet_name=k, index=False)
```

(1) 使用 pandas 的 ExcelWriter 将整理好的数据保存到 Excel 文件中。

(2) 对于每个表格，如果数据长度与行业信息长度不匹配，则补齐空值。

(3) 将每个表格的数据保存为一个单独的 sheet。

任务 2：处理采集的数据——步骤 2 数据处理

数据处理说明如下：

- (1) 清除原网页中未记载数据，即指标数据为空的数据。
- (2) 去除量纲较大的数据列，如成长能力指标、每股指标、盈利能力指标、运营能力指标、偿债能力指标、净利润(元)、扣非净利润(元)、营业总收入(元)等。
- (3) 将百分数转化为浮点数
- (4) 货币计数转化为浮点数
- (5) 选择违规企业的数据和未违规股票的 80%作为训练集，选择【尔康】和未违规股票的 20%作为测试集。

详细代码如下：

```
import pandas as pd # 引入数据框
from sklearn.model_selection import train_test_split # 用来划分训练集与测试集
```

- (1) Pandas：用于数据处理和分析，特别是数据清洗和转换。
- (2) train_test_split：从 sklearn.model_selection 模块导入，用于将数据集分割为训练集和测试集。

```
def _data_cleaning(df):
    # 清除原网页中未记载数据
    for i in range(3, 32):
        df = df.drop(index=df[df.iloc[:, i] == '--'].index)
```

这个循环遍历数据框的第 4 列到第 32 列（索引从 3 开始），删除所有值为 '--' 的行。这通常用于去除缺失或无效数据。

```
# 去除指标不需要列
df = df.drop('成长能力指标', axis=1)
df = df.drop('每股指标', axis=1)
df = df.drop('盈利能力指标', axis=1)
df = df.drop('运营能力指标', axis=1)
df = df.drop('偿债能力指标', axis=1)
df = df.drop('净利润(元)', axis=1)
df = df.drop('扣非净利润(元)', axis=1)
df = df.drop('营业总收入(元)', axis=1)
```

这些行删除了一些不需要的列，以简化数据框。被删除的列名表明它们可能与成长能力、盈利能力等相关，但在后续分析中可能并不需要。

```
# 百分数转化为浮点数
```

```

df['净利润同比增长率'] = df['净利润同比增长率
'].str.rstrip('%').astype('float') / 100.0
df['扣非净利润同比增长率'] = df['扣非净利润同比增长率
'].str.rstrip('%').astype('float') / 100.0
df['营业总收入同比增长率'] = df['营业总收入同比增长率
'].str.rstrip('%').astype('float') / 100.0
df['销售净利率'] = df['销售净利率'].str.rstrip('%').astype('float') /
100.0
df['销售毛利率'] = df['销售毛利率'].str.rstrip('%').astype('float') /
100.0
df['净资产收益率'] = df['净资产收益率'].str.rstrip('%').astype('float')
/ 100.0
df['净资产收益率-摊薄'] = df['净资产收益率-摊薄
'].str.rstrip('%').astype('float') / 100.0
df['资产负债率'] = df['资产负债率'].str.rstrip('%').astype('float') /
100.0

```

(1) 将某些百分比格式的字符串转换为浮点数。通过 `str.rstrip('%')` 方法去掉字符串末尾的百分号，然后再将其转换为浮点数并除以 100，得到实际的比例值。

```

# 货币计数转化为浮点数
df['基本每股收益(元)'] = df['基本每股收益(元)'].apply(
    lambda d: float(d.replace('$', '').replace(',','')))
df['每股净资产(元)'] = df['每股净资产(元)'].apply(
    lambda d: float(d.replace('$', '').replace(',','')))
df['每股资本公积金(元)'] = df['每股资本公积金(元)'].apply(
    lambda d: float(d.replace('$', '').replace(',','')))
df['每股未分配利润(元)'] = df['每股未分配利润(元)'].apply(
    lambda d: float(d.replace('$', '').replace(',','')))
df['每股经营现金流(元)'] = df['每股经营现金流(元)'].apply(
    lambda d: float(d.replace('$', '').replace(',','')))
df['营业周期(天)'] = df['营业周期(天)'].apply(
    lambda d: float(d.replace('$', '').strip('万').replace(',','')))
df['存货周转率(次)'] = df['存货周转率(次)'].apply(
    lambda d: float(d.replace('$', '').strip('万').replace(',','')))
df['存货周转天数(天)'] = df['存货周转天数(天)'].apply(
    lambda d: float(d.replace('$', '').strip('万').replace(',','')))
df['应收账款周转天数(天)'] = df['应收账款周转天数(天)'].apply(
    lambda d: float(d.replace('$', '').replace(',','')))
df['流动比率'] = df['流动比率'].apply(lambda d: float(
    d.replace('$', '').replace(',','')))
df['速动比率'] = df['速动比率'].apply(lambda d: float(
    d.replace('$', '').replace(',','')))
df['保守速动比率'] = df['保守速动比率'].apply(
    lambda d: float(d.replace('$', '').replace(',','')))

```

```
df['产权比率'] = df['产权比率'].apply(lambda d: float(
    d.replace('$', '').replace(',', '')))
return df
```

(1) 针对不同的货币数据列（如每股收益、每股净资产等），使用 `apply` 方法结合 `lambda` 函数，对字符串进行清理（去掉\$和,）并转换为浮点数。

(2) 对于营业周期、存货周转率等指标，代码同样使用 `apply` 和 `lambda` 函数来进行字符串处理和转换，确保这些数据都是以浮点数的形式存在。

读取文件

```
df = pd.read_excel('[UserFolderPath]/数据采集结果.xlsx', sheet_name='主要指标', dtype=str)
```

(1) 使用 `pandas` 的 `read_excel` 函数从 Excel 文件中读取数据。

(2) 指定 `sheet_name='主要指标'` 来选择特定的工作表。

(3) 设置 `dtype=str` 确保所有读取的数据都以字符串形式存储，这有助于后续的数据清洗。

```
df = df.dropna(how='all') # 去除空白值
df = _data_cleaning(df) # 去除未记载数据和转换格式
```

(1) 使用 `dropna(how='all')` 去除所有值都为空的行。

(2) 调用 `_data_cleaning` 函数（未在代码段中给出，但假设它执行了数据清洗任务，如删除无效数据、转换数据类型等）。

```
df.rename(columns={'科目\年度': '时间'}, inplace=True)
```

(1) 重命名列名，将 '科目\年度' 改为 '时间'，以更清晰地表达其含义。

将尔康制药（违规）和中国医药（未违规）单独提出来作为预测集来测试泛化性能，泛化精度越高，模型实战效果越好

```
df_pr_1 = df[df['股票'] == '尔康制药']
df_pr_1 = pd.concat([df_pr_1[df_pr_1['时间'].str.contains('2015')], # 选取违规年份
                    df_pr_1[df_pr_1['时间'].str.contains('2016')]])
df_pr_1['类别'] = 1
df_pr_0 = pd.concat(
    [
        df.query("股票=='中国医药' & 时间.str.contains('2015')"),
        df.query("股票=='中国医药' & 时间.str.contains('2016')"),
    ]
)
df_pr_0['类别'] = 0
df_pr = pd.concat([df_pr_1, df_pr_0], axis=0)
df_pr.to_csv('[UserFolderPath]/尔康&中国医药.csv', encoding='gbk')
```

(1) 提取特定股票（尔康制药和中国医药）在特定年份（违规和未违规年份）的数据。

(2) 为这些数据添加 '类别' 列，分别标记为 1（违规）和 0（未违规）。

将这两个数据集合并，并保存为 CSV 文件。

```
# 删除尔康制药的留下来做训练集与测试集
df_total = df.drop(index=df[df['股票'] == '尔康制药'].index).drop(index=df[df['股票'] == '中国医药'].index)
# 找到违规的打上标签 1
df_1 = pd.concat([
    df_total.query("股票 == '复星医药' & 时间.str.contains('2018')"),
    df_total.query("股票 == 'ST 康美' & 时间.str.contains('2016')"),
    df_total.query("股票 == 'ST 康美' & 时间.str.contains('2017')"),
    df_total.query("股票 == 'ST 康美' & 时间.str.contains('2018')"),
    df_total.query("股票 == '思创医惠' & 时间.str.contains('2019')"),
    df_total.query("股票 == '思创医惠' & 时间.str.contains('2020')"),
    df_total.query("股票 == 'ST 太安' & 时间.str.contains('2018')"),
    df_total.query("股票 == 'ST 太安' & 时间.str.contains('2019')"),
    df_total.query("股票 == 'ST 太安' & 时间.str.contains('2020')"),
    df_total.query("股票 == 'ST 太安' & 时间.str.contains('2021')"),
    df_total.query("股票 == '亚太药业' & 时间.str.contains('2016')"),
    df_total.query("股票 == '亚太药业' & 时间.str.contains('2017')"),
    df_total.query("股票 == '亚太药业' & 时间.str.contains('2018')")
])
df_1['类别'] = 1
# 找到未违规的打上标签 0
df_0 = df_total.drop(index=df[df['股票'] == '复星医药'].index)
df_0 = df_0.drop(index=df[df['股票'] == 'ST 康美'].index)
df_0 = df_0.drop(index=df[df['股票'] == '思创医惠'].index)
df_0 = df_0.drop(index=df[df['股票'] == 'ST 太安'].index)
df_0 = df_0.drop(index=df[df['股票'] == '亚太药业'].index)
df_0['类别'] = 0
# 合并数据集
df_total = pd.concat([df_1, df_0], axis=0)
```

(1) 删除尔康制药和中国医药的数据，因为它们的数据已经被用作预测集。

(2) 分别查询和标记违规（df_1）和未违规（df_0）的股票数据。

(3) 合并这两个数据集，形成完整的训练集和测试集的基础。

```
# 划分训练集与测试集
X = df_total.iloc[:, 3:24]
y = df_total.iloc[:, 24:25]
```

```
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.2)
df_train = pd.concat([X_train, y_train], axis=1)
df_test = pd.concat([X_test, y_test], axis=1)
```

(1) 使用 `iloc` 选择特征列和目标列。

(2) 使用 `train_test_split` 函数将数据集划分为训练集和测试集，测试集大小为 20%。

(3) 将训练集和测试集的特征和目标合并回各自的 `DataFrame` 中。

```
# 存储
df_train.to_csv('[UserFolderPath]/训练集.csv', encoding='gbk')
df_test.to_csv('[UserFolderPath]/测试集.csv', encoding='gbk')
```

(1) 将训练集和测试集保存为 CSV 文件，以便后续使用。

任务 3：审计风险预测

代码需求说明：

(1) 实现一个全连接神经网络模型的训练和预测，并对测试集和新数据进行了精度和预测结果的评估。

(2) 运用模型预测被审计单位最近年度的审计风险。

```
# 导入需要的库
'''
使用 tensorflow 的顺序模型实现全连接神经网络
'''
import pandas as pd
import numpy as np
from keras.models import Sequential
from tensorflow.python.keras.layers.core import Dense, Activation
import tensorflow as tf
import os
```

(1) `pandas`: 用于数据读取和处理。

(2) `numpy`: 用于数值计算。

(3) `Sequential`: Keras 中的顺序模型，用于线性堆叠网络层。

(4) `Dense`: 全连接层。

(5) `Activation`: 激活函数层。

(6) `tensorflow`: 一个开源的机器学习库。

(7) `os`: 用于设置环境变量。

```
# 设置环境变量 TF_CPP_MIN_LOG_LEVEL
os.environ['TF_CPP_MIN_LOG_LEVEL'] = '2' # 只显示警告和错误信息
这行代码用于减少 TensorFlow 的日志输出，只显示警告和错误信息。
```

```
# 混淆矩阵可视化
```

```

def cm_plot(y, yp): # 混淆矩阵可视化
    from sklearn.metrics import confusion_matrix # 导入混淆矩阵函数
    cm = confusion_matrix(y, yp) # 混淆矩阵
    import matplotlib.pyplot as plt # 导入作图库
    plt.matshow(cm) # 画混淆矩阵图, 配色风格使用 cm.Greens, 更多风格请参考官
    网。
    plt.colorbar() # 颜色标签
    for x in range(len(cm)): # 数据标签
        for y in range(len(cm)):
            plt.annotate(cm[x, y], xy=(x, y),
                horizontalalignment='center',
                verticalalignment='center')
    plt.xlabel('True label') # 坐标轴标签
    plt.ylabel('Predicted label') # 坐标轴标签
    return plt

```

该函数使用 `sklearn.metrics.confusion_matrix` 计算混淆矩阵，并使用 `matplotlib` 进行可视化。混淆矩阵是评估分类模型好坏的重要工具。

使用随机种子使得训练过程可复现，一般情况每次训练的精度会有略微不同，这是由于算法特性所导致

```
tf.random.set_seed(1)
```

这行代码确保每次运行代码时，初始化的随机数是相同的，使得实验结果可复现。

导入训练集

```
filename = '[UserFolderPath]/训练集.csv'
```

```
data = pd.read_csv(filename, encoding='gbk')
```

去除空集重新排序

```
data = data.dropna()
```

```
data = data.sort_values('类别', ascending=False)
```

分开特征矩阵和类别矩阵

```
X = data.iloc[:, 1:22]
```

```
X.astype(np.float32)
```

```
y = data.iloc[:, 22].astype(int)
```

- (1) 使用 `pandas` 读取 CSV 文件。
- (2) 去除空值并排序。
- (3) 分割特征矩阵 `X` 和标签矩阵 `y`。

注意：`X.astype(np.float32)` 这行代码没有重新赋值给 `X`，因此 `X` 的数据类型并没有改变。

实例化模型

```
model = Sequential() # 建立顺序模型
```

```
model.add(Dense(input_dim=21, units=64)) # 全连接层, 21 进 64 出
```

```
model.add(Activation('relu')) # 使用 relu 激活函数
```

```
model.add(Dense(input_dim=64, units=1)) # 全连接层, 64 进 1 出
```

```
model.add(Activation('sigmoid')) # 由于是 0-1 输出, 用 sigmoid 函数作为激活函数
```

- (1) 使用 `Sequential` 模型。
- (2) 添加两个全连接层，第一个 64 个神经元，第二个 1 个神经元（用于二分类）。

(3) 第一个全连接层后使用 `relu` 激活函数，第二个全连接层后使用 `sigmoid` 激活函数。

```
# 使用二分类交叉熵作为损失函数，优化器梯度下降法
```

```
model.compile(loss='binary_crossentropy', optimizer='adam')
```

(1) 使用二分类交叉熵作为损失函数。

(2) 使用 Adam 优化器。

```
# 开始训练
```

```
# 训练次数 100 次，次数过多会造成过拟合，即测试精度很高但泛化精度降低，训练批次为 10 每次
```

```
model.fit(X, y, epochs=10, batch_size=2, verbose=0)
```

(1) 使用训练数据训练模型。

(2) `epochs=10` 表示训练 10 个周期。

(3) `batch_size=2` 表示每次更新权重时使用的样本数为 2。

(4) `verbose=0` 表示不在控制台输出训练过程。

```
# 测试集精度
```

```
test = pd.read_csv('[UserFolderPath]/测试集.csv', encoding='gbk')
```

```
test = test.dropna()
```

```
test = test.sort_values('类别', ascending=False)
```

```
X = test.iloc[:, 1:22]
```

```
y = test.iloc[:, 22]
```

```
yp = (model.predict(X, batch_size=10, verbose=0) >
```

```
0.5).astype("int32").reshape(len(y))
```

```
cm_plot(y, yp).savefig('[UserFolderPath]/混淆矩阵可视化 1.png')
```

```
accuracy = float((yp == y).astype(int).sum()) / float(y.size)
```

```
print(f'测试集的精度为{accuracy}')
```

(1) 使用测试数据评估模型性能。

(2) 将预测结果大于 0.5 的视为类别 1，否则视为类别 0。

(3) 计算测试集的精度。

```
# 泛化性能测试实战
```

```
erkang = pd.read_csv('[UserFolderPath]/尔康&中国医药.csv', encoding='gbk')
```

```
erkang.iloc[:, 4:25]
```

```
X = erkang.iloc[:, 4:25]
```

```
X.astype(np.float32)
```

```
y = np.array([[1], [1],
```

```
[0], [0]])
```

```
# 按照真实情况，尔康制药（财务造假，类别为 1），中国医药（未造假，类别为 0）
```

```
# 进行预测
```

```
y_predict = model.predict(X, verbose=0)
```

```
yp = y_predict.round(0).astype(int)
```

```
print(f'模型的预测结果为: {yp}')
```

```
cm_plot(y, yp).savefig('[UserFolderPath]/混淆矩阵可视化 2.png')
```

(1) 对新的数据集进行预测。

(2) 将预测结果四舍五入后转换为整数。

```
# 输出结果
```

```

ls = []
for i in yp:
    k = '违规' if i == 1 else '未违规'
    ls.append(k)
erkang['是否违规'] = ls
erkang.to_csv('[UserFolderPath]/是否财务造假.csv')

```

(1) 将预测结果添加到数据框中并保存为新的 CSV 文件。

项目八：银行交易异常识别

实训步骤

步骤 1：数据采集

需求说明：

- (1) 导入必要的库和模块。
- (2) 读取银行对账单数据。

```

import pandas as pd
from sklearn.ensemble import RandomForestClassifier
from sklearn.model_selection import train_test_split
from sklearn.metrics import classification_report
from sklearn.preprocessing import LabelEncoder

```

```

# 读取银行对账单数据
bank_statement = pd.read_csv('银行对账单表 2022.csv', encoding='ANSI',
dtype={'业务产品种类': str}) # 银行对账单数据

```

- (1) `pd.read_csv` 是 `pandas` 库中用于读取 CSV 文件的函数。
- (2) `encoding='ANSI'` 指定了文件的编码方式。
- (3) `dtype={'业务产品种类': str}` 设置了特定列的数据类型，确保“业务产品种类”列被正确读取为字符串类型。

步骤 2：数据处理

数据处理步骤涉及数据清洗、转换和准备。

需求说明：

- (1) 导入必要的库和模块。

(2) 读取 CSV 文件，并指定编码为 ANSI。

1、特征工程

特征工程是数据处理的重要部分，它涉及从原始数据中提取和创建有意义的特征。

```
# 对非数值型特征进行编码
label_encoder = LabelEncoder()
#清理数据，将对方户名，摘要列的所有数据转化为字符串
for col in ['业务产品种类', '对方户名', '摘要']:
    bank_statement[col] = bank_statement[col].astype(str)
bank_statement['业务产品种类'] =
label_encoder.fit_transform(bank_statement['业务产品种类'])
bank_statement['对方户名'] = label_encoder.fit_transform(bank_statement['
对方户名'])
bank_statement['摘要'] = label_encoder.fit_transform(bank_statement['摘要
'])
```

```
# 选择特征和目标列
features = ['日期', '业务产品种类', '对方户名', '摘要', '借方发生额', '贷方
发生额', '交易金额', '余额']
target = '审计结果' # 目标列名称
```

(1) for 循环：将列名为 '业务产品种类'、'对方户名' 和 '摘要' 的所有数据转换为字符串类型。这是因为有些机器学习算法要求输入的特征必须是数值型的，因此需要先将非数值型的特征数据转换为字符串类型。

(2) LabelEncoder 是用来将非数值型的分类特征转换为数值标签的工具。对于每一个列名为 '业务产品种类'、'对方户名' 和 '摘要' 的特征列，使用 fit_transform() 方法对其进行编码转换。每个不同的字符串值会被映射为一个唯一的整数值。

(3) features 列表定义了用于训练模型的特征列，包括 '日期'、'业务产品种类'、'对方户名'、'摘要'、'借方发生额'、'贷方发生额'、'交易金额' 和 '余额'。

(4) target 变量指定了模型要预测的目标列，即 '审计结果' 列。在机器学习中，目标列通常是我们希望模型预测或分类的列。

2、数据清理

```
# 清理数据，移除千位分隔符并将字符串转换为浮点数
for col in ['借方发生额', '贷方发生额', '交易金额', '余额']:
    bank_statement[col] = bank_statement[col].str.replace(',',
).astype(float)
```

```
# 将日期字段转换为时间戳
bank_statement['日期'] = pd.to_datetime(bank_statement['日期
']).astype('int64') // 10**9 # 转换为秒级时间戳
```

```
# 填充 NaN 值为 0
bank_statement.fillna(0, inplace=True) # 在这里添加这行代码
```

- (1) 使用 `.str.replace(',', '')` 移除千位分隔符。
- (2) 使用 `.astype(float)` 将字符串转换为浮点数。
- (3) 使用 `fillna(0, inplace=True)` 填充所有 NaN 值为 0 (`inplace=True` 表示直接在原 DataFrame 上进行修改)。

(4) 代码还将日期字段转换为时间戳，有助于机器学习算法处理日期数据：

- ① `pd.to_datetime`：将日期字符串转换为 `datetime` 对象。
- ② `astype('int64')`：将 `datetime` 对象转换为时间戳的整数表示。
- ③ `// 10**9`：将时间戳从纳秒转换为秒。

步骤 3：模型构建

需求说明：

- (1) 划分训练集和测试集。
- (2) 构建随机森林模型
- (3) 拟合模型。
- (4) 在测试集上进行预测
- (5) 输出模型评价结果。

1. 划分训练集和测试集

```
# 划分训练集和测试集
X_train, X_test, y_train, y_test =
train_test_split(bank_statement[features], bank_statement[target],
test_size=0.2, random_state=42)
```

(1) `train_test_split` 是 Scikit-learn 库中提供的一个非常实用的函数，用于将数据集划分为训练集和测试集。

(2) `bank_statement[features]`：这通常是一个 Pandas DataFrame 的列子集，代表特征数据。
`features` 是一个列表，包含了需要从 `bank_statement` 中选择的列名。

(3) `bank_statement[target]`: 这是 Pandas DataFrame 中的一列，代表目标变量或标签。`target` 是目标列的名称。

(4) `test_size=0.2`: 这个参数指定了测试集应该包含原始数据集的比例。在这个例子中，20% 的数据将被用作测试集，而剩下的 80% 将被用作训练集。

(5) `random_state=42`: 这个参数用于设置随机种子，以确保每次运行代码时数据划分的结果都是相同的（在可重复性的背景下非常重要）。

(6) 返回值:

`X_train`: 包含 80% 原始数据的特征，用于训练模型。

`X_test`: 包含 20% 原始数据的特征，用于测试模型。

`y_train`: 与 `X_train` 对应的目标变量或标签。

`y_test`: 与 `X_test` 对应的目标变量或标签。

2. 构建随机森林模型

```
# 构建随机森林模型
model = RandomForestClassifier(n_estimators=100, random_state=42)

# 拟合模型
model.fit(X_train, y_train)
```

(1) `RandomForestClassifier` 是 `sklearn.ensemble` 中的一个工具，用于构建随机森林分类器。

(2) `n_estimators=100` 指定了森林中树的数量。

(3) `random_state=42` 确保结果的可复现性。

(4) `model.fit(X_train, y_train)` 是机器学习流程中的一个关键步骤，它负责训练模型。具体来说，这行代码告诉模型使用提供的训练数据（`X_train` 作为特征，`y_train` 作为目标变量或标签）来学习如何预测或分类数据：

① `model`: 这是之前已经初始化的机器学习模型对象，在这个例子中是一个 `RandomForestClassifier` 对象。

② `X_train`: 这是一个包含训练数据的特征矩阵。通常，它是从原始数据集中划分出来的，只包含用于训练模型的数据点。

③ `y_train`: 这是一个与 `X_train` 相对应的目标变量数组或标签数组。它包含了与 `X_train` 中每个数据点相关联的正确分类或预测值。

3. 评估模型

在测试集上评估模型的性能，并输出分类报告和特征重要性。

```
# 在测试集上进行预测
y_pred = model.predict(X_test)

# 输出分类报告
print("分类报告: ")
print(classification_report(y_test, y_pred))

# 输出特征重要性
print("特征重要性: ")
feature_importance = pd.Series(model.feature_importances_,
                                index=features)
print(feature_importance)
```

- (1) model.predict(X_test): 在测试集上进行预测。
- (2) classification_report: 显示了模型的精确度、召回率、F1 分数等指标。
- (3) model.feature_importances_ : 提供了每个特征对模型预测的贡献程度。

运行结果:

```
分类报告:
              precision    recall  f1-score   support

   异常         0.96      0.76      0.85         84
   正常         0.92      0.99      0.95        226

 accuracy                   0.93         310
 macro avg                   0.94         310
weighted avg                   0.93         310

特征重要性:
日期          0.083106
业务产品种类  0.007375
对方户名      0.150325
摘要          0.181773
借方发生额    0.201748
贷方发生额    0.038008
交易金额      0.220111
余额          0.117554
dtype: float64
预测结果:
['正常' '正常' '正常' ... '正常' '正常' '异常']
```

结果分析:

1、分类报告

这是一个二分类问题的结果，分为“异常”和“正常”两个类别。我们可以根据报告中的各项指标来评估模型的性能:

1.1 精确度 (Precision) :

(1) 异常类别: 0.96, 意味着在模型预测为异常的样本中, 有 96%的样本是真正的异常。

(2) 正常类别: 0.92, 意味着在模型预测为正常的样本中, 有 92%的样本是真正的正常。

1.2 召回率 (Recall) :

(1) 异常类别: 0.76, 意味着在所有真正的异常样本中, 模型找出了 76%的样本。

(2) 正常类别: 0.99, 意味着在所有真正的正常样本中, 模型找出了 99%的样本。

1.3 F1 分数 (F1-score) :

(1) F1 分数是精确度和召回率的调和平均值, 用于综合评估模型性能。

(2) 异常类别: 0.85, 表明模型在异常类别的预测上性能较为均衡。

(3) 正常类别: 0.95, 表明模型在正常类别的预测上性能较好。

1.4 准确率 (Accuracy) : 0.93, 意味着在所有样本中, 模型有 93%的预测是正确的。

2、特征重要性

根据特征重要性的分析, 以下是对预测影响最大的特征 (按重要性排序) :

(1) 交易金额(0.220111): 这是最重要的特征, 对分类异常和正常交易都有显著影响。

(2) 摘要(0.181773)和对方户名(0.150325): 这些特征也对分类起到了重要作用, 特别是在识别异常交易方面。

(3) 余额 (0.117554)和日期(0.083106): 这些特征的重要性较低, 但仍然对模型的整体性能有一定影响。

(4) 借方发生额(0.201748)和贷方发生额(0.038008): 借方和贷方发生额的重要性分别为第二和第四位, 显示出这些金额变动对于异常检测的重要性。

3、结论与建议

最后提到的预测结果显示大多数交易被预测为正常, 但也有一部分被预测为异常。模型在整体上表现良好, 尤其是对于正常交易的识别准确率很高。

步骤 4: 异常数据识别

1、数据读取

```
# 读取新的银行对账单数据
```

```
new_bank_statement = pd.read_csv('银行对账单表 2021 预测.csv',
encoding='ANSI')
```

2、数据处理

(1) 定义特征列

```
# 定义您想要进行预测的特征列
features = ['日期', '业务产品种类', '对方户名', '摘要', '借方发生额', '贷方发生额', '交易金额', '余额']
```

(2) 对非数值型特征进行编码

```
# 对非数值型特征进行编码
label_encoder = LabelEncoder()

for col in ['业务产品种类', '对方户名', '摘要']:
    if col in new_bank_statement.columns: # 确保列存在于 DataFrame 中
        new_bank_statement[col] = new_bank_statement[col].astype(str)
        new_bank_statement[col] =
label_encoder.fit_transform(new_bank_statement[col])
```

(1) 使用 LabelEncoder()对非数值型特征列进行编码，将其转换为数值型，包括'业务产品种类'、'对方户名'和'摘要'。

(3) 清理数据

(1) 检查并处理借方发生额、贷方发生额、交易金额和余额列，移除千位分隔符并将其转换为浮点数。

(2) 将日期字段转换为 Unix 时间戳。

(3) 填充 NaN 值为 0。

```
# 清理数据，移除千位分隔符并将字符串转换为浮点数
for col in ['借方发生额', '贷方发生额', '交易金额', '余额']:
    if new_bank_statement[col].dtype == str: # 检查列的数据类型是否为字符串
        new_bank_statement[col] =
new_bank_statement[col].str.replace(',', '').astype(float)
    elif new_bank_statement[col].dtype in [float, int]: # 如果列已经是数值类型，则不需要替换逗号
        # 在这里你可能什么都不需要做，除非你需要处理其他非数值数据
        pass
    else:
        # 处理其他数据类型的情况，如果需要的话
        print(f"Column '{col}' has unsupported dtype:
{new_bank_statement[col].dtype}")
```

(1) 代码首先通过一个 `for` 循环遍历列表 ['借方发生额', '贷方发生额', '交易金额', '余额'] 中的每个元素（即列名）。这些列被假定为可能包含以字符串形式存储的数值，这些数值可能包含千位分隔符。

(2) 对于每一列，代码首先检查该列的数据类型 (`dtype`) 是否为字符串 (`str`)。这是通过 `new_bank_statement[col].dtype == str` 来实现的。如果列的数据类型是字符串，那么代码将继续执行下一步的数据清洗操作。

(3) 如果列的数据类型是字符串，代码会执行两个操作：

① 使用 `str.replace(',', '')` 方法移除字符串中的逗号（即千位分隔符）。这一步是必要的，因为逗号会阻止字符串被直接转换为浮点数。

② 接着，使用 `astype(float)` 将清洗后的字符串转换为浮点数。这样，原本以字符串形式存储的数值数据就被转换成了适合数值计算的浮点数类型。

(4) 如果列的数据类型已经是浮点数 (`float`) 或整数 (`int`)，则代码中的 `elif` 分支会执行。在这个例子中，如果列已经是数值类型，那么就不需要进行任何操作（通过 `pass` 语句表示）。这可能是因为数值类型的数据已经是以正确的格式存储，无需额外的清洗或转换。

(5) 如果列的数据类型既不是字符串也不是数值类型（浮点数或整数），则代码会打印一条消息，指出该列具有不支持的数据类型。这通常是一个警告，表明可能需要对这些列进行额外的处理或检查，以确保数据的准确性和完整性。

```
# 将日期字段转换为时间戳（这里转换为 Unix 时间戳秒数）
new_bank_statement['日期'] = pd.to_datetime(new_bank_statement['日期'])
new_bank_statement['日期'] = new_bank_statement['日期'].astype('int64') // 10**9
```

(6) 代码首先使用 `pd.to_datetime` 方法将“日期”列从字符串转换为日期时间对象。然后，它使用 `astype('int64')` 将日期时间对象转换为 Unix 时间戳（以纳秒为单位）。最后，通过整除 `10**9` 将时间戳从纳秒转换为秒。

```
# 填充 NaN 值为 0（注意：这里假设填充 0 是合理的，实际中可能需要根据具体情况判断）
new_bank_statement.fillna(0, inplace=True)
```

(7) 这行代码使用 `fillna` 方法将所有 NaN 值（即缺失值）替换为 0。`inplace=True` 参数表示直接在原始 `DataFrame` 上进行修改，而不是返回一个新的 `DataFrame`。

3、数据预测

(1) 选择特征列

```
# 选择特征列（确保 features 列表中只包含 new_bank_statement 中的列）
new_X = new_bank_statement[features]
```

从 `new_bank_statement` 这个 DataFrame 中选择由 `features` 列表指定的列，并将这些列的数据赋值给新的 DataFrame `new_X`。

(2) 进行预测

```
# 在新数据上进行预测
# 注意: features 中不应该包含'审计结果'列
new_y_pred = model.predict(new_X)
```

使用已经训练好的模型 `model` 对 `new_X` 中的数据进行预测，并将预测结果赋值给 `new_y_pred`。注释中强调了 `features` 列表中不应该包含'审计结果'列，因为'审计结果'通常是模型要预测的目标变量，而不是用于预测的特征。

(3) 输出结果

```
# 输出预测结果
print("预测结果: ")
print(new_y_pred)
```

```
# 如果要将预测结果保存为新的列或 CSV 文件
new_bank_statement['预测结果'] = new_y_pred
new_bank_statement.to_csv('银行对账单表_预测结果.csv', index=False)
```

(1) 先将预测结果 `new_y_pred` 作为新的列'预测结果'添加到原始的 `new_bank_statement` DataFrame 中。

(2) 然后，使用 `to_csv` 方法将更新后的 `new_bank_statement` DataFrame 保存到名为'银行对账单表_预测结果.csv'的 CSV 文件中。`index=False` 参数确保在保存时不包含 DataFrame 的索引。

项目九：时间序列预测审计应用

实训步骤

1、数据采集

需求说明：

1、打开同花顺个股中（<https://stockpage.10jqka.com.cn/>），找到对应股票的财务分析界面，采集一家上市公司历年营业收入数据。

2、采集的数据保存在“数据爬取结果”文件夹中。

具体代码修改如下：

```
import pandas as pd # 使用数据框进行数据读取与存储
from selenium import webdriver # 导入 selenium 库进行爬虫
from selenium.webdriver.chrome.service import Service
from selenium.webdriver.common.by import By
from selenium.webdriver.common.action_chains import ActionChains
from selenium.webdriver import Keys
import time # 使用 time 库来实现暂停，减少爬虫过程中的报错

class HAND():
    def __init__(self, url, name) -> None:
        self.url = url
        self.name = name
    def run(self):
        # 设置数据格式
        dic = {
            '年份': [],
            '营业收入(元)': []
        }
        # 实例化 webdriver 对象，用来模拟浏览器操作
        wd = webdriver.Chrome(
            service=Service(r'chromedriver.exe')
        )
        wd.implicitly_wait(20) # 隐式等待，防止网速问题造成报错
        url = self.url
        url = str(url)
        wd.get(url)
        mainWindow = wd.current_window_handle # 设置为操作窗口
        wd.switch_to.frame('dataifm')
        wd.find_element(By.XPATH, "//a[text()='按单季度']").click() #确定
        爬取的数据：“按报告期” “按年度” “按单季度”

        _date_line_of_this_stock = wd.find_elements(
            # 获取所爬股票的信息时间轴
            By.XPATH, '//table[@class="top_thead" and @style="width:
100%;"]//tr/*')
        _len_of_times = len(wd.find_elements(
            # 时间轴总长度
            By.XPATH, '//table[@class="top_thead" and @style="width:
100%;"]/tbody/tr/*'))
        for i in range(0, _len_of_times // 6 + 1):
            _needed_times = 6 if _len_of_times - \
                (i) * 6 >= 6 else _len_of_times % 6
```

```

_value_tb = wd.find_elements(
    By.XPATH, '//table[@class="tbody" and @style="top: 0px;
width: 100%;"]/tbody/*')
_value = _value_tb[5] # 第 height 行的时间序列数据
_value = _value.find_elements(By.XPATH, './*') # 抓取一整行
for times in range(0, _needed_times):
    dic['年份
'].append(_date_line_of_this_stock[times+6*i].text)
    dic['营业收入(元)'].append(_value[times+6*i].text)
    _click_times = 0
    while _click_times < 6 and _click_times < _len_of_times - (i+1)
* 6:
        wd.find_element(
            By.XPATH, '//div[@class="data-icon data-icon-next
"]').click()
        time.sleep(0.2)
        _click_times += 1
    df = pd.DataFrame(dic) # 保存为数据框
    name = self.name
    path = '数据爬取结果/' + str(name) + '.csv'
    df.to_csv(path, encoding='gbk')
    print(df)
if __name__ == '__main__':
    a = HAND('https://stockpage.10jqka.com.cn/002612/finance/', '朗姿股份
').run()
    #a = HAND(input('输入爬取的网址: '), input('输入公司名称:')).run()

```

采集结果:

序号	年份	营业收入(元)
1	2010/12/31	1.69 亿
2	2011/3/31	2.32 亿
3	2011/6/30	2.00 亿
4	2011/9/30	1.88 亿
5	2011/12/31	2.16 亿
6	2012/3/31	3.17 亿
7	2012/6/30	2.30 亿
8	2012/9/30	2.36 亿
9	2012/12/31	3.35 亿
10	2013/3/31	3.70 亿
11	2013/6/30	3.14 亿
12	2013/9/30	3.19 亿
13	2013/12/31	3.75 亿
14	2014/3/31	3.34 亿
15	2014/6/30	2.84 亿
16	2014/9/30	2.68 亿

17	2014/12/31	3.50 亿
18	2015/3/31	3.27 亿
19	2015/6/30	2.43 亿
20	2015/9/30	2.67 亿
21	2015/12/31	3.08 亿
22	2016/3/31	2.66 亿
23	2016/6/30	2.06 亿
24	2016/9/30	3.53 亿
25	2016/12/31	5.42 亿
26	2017/3/31	5.36 亿
27	2017/6/30	5.21 亿
28	2017/9/30	5.87 亿
29	2017/12/31	7.09 亿
30	2018/3/31	6.46 亿
31	2018/6/30	6.52 亿
32	2018/9/30	6.98 亿
33	2018/12/31	6.65 亿
34	2019/3/31	7.22 亿
35	2019/6/30	6.76 亿
36	2019/9/30	7.31 亿
37	2019/12/31	8.78 亿
38	2020/3/31	5.87 亿
39	2020/6/30	6.42 亿
40	2020/9/30	7.68 亿
41	2020/12/31	8.80 亿
42	2021/3/31	8.92 亿
43	2021/6/30	8.97 亿
44	2021/9/30	8.61 亿
45	2021/12/31	11.82 亿
46	2022/3/31	9.39 亿
47	2022/6/30	9.65 亿
48	2022/9/30	11.22 亿
49	2022/12/31	8.52 亿
50	2023/3/31	11.39 亿
51	2023/6/30	11.74 亿
52	2023/9/30	13.81 亿

2、数据处理

采集的数据可能存在空值和单位，需要先去除数据框中的空值，再处理数据框中未记录的值的，最后将所有带单位的数转化为纯浮点数，结果保存在“数据处理结果”文件夹。

需求说明:

- 1、删除空值行;
- 2、删除营业收入为'--'的行;
- 3、转换营业收入为数值, 并将数据类型转换为浮点数。

代码:

```
import pandas as pd

class DEAL():
    def __init__(self, name) -> None:
        self.name = name

    def run(self):
        name = self.name
        path = '数据爬取结果/' + str(name) + '.csv'
        df = pd.read_csv(path, encoding='gbk')
        df = df.dropna()
        df = df.drop(index=df[df['营业收入(元)'] == '--'].index)
        df['营业收入(元)'] = df['营业收入(元)'].apply(lambda x:
float(x.replace('万', ''))
* 10000 if '万' in x else
float(x.replace('亿', '')) * 100000000)
        df = df[['年份', '营业收入(元)']]
        path = '数据处理结果/' + str(name) + '.csv'
        df.to_csv(path, encoding='gbk')
        print(df)

if __name__ == "__main__":
    #a = DEAL(input('请输入股票名称:')).run()
    a = DEAL('朗姿股份').run()
```

(1) 定义了一个名为 DEAL 的类, 其中包含构造函数 __init__ 和方法 run。

(2) 构造函数 __init__():

- 该函数接受一个参数 name, 并将其存储在实例属性 self.name 中。
- 这个参数代表股票的名称。

(3) 方法 run: 读取股票数据的 CSV 文件, 进行数据处理, 并将处理后的数据保存到新的 CSV 文件中。

- 构建文件路径: 使用传入的股票名称构建数据文件的路径。
- 使用 pd.read_csv 方法读取 CSV 文件为 DataFrame。

(4) 数据处理:

- 删除空值行: 使用 `df.dropna()` 方法删除 DataFrame 中的空值行。
- 删除营业收入为'--'的行: 使用 `df.drop(index=df[df['营业收入(元)'] == '--'].index)` 来删除“营业收入(元)”列中值为'--'的行。
- 转换营业收入为数值: 对“营业收入(元)”列的数据进行处理, 将'万'转换为 10000, 将'亿'转换为 100000000, 并将数据类型转换为浮点数。

(5) 选择列:

- 选择保留“年份”和“营业收入(元)”两列的数据: 使用 `df[['年份', '营业收入(元)']]` 选择指定的列。

(6) 保存数据:

- 构建处理结果保存路径: 构建新的数据文件路径。
- 使用 `df.to_csv` 方法将处理后的数据保存为新的 CSV 文件, 指定编码为'gbk'。

(7) 主程序: 在 `__main__` 部分, 创建一个 DEAL 类的实例, 并调用其 `run` 方法来处理名为“朗姿股份”的股票数据。

3、数据分析-时间序列预测

需求说明:

- 1、运用图形法和 ADF 检验法检验时间序列的平稳性, 并输出平稳性检验结果;
- 2、对不平稳的时间序列, 运用差分变换, 并输出差分后图形;
- 3、寻找最优 pq 值, 并输出结果;
- 4、检测拟合效果;
- 5、运用时间序列模型进行预测。

步骤 1: 读取数据

然后运行代码, 代码将会将年份列时间序列化并设置为索引。

```
import numpy as np
import statsmodels.api as sm
from statsmodels.tsa.stattools import adfuller
import warnings
import pandas as pd
import matplotlib.pyplot as plt
from statsmodels.tsa.arima.model import ARIMA
```

```
from statsmodels.stats.diagnostic import acorr_ljungbox
```

(1) numpy: Python 的一个库, 支持大量的维度数组与矩阵运算, 此外也针对数组运算提供大量的数学函数库。

(2) statsmodels.api: statsmodels 是一个 Python 模块, 提供了许多统计模型、统计检验和数据探索的功能, 这里通过 api 接口导入, 方便后续使用。

(3) from statsmodels.tsa.stattools import adfuller: 从 statsmodels 的 tsa.stattools 模块中导入 adfuller 函数, 该函数用于进行 ADF (Augmented Dickey-Fuller) 单位根检验, 以判断时间序列数据是否是平稳的。

(4) warnings: Python 的警告控制模块, 可以用来控制警告信息的输出。

(5) pandas: 一个开源的、BSD 许可的库, 提供了高性能、易用的数据结构和数据分析工具。

(6) matplotlib.pyplot: matplotlib 的 pyplot 模块提供了一个类似于 MATLAB 的绘图系统, 是 Python 中常用的绘图库之一。

(7) from statsmodels.tsa.arima.model import ARIMA: 从 statsmodels 的 tsa.arima.model 模块中导入 ARIMA 类, 用于构建 ARIMA (自回归积分滑动平均) 模型。

(8) from statsmodels.stats.diagnostic import acorr_ljungbox: 从 statsmodels 的 stats.diagnostic 模块中导入 acorr_ljungbox 函数, 该函数用于进行 Ljung-Box 检验, 以检验时间序列数据中是否存在自相关性。

```
plt.rcParams['font.sans-serif'] = ['SimHei'] # 用来正常显示中文标签
plt.rcParams['axes.unicode_minus'] = False # 用来正常显示负号
```

(9) plt.rcParams['font.sans-serif'] = ['SimHei']: 这行代码配置了 matplotlib 的字体设置, 使其能够显示中文。'SimHei' 是字体名称, 表示使用黑体显示中文。

(10) plt.rcParams['axes.unicode_minus'] = False: 这行代码配置了 matplotlib 的负号显示方式, 避免在使用中文时出现方块或其他异常显示。

```
class MOD():
    def __init__(self, name) -> None:
        self.name = name

    def read(self): # 读取数据, 调整索引
        name = self.name
        path = '数据处理结果/' + str(name) + '.csv'
        df = pd.read_csv(path, encoding='gbk')
        df = df[['年份', '营业收入(元)']]
        df['年份'] = pd.to_datetime(df['年份'])
```

```
df.set_index("年份", inplace=True)
df["营业收入(元)"] = df["营业收入(元)"].astype(float)
self.df = df
```

(1) 定义了一个名为 MOD 的类。

(2) def __init__(self, name) -> None:

①__init__方法是类的初始化方法，当创建类的新实例时自动调用。

②它接受一个参数 name，这个参数被用来初始化实例的 name 属性。

③方法的返回类型注解为 None，表明这个方法不返回任何值（实际上，在 Python 中，如果方法没有明确的返回值，它会隐式地返回 None）

(3) def read(self): read 方法是一个实例方法，它不接受除了 self 之外的任何参数。self 是对类实例自身的引用，允许访问实例的属性和其他方法。方法的主要目的是读取与实例的 name 属性相关联的 CSV 文件，并对其进行预处理。

(4) name = self.name : 从实例的 name 属性中获取公司名称或股票代码。

(5) path = '数据处理结果/' + str(name) + '.csv' : 构造 CSV 文件的路径，这里假设所有 CSV 文件都存储在名为数据处理结果的文件夹中，文件名由 name 和.csv 扩展名组成。

(6) df = pd.read_csv(path, encoding='gbk') : 使用 pd.read_csv() 函数读取 CSV 文件，指定 encoding='gbk' 以处理可能存在的中文编码问题（这通常是在处理中国内地的数据时需要的，因为 GBK 是一种常用的中文字符编码）。

(7) df = df[['年份', '营业收入(元)']] : 从读取的 DataFrame 中选择特定的列（这里只选择年份和营业收入(元)两列）。

(8) 将年份列转换为 datetime 类型，这对于时间序列分析非常重要，因为它允许根据时间进行索引和排序。

(9) 将年份列设置为 DataFrame 的索引。这里使用了 inplace=True 参数，意味着这个操作会直接修改原始的 DataFrame，而不是返回一个新的 DataFrame。

(10) 将营业收入(元)列的数据类型转换为 float，以便进行数值计算。

步骤 2: 平稳性检验

定义一个名为 pingwen 的方法，目的是对时间序列数据进行平稳性检验，包括通过图形法和 ADF (Augmented Dickey-Fuller) 检验法。

```
def pingwen(self): # 检验平稳性，分图形法和 ADF 检验法
    df = self.df
    plt.figure(facecolor='white', figsize=(15, 10))
```

```

plt.plot(df.index, df['营业收入(元)'], label='Time Series')
plt.legend(loc='best')
plt.show()
a = adfuller(df["营业收入(元)"], autolag='AIC')[1]
print('平稳性检验结果为: ', a)
if a > 0.05:
    self.chafen()

```

(1) `df = self.df`: 这行代码从类的实例中获取之前通过 `read` 方法加载并处理过的 `DataFrame`, 该 `DataFrame` 包含时间序列数据。

图形法检验平稳性:

(2) 使用 `matplotlib` 库创建一个图形窗口, 设置背景色为白色, 并指定图形的大小为 15x10 英寸。

(3) 通过 `plt.plot(df.index, df['营业收入(元)'], label='Time Series')` 绘制时间序列图, 其中 `x` 轴是 `DataFrame` 的索引 (在这里是时间), `y` 轴是 '营业收入(元)' 列的值。

(4) 使用 `plt.legend(loc='best')` 添加图例, 并自动选择最佳位置显示。

(5) 使用 `plt.show()` 显示图形。这样, 用户就可以直观地观察时间序列是否有明显的趋势或季节性变化, 从而初步判断其是否平稳。

ADF 检验法:

(1) 使用 `adfuller` 函数从 `statsmodels.tsa.stattools` 模块对 '营业收入(元)' 列进行 ADF 单位根检验。这个函数返回一个包含多个统计量的元组, 其中第一个元素是 ADF 统计量, 第二个元素是 `p` 值 (即我们关心的 `a`)。

(2) `autolag='AIC'` 参数指示 `adfuller` 函数自动选择滞后阶数, 以最小化 AIC (赤池信息量准则) 值。

(3) `print('平稳性检验结果为: ', a)`: 打印出 ADF 检验的 `p` 值。在 ADF 检验中, 如果 `p` 值小于显著性水平 (通常是 0.05), 则拒绝原假设 (即时间序列存在单位根, 不平稳)。但在这段代码中, `p` 值被用来直接判断平稳性, 这是不正确的。正确的做法应该是检查 `p` 值是否小于显著性水平。

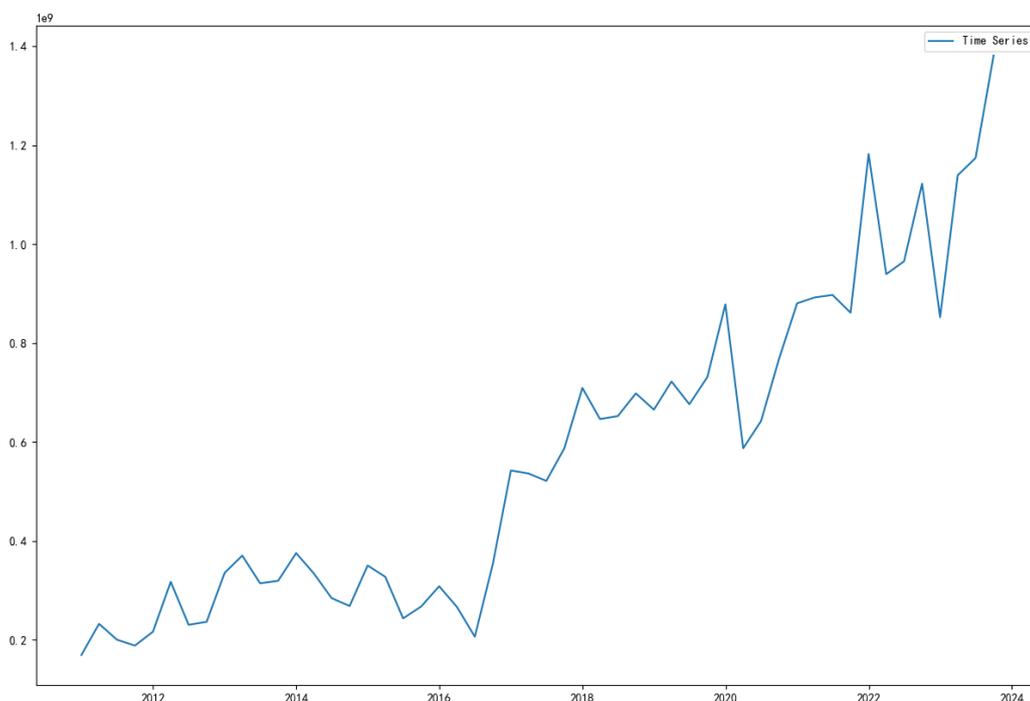
非平稳性处理:

如果 `p` 值大于 0.05 (即 ADF 检验未能拒绝原假设, 认为时间序列可能不平稳), 则调用 `self.chafen()` 方法。

运行结果:

(1) 图示法

运行“平稳性检验”中的“图示法”代码，代码会绘制一个营业收入关于时间的折线图，观察折线图，发现指标有明显地随着时间上升的趋势，序列可能是非平稳的。



(2) ADF 法

进一步进行 ADF 检验验证猜想，运行“adfuller 统计量”代码，输出结果为统计量的 P 值，当 P 值小于 0.05 时，说明序列为平稳序列，反之为非平稳序列。本例中输出结果 0.74，大于 0.05 所以再一次印证了序列为非平稳的。

平稳性检验结果为: 0.744526277301529

步骤 3: 差分变换

对于不平稳的序列，我们要进行差分变换消除其随机影响，运行“差分变换”部分代码，代码会对序列进行一阶差分变换，并输出一阶差分变换的 ADF 检验结果，每运行一次即进行一次差分变换。

```
def chafen(self): # 进行差分，并输出差分后图形
    df = self.df
    df_p = df.iloc[0:3, :] # 留下最近三年
    df_t = df.iloc[3:100, :] # 其他用来训练训练
    train_diff = df_t.copy() # 差分训练集
    for i in range(10):
```

```

train_diff['营业收入(元)'] = train_diff['营业收入
(元)'].diff() # 差分
train_diff = train_diff.dropna() # 去除空值
b = adfuller(train_diff["营业收入(元)"], autolag='AIC')[1] #
差分后再 ADF 检验
pvalue = float(acorr_ljungbox(train_diff['营业收入(元)'],
lags=1)['lb_pvalue'])
print(f'第{i + 1}次差分后的 ADF 检验的 p 值为{b}')
print(f'第{i + 1}次差分后的白噪声检验结果的 p 值为{pvalue}')
if b < 0.05 and pvalue < 0.05:
    print('已经符合平稳序列并且非白噪声序列')
    print(f'差分次数为{i + 1}')
    break
if pvalue > 0.05:
    print('数据无法满足要求')
    break

self.chafen_times = i + 1

```

数据分割:

(1) 定义了一个名为 `chafen` 的方法，旨在通过对时间序列数据进行差分处理，并检查差分后的数据是否满足平稳性和非白噪声的条件。

(2) `df = self.df`: 从类的实例中获取 DataFrame。

(3) `df_p = df.iloc[0:3, :]`: 选择前 3 行数据作为测试集。

(4) `df_t = df.iloc[3:100, :]`: 选择从第 4 行到第 100 行的数据作为训练集。这里同样假设了 DataFrame 的索引是有意义的，并且第 4 行到第 100 行包含了足够的用于训练的数据。

差分处理:

(1) 在一个循环中，最多进行 10 次差分处理。

(2) `train_diff['营业收入(元)'] = train_diff['营业收入(元)'].diff()`: 对'营业收入(元)'列进行一阶差分。差分是时间序列分析中常用的一种技术，用于消除趋势和季节性，使数据更接近平稳。

(3) `train_diff = train_diff.dropna()`: 每次差分后，都会去除由于差分操作而产生的 NaN 值。

平稳性检验:

(1) 使用 ADF (Augmented Dickey-Fuller) 检验来检查差分后的数据是否平稳。如果 ADF 检验的 p 值小于 0.05，则认为时间序列在统计上显著地拒绝了存在单位根的原假设，即时间序列是平稳的。

白噪声检验:

(1)使用 Ljung-Box 检验(通过 `acorr_ljungbox` 函数)来检查差分后的数据是否是白噪声。白噪声是指时间序列中的各个随机变量之间互不相关,即没有自相关性。如果 Ljung-Box 检验的 p 值小于 0.05,则认为时间序列存在自相关性,不是白噪声。

条件判断和输出结果:

(1) 如果差分后的数据既满足平稳性条件 (ADF 检验的 p 值小于 0.05) 又满足非白噪声条件 (Ljung-Box 检验的 p 值小于 0.05), 则输出相应的信息, 并记录差分次数。

(2) 如果 Ljung-Box 检验的 p 值大于 0.05, 即数据是白噪声, 则输出“数据无法满足要求”并中断循环。

存储差分次数:

`self.chafen_times = i + 1`: 将满足条件的差分次数存储在类的实例变量中。如果循环正常结束 (即没有因为满足条件而中断), 那么这将记录最后一次尝试的差分次数。

```
plt.figure(facecolor='white', figsize=(15, 8))
plt.plot(train_diff.index,
         train_diff['营业收入(元)'], label='Time Series after
diff')
plt.legend(loc='best')
plt.show()
```

(1) 这段代码使用了 `matplotlib` 库来绘制一个时间序列图, 具体是对差分处理后的‘营业收入(元)’数据进行可视化。

```
# 绘制自相关图和偏自相关图
fig, axes = plt.subplots(2, 1, figsize=(15, 8))
sm.graphics.tsa.plot_acf(train_diff['营业收入(元)'], lags=20,
ax=axes[0])
sm.graphics.tsa.plot_pacf(train_diff['营业收入(元)'], lags=20,
ax=axes[1])
plt.show()
```

```
self.df_p = df_p
self.df_t = df_t
self.train_diff = train_diff
```

绘制自相关图和偏自相关图:

(1) `fig, axes = plt.subplots(2, 1, figsize=(15, 8))`: 这行代码使用 `matplotlib` 的 `subplots` 函数创建了一个包含两个子图的图形。2, 1 表示图形有 2 行 1 列, 即垂直排列的两个子图。`figsize=(15, 8)` 设置了整个图形的大小。`fig` 是整个图形的对象, 而 `axes` 是一

个包含两个子图轴 (Axes) 对象的数组。

(2) `sm.graphics.tsa.plot_acf(train_diff['营业收入(元)'], lags=20, ax=axes[0])`: 这行代码使用 `statsmodels` (通常缩写为 `sm`) 库中的 `plot_acf` 函数来绘制 `train_diff` `DataFrame` 中 '营业收入(元)' 列的自相关图。`lags=20` 参数指定了计算自相关系数的滞后阶数, 这里设置为 20 阶。`ax=axes[0]` 参数指定了自相关图应该绘制在哪个子图上, 这里是第一个子图 (垂直排列的上方子图)。

(3) `sm.graphics.tsa.plot_pacf(train_diff['营业收入(元)'], lags=20, ax=axes[1])`: 与上一行代码类似, 这行代码使用 `plot_pacf` 函数来绘制偏自相关图。偏自相关图用于帮助识别时间序列中的 AR(自回归)模型的阶数。同样地, `lags=20` 指定了滞后阶数, 而 `ax=axes[1]` 指定了绘制偏自相关图的子图 (垂直排列的下方子图)。

(4) `plt.show()`: 这行代码用于显示整个图形, 包括两个子图: 自相关图和偏自相关图。

存储 DataFrame 对象:

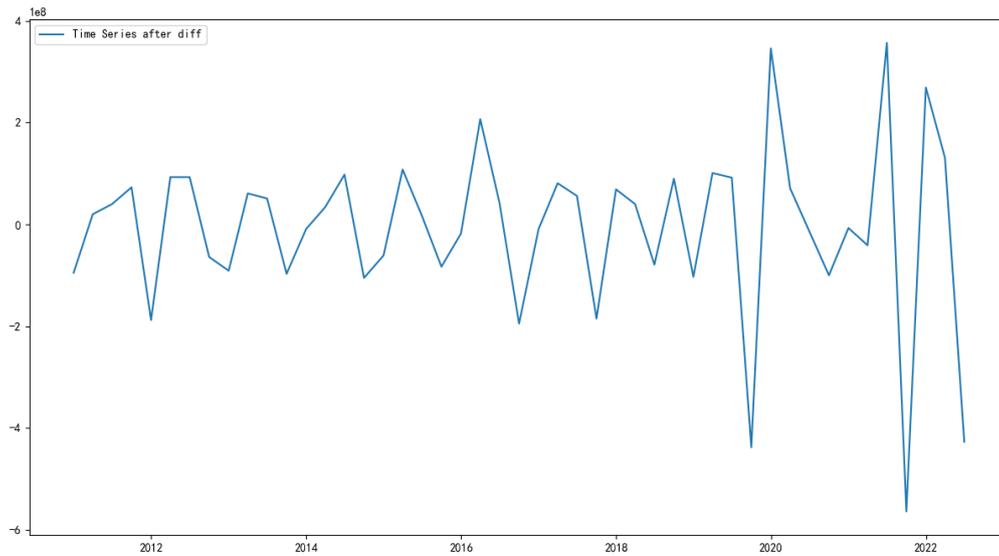
(1) `self.df_p = df_p`: 将 `df_p` `DataFrame` 对象存储在类的实例变量 `df_p` 中。

(2) `self.df_t = df_t`: 将 `df_t` `DataFrame` 对象存储在类的实例变量 `df_t` 中。

(3) `self.train_diff = train_diff`: 将差分后的训练数据集 `train_diff` 存储在类的实例变量 `train_diff` 中。

运行结果:

(1) 从本例的结果来看, 第 2 次差分后的 ADF 检验的 `p` 值为 0.000531675345835194, 远远小于 0.05, 所以模型已经平稳, 输出差分后图像可以看出已无随时间上涨的现象。



(2) 第 2 次差分后的白噪声检验结果的 p 值为 0.0011759742911574406，明显小于 0.05，说明序列为非白噪声序列，可以进行建模。

差分变换结果：

第 1 次差分后的 ADF 检验的 p 值为 0.2595621038194392

第 1 次差分后的白噪声检验结果的 p 值为 0.02357795299685775

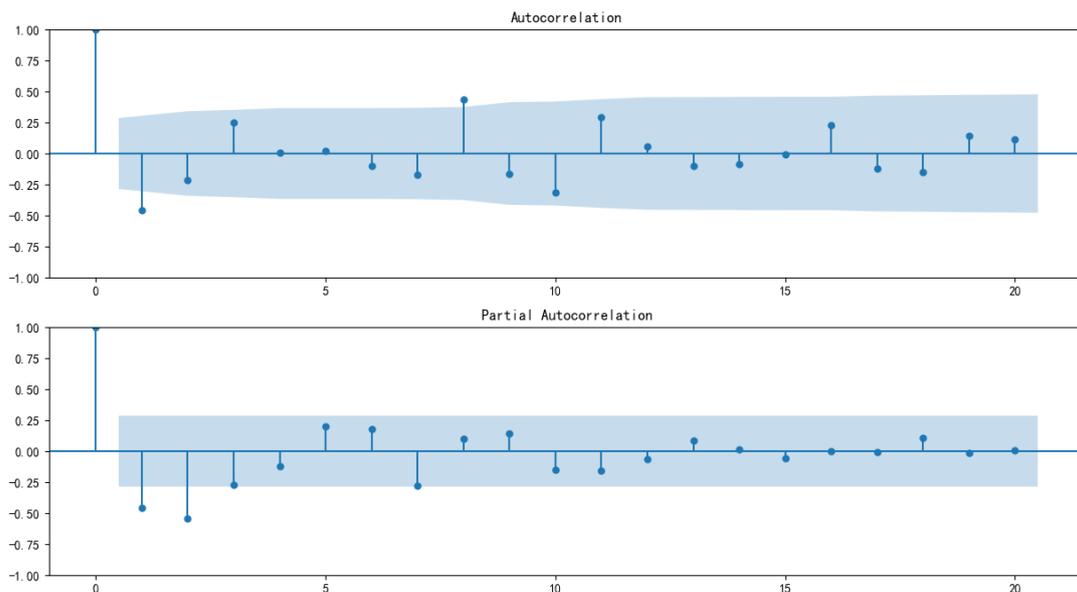
第 2 次差分后的 ADF 检验的 p 值为 0.000531675345835194

第 2 次差分后的白噪声检验结果的 p 值为 0.0011759742911574406

已经符合平稳序列并且非白噪声序列

差分次数为 2

(3) 自相关和偏自相关图：



从自相关和偏自相关图上可知，P 值为 1，q 值为 3。输入到代码运行结果中并按 enter 键。

步骤 4：定阶

ARMA 模型有三个参数 p 、 i 、 q ，其中 i 为差分次数，在上一个步骤中已经确定为 2。上一步骤当中由自相关和偏自相关图反应的 p 值为 1， q 值为 3，下面运用信息准则方法运行“定阶”部分代码，代码会通过暴力枚举的方式找出合适的 p 、 q 值。

```
def dingjie(self): # 并寻找最优 pq 值
    train_diff = self.train_diff
    df_t = self.df_t
    t = self.chafen_times
    warnings.filterwarnings("ignore")
    pmax = 8
    qmax = 8
    aic_matrix = [] # aic 矩阵
    for p in range(1, pmax + 1):
        tmp = []
        for q in range(1, qmax + 1):
            try: # 存在部分报错，所以用 try 来跳过报错。
                model = ARIMA(
                    endog=df_t['营业收入(元)'], order=(p, t, q))
                results = model.fit()
                tmp.append(results.aic)
                print('ARIMA p:{} q:{} - AIC:{}'.format(p, q,
                    results.aic))
            except:
```

```

        tmp.append(None)
    aic_matrix.append(tmp)
    aic_matrix = pd.DataFrame(aic_matrix) # 从中可以找出最小值
    p_aic, q_aic = aic_matrix.stack().idxmin() # 先用 stack 展平, 然后用 idxmin 找出最小值位置。
    print(u'AIC 最小的 p 值和 q 值为: %s、%s' % (p_aic + 1, q_aic + 1))
    self.p_aic = p_aic + 1 # 由于 range(1, pmax+1), 所以需要 +1
    self.q_aic = q_aic + 1

```

初始化和参数设置:

(1) train_diff、df_t、t: 这些是从类的实例变量中获取的, 分别代表差分后的训练数据集、原始训练数据集和差分次数。

(2) pmax 和 qmax: 分别设置为 8, 表示在寻找最佳 p 和 q 值时, 它们的最大测试值。

(3) aic_matrix: 一个空列表, 用于存储不同 p 和 q 组合下的 AIC (赤池信息量准则) 值。

遍历所有 p 和 q 组合:

(1) 使用两层嵌套循环遍历所有可能的 p 和 q 值 (从 1 到 pmax 和 qmax)。

(2) 对于每一对 p 和 q, 尝试创建一个 ARIMA 模型并拟合数据。这里使用了 try-except 语句来捕获并忽略任何可能发生的错误 (例如, 如果某些 p 和 q 组合导致模型拟合失败)。

(3) 如果模型成功拟合, 则将其 AIC 值添加到当前 p 值的临时列表 tmp 中。

如果发生错误, 则在该位置添加 None。

构建 AIC 矩阵并找到最小值:

(1) 将所有 AIC 值 (包括可能的 None 值) 存储在 aic_matrix 列表中, 并将其转换为 DataFrame 以便更容易地处理。

(2) 使用 stack() 方法将 DataFrame 展平为一维 Series, 然后使用 idxmin() 方法找到 AIC 值最小的索引 (这是 p 和 q 的组合)。

(3) 通过索引计算出最小 AIC 值对应的 p 和 q 值 (注意要加 1, 因为循环是从 1 开始的, 而索引是从 0 开始的)。

运行结果:

AIC 最小的 p 值和 q 值为: 1、4。

步骤 5: 检测拟合效果

可以对模型进行差分还原来检查模型的拟合效果, 运行“一阶差分还原”代码, 代码会输出原始数据与预测数据的折线图。

```
# 用户手动输入的 p 和 q 值
```

```
self.p_manual = int(input("请输入自相关图显示的 p 值: "))
```

```
self.q_manual = int(input("请输入偏自相关图显示的 q 值: "))
```

```
print(u'根据 ACF 和 PACF 图选择的 p 值和 q 值为:%s,%s' % (self.p_manual, self.q_manual))
```

用户手动输入:

(1) 通过 `input()` 函数获取用户输入的 `p` 和 `q` 值, 并将它们存储在类的实例变量 `self.p_manual` 和 `self.q_manual` 中。

输出和存储结果:

(1) 打印出基于 AIC 准则和用户手动选择的 `p` 和 `q` 值。

(2) 将基于 AIC 准则选择的 `p` 和 `q` 值(加 1 后)存储在类的实例变量 `self.p_aic` 和 `self.q_aic` 中。

运行结果:

AIC 最小的 `p` 值和 `q` 值为: 1、4

请输入自相关图显示的 `p` 值: 1

请输入偏自相关图显示的 `q` 值: 3

根据 ACF 和 PACF 图选择的 `p` 值和 `q` 值为: 1、3

```
def huanyuan(self): # 检测拟合效果
```

```
df_t = self.df_t
```

```
t = self.chafen_times
```

(1) 定义一个名为 `huanyuan` 的方法, 用于比较基于 AIC 准则自动选择的 ARIMA 模型参数 (`p_aic`, `q_aic`) 与用户手动选择的 ARIMA 模型参数 (`p_manual`, `q_manual`) 的拟合效果。

(2) 从类的实例变量中获取 `df_t` (原始训练数据集) 和 `chafen_times` (差分次数)。

```
# 使用 AIC 确定的 p 和 q 值进行建模
```

```
model_aic = ARIMA(endog=df_t['营业收入(元)'], order=(self.p_aic, t, self.q_aic))
```

```
result_ARIMA_aic = model_aic.fit()
```

```
predict_diff_aic = result_ARIMA_aic.predict()
```

```
predict_aic = predict_diff_aic.shift(-1)
```

(1) 创建一个 ARIMA 模型 `model_aic`, 使用基于 AIC 准则自动选择的 `p_aic` 和 `q_aic` 值作为参数。

(2) 拟合模型并获取预测结果 `predict_diff_aic`。注意, `predict_diff_aic` 是差分后的预测值, 因为 ARIMA 模型是对差分后的序列进行建模的。

(3) 使用 `shift(-1)` 尝试将预测值向前移动一位以与原始序列对齐，但这是不正确的做法，因为 `shift(-1)` 会将 NaN 值置于预测序列的开头，并且无法直接用于与原始序列比较。正确的做法应该是先对预测结果进行反差分（如果原始数据经过了差分处理），然后再与原始数据比较。

```
# 使用用户手动输入的 p 和 q 值进行建模
model_manual = ARIMA(endog=df_t['营业收入(元)'],
order=(self.p_manual, t, self.q_manual))
result_ARIMA_manual = model_manual.fit()
predict_diff_manual = result_ARIMA_manual.predict()
predict_manual = predict_diff_manual.shift(-1)
```

(1) 创建并拟合另一个 ARIMA 模型 `model_manual`，使用用户手动选择的 `p_manual` 和 `q_manual` 值作为参数。

(2) 获取预测结果 `predict_diff_manual`，并尝试通过 `shift(-1)` 向前移动。

```
plt.figure(figsize=(15, 8), facecolor='white')
predict_aic.plot(color='blue', label='Predict (AIC)')
predict_manual.plot(color='green', label='Predict (Manual)')
df_t['营业收入(元)'].sort_index().plot(color='red',
label='Original')
err_aic = np.sqrt(np.mean((predict_aic.dropna() - df_t['营业收入(元)'].dropna()) ** 2))
err_manual = np.sqrt(np.mean((predict_manual.dropna() - df_t['营业收入(元)'].dropna()) ** 2))
plt.legend(loc='best')
plt.title('Error (AIC): %.4f, Error (Manual): %.4f' % (err_aic,
err_manual))
plt.show()
```

(1) 使用 `matplotlib` 绘制图形，展示原始数据、基于 AIC 的预测结果和用户手动选择的预测结果。

(2) 计算并显示两种预测方法的均方根误差（RMSE），以评估它们的拟合效果。这里，误差是通过计算预测值与原始数据之间的差异的平方的平均值的平方根来得到的。

(3) 注意：由于 `predict_aic` 和 `predict_manual` 可能包含 NaN 值（由于 `shift(-1)` 的使用），因此在计算误差之前需要先使用 `dropna()` 方法删除这些 NaN 值。

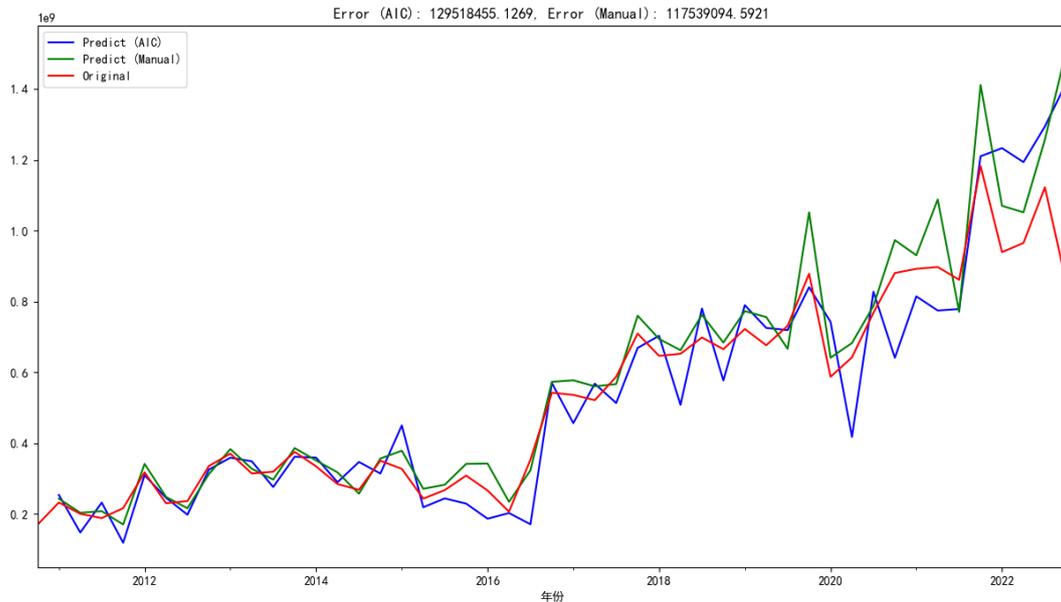
```
print(f"AIC 模型的误差: {err_aic}")
print(f"手动选择模型的误差: {err_manual}")
```

运行结果：

AIC 模型的误差: 129518455.12690397

手动选择模型的误差: 117539094.59205753

本例中两条线都与原始数据基本贴合, 拟合效果较好。



步骤 6: 模型选择

基于预测结果的准确性 (通过误差来衡量) 来自动选择最佳的 ARIMA 模型参数。

```
# 选择误差较小的模型
if np.isnan(err_aic):
    print("由于 AIC 模型的误差为 NaN, 选择手动输入的模型进行预测")
    self.p = self.p_manual
    self.q = self.q_manual
elif np.isnan(err_manual):
    print("由于手动选择模型的误差为 NaN, 选择 AIC 模型进行预测")
    self.p = self.p_aic
    self.q = self.q_aic
elif err_aic < err_manual:
    print("选择 AIC 模型进行预测")
    self.p = self.p_aic
    self.q = self.q_aic
else:
    print("选择手动输入的模型进行预测")
    self.p = self.p_manual
    self.q = self.q_manual
```

检查误差是否为 NaN:

(1) 检查两种模型的误差 (`err_aic` 和 `err_manual`) 是否为 NaN (即不是一个有效的数值)。如果基于 AIC 的模型误差是 NaN, 那么打印一条消息说明这一点, 并选择用户手动输入的模型参数 (`self.p_manual` 和 `self.q_manual`) 用于后续的预测。

(2) 如果手动选择模型的误差是 NaN, 那么打印一条消息, 并选择基于 AIC 的模型参数 (`self.p_aic` 和 `self.q_aic`)。

比较误差大小:

(1) 如果两种模型的误差都不是 NaN, 代码接下来会比较这两个误差值。

(2) 如果基于 AIC 的模型的误差 (`err_aic`) 小于手动选择模型的误差 (`err_manual`), 则打印一条消息说明选择了 AIC 模型, 并将该模型的参数 (`self.p_aic` 和 `self.q_aic`) 设置为类的实例变量 `self.p` 和 `self.q`, 以便后续使用。否则 (即如果 `err_aic` 不小于 `err_manual`), 打印一条消息说明选择了手动输入的模型, 并将该模型的参数设置为类的实例变量。

运行结果:

选择手动输入的模型进行预测

步骤 7: 进行预测

运行“预测”部分代码, 输出最近三个季度的预测值和这三个季度的真实值, 以及二者之差和真实值的比例, 以及预测结果的平均误差。

```
def yuce(self): # 进行预测
    df_t = self.df_t
    df_p = self.df_p
    p = self.p
    q = self.q
    t = self.chafen_times
    model = ARIMA(endog=df_t['营业收入(元)'].sort_index(), order=(p, t,
q))
    result_ARIMA = model.fit()
    y_forecasted = result_ARIMA.forecast(3)
    y_forecasted.index = df_p.index # 确保索引一致
    y_truth = df_p['营业收入(元)'].sort_index()
    error_rate = (abs(np.array(y_forecasted) -
np.array(y_truth)) / y_truth).mean()
    print("=====")
    print('预测结果\n', y_forecasted)
```

```

print("=====
")
print('真实结果\n', y_truth)
print("=====
")
print('差距比例\n', (y_forecasted - y_truth) / y_truth)
print("=====
")
print('平均差距比例: {}'.format(round(error_rate, 4)))
print("=====
")

```

初始化:

- (1) df_t: 包含训练数据的 DataFrame, 用于拟合 ARIMA 模型。
- (2) df_p: 包含预测目标 (通常是未来时间段的数据) 的 DataFrame, 用于与预测结果进行比较。
- (3) p 和 q: 之前通过某种方式 (如 AIC 准则或用户手动选择) 确定的 ARIMA 模型的参数。
- (4) t: 差分次数, 用于将非平稳数据转换为平稳数据以进行建模。

建模:

- (1) 使用 ARIMA 类创建一个 ARIMA 模型实例, 其中 endog 参数设置为训练数据 df_t['营业收入(元)'].sort_index() (确保索引是有序的)。order 参数设置为 (p, t, q), 即之前确定的 ARIMA 模型的参数。
- (2) 调用 fit 方法拟合模型, 得到拟合结果 result_ARIMA。

预测:

- (1) 使用 forecast 方法对未来的数据点进行预测, 预测期数设置为 3 (这意味着将预测未来三个时间点的值)。forecast 方法返回的是一个包含预测值的数组, 默认索引是从 0 开始的整数索引。
- (2) 将预测值的索引设置为与 df_p (即预测目标数据的 DataFrame) 的索引一致, 以便后续能够正确地将预测值与实际值进行比较。

计算误差:

- (1) 提取 df_p 中的实际值 y_truth, 并同样确保索引是有序的。
- (2) 计算预测值 y_forecasted 与实际值 y_truth 之间的绝对误差, 并将这个误差数组除以实际值数组, 得到每个预测点的误差比例。
- (3) 计算这些误差比例的平均值 error_rate, 作为预测结果的总体误差度量。

输出结果:

(1) 运用 print, 打印出预测结果、真实结果、差距比例以及平均差距比例, 以使用户评估模型的预测性能。

输入要分析的公司名称:

```
if __name__ == '__main__':  
    a = MOD('朗姿股份') # a = MOD('苏泊尔')  
    a.read()  
    a.pingwen()  
    a.dingjie()  
    a.huanyuan()  
    a.yuce()
```

运行结果:

```
=====
```

预测结果

年份	
2023-09-30	9.741832e+08
2023-06-30	1.025552e+09
2023-03-31	1.023983e+09

Name: predicted_mean, dtype: float64

```
=====
```

真实结果

年份	
2023-03-31	1.139000e+09
2023-06-30	1.174000e+09
2023-09-30	1.381000e+09

Name: 营业收入(元), dtype: float64

```
=====
```

差距比例

年份	
2023-03-31	-0.100981
2023-06-30	-0.126446
2023-09-30	-0.294581

Freq: QE-DEC, dtype: float64

```
=====
```

平均差距比例: 0.1766

```
=====
```

可以看出平均误差只有 0.1766, 模型的预测效果较好。

项目十：社交网络审计应用

实训 1：识别关联方

实训步骤

（一）数据收集

在大数据时代，识别企业关联方可以从多个渠道获取信息，包括但不限于以下几种：

（1）公开信息数据库。利用公开的企业注册信息数据库、财务报表数据库等，查找企业的关联方信息。

（2）社交媒体和网络分析。通过分析企业高管、员工在社交媒体上的互动关系、网络上的信息发布等，识别可能的关联方。

（3）商业关系数据库。利用商业数据库，如 Dun & Bradstreet、LexisNexis 等，获取企业与其他实体之间的商业关系信息。

（4）金融数据。分析企业的金融交易记录，包括资金流动、股权交易等，识别可能的关联方。

（5）供应链和合作伙伴。分析企业的供应链关系和合作伙伴关系，了解可能的关联方。

（6）舆情监测。通过舆情监测工具，跟踪媒体报道、新闻事件等，获取与企业相关的关联方信息。

（7）专业调查和调研。进行专业调查和调研，采集企业关联方的信息，包括股东信息、高管信息等。

本案例数据主要是根据公司年报以及企查查等公开平台的数据整理得到。

1、节点层级表

0 层级表示要分析的审计对象；1 层级表示分析对象的股东名称；2 层级表示审计对象投资的子公司或者参股公司。

层级	节点
0	朗姿股份有限公司
1	申东日

1	申今花
1	朗姿股份有限公司-第三期员工持股计划
1	上海烜鼎资产管理有限公司-烜鼎长红七号私募证券投资基金
1	上海烜鼎资产管理有限公司-烜鼎长红六号私募证券投资基金
1	香港中央结算有限公司
1	银华基金-中国人寿保险股份有限公司一分红险-银华基金国寿股份均衡股票型组合单一资产管理计划（可供出售）
1	赵志海
1	钱密林
1	北京合源融微股权投资中心（有限合伙）
2	北京卓可服装有限公司
2	天津朗姿供应链管理有限公司
2	朗姿医疗管理有限公司
2	朗姿国际贸易有限公司
2	上海永初贸易有限公司
2	西藏哗叽服饰有限公司
2	成都朗姿商业发展有限公司
2	北京朗姿服饰有限公司
2	广州若羽臣科技股份有限公司
2	北京朗姿韩亚资产管理有限公司
2	北京麦可利商业管理有限公司
2	阿咖邦贸易（北京）有限公司
...	...

2、投资信息表

全面梳理出与审计对象相关的母子公司的投资或参股公司，以及子公司的股东、投资比例等。

母公司	子公司	投资比例
朗姿股份有限公司	北京卓可服装有限公司	100.00
朗姿股份有限公司	天津朗姿供应链管理有限公司	100.00
朗姿股份有限公司	朗姿医疗管理有限公司	100.00
朗姿股份有限公司	朗姿国际贸易有限公司	100.00
朗姿股份有限公司	上海永初贸易有限公司	100.00
朗姿医疗管理有限公司	四川米兰柏羽医学美容医院有限公司	100.00
朗姿医疗管理有限公司	陕西高一生医疗美容医院有限公司	100.00
朗姿股份有限公司	西藏哗叽服饰有限公司	100.00
朗姿医疗管理有限公司	昆明米兰柏羽医疗美容医院有限公司	100.00
朗姿医疗管理有限公司	西安市新城区晶肤医疗美容有限公司	100.00
朗姿医疗管理有限公司	青岛米兰柏羽整形美容医院有限公司	100.00
朗姿股份有限公司	成都朗姿商业发展有限公司	100.00
朗姿医疗管理有限公司	西安未央晶肤医疗美容诊所有限公司	100.00
朗姿医疗管理有限公司	成都青羊光华晶肤医疗美容诊所有限公司	100.00

四川米兰柏羽医学美容医院有限公司	成都米兰柏羽美容服务有限公司	100.00
朗姿医疗管理有限公司	长沙市芙蓉区晶肤医疗美容有限公司	100.00
朗姿医疗管理有限公司	成都翔丹企业管理咨询合伙企业（有限合伙）	100.00
朗姿医疗管理有限公司	重庆米兰柏羽医疗美容医院有限公司	100.00
朗姿医疗管理有限公司	四川晶肤医学美容医院有限公司	100.00
朗姿医疗管理有限公司	北京朗姿医疗管理有限公司	100.00
朗姿医疗管理有限公司	西藏朗姿医疗科技有限公司	100.00
朗姿医疗管理有限公司	成都锦江米兰柏羽医疗美容医院有限公司	100.00
武汉五洲整形外科医院有限公司	武汉五洲美容服务有限公司	100.00
北京朗姿医疗管理有限公司	武汉五洲整形外科医院有限公司	90.00
朗姿股份有限公司	北京朗姿服饰有限公司	74.87
服装控股有限公司（香港）	北京朗姿服饰有限公司	25.13
朗姿股份有限公司	广州若羽臣科技股份有限公司	10.54
朗姿股份有限公司	北京朗姿韩亚资产管理有限公司	29.23
朗姿股份有限公司	北京麦可利商业管理有限公司	20.00
北京朗姿韩亚资产管理有限公司	北京麦可利商业管理有限公司	35.00
阿卡邦株式会社	阿咖邦贸易（北京）有限公司	66.02
朗姿股份有限公司	阿咖邦贸易（北京）有限公司	33.98
北京朗姿服饰有限公司	北京朗姿服饰有限公司深圳分公司	100.00
申东日	朗姿股份有限公司	47.82
申金花	朗姿股份有限公司	6.76
朗姿股份有限公司-第三期员工持股计划	朗姿股份有限公司	2.13
上海烜鼎资产管理有限公司-烜鼎长红七号私募证券投资基金	朗姿股份有限公司	1.00
（仅提供部分）		

（二）绘制被审计单位社交网络图

1、导入库

首先导入了需要使用的库，包括 Pyecharts、Pandas、Numpy、pagerank 等。

```
from pyecharts.charts import Graph, Tab
from pyecharts import options as opts
from pyecharts.commons.utils import JsCode
import pandas as pd
import numpy as np
from fast_pagerank import pagerank
from fast_pagerank import pagerank_power
```

```

from pyecharts.components import Table
import networkx as nx
import time
from scipy import sparse
import operator

```

2、定义网络绘制函数

```

def networks(data1, data2):
    data = data1
    data = data.rename(columns={'母公司': '输出', '子公司': '输入'}) # 重
命名列标签

    project = [str(u) for u in list(set(data2['节点']))] # 项目标签

    categories = [str(u) for u in list(set(data['输出']))] # 网络节点, 无
重复项
    x = [str(u) for u in list(set(data['输入']))] # 节点补录
    xc = list(set(x + categories)) # 完整节点

```

(1) 定义了一个名为 networks 的函数，用于绘制被审计单位的社交网络。该函数接受两个参数 data1 和 data2，分别表示输入数据和节点数据。在函数中对输入的数据进行了一些处理，包括重命名列标签、获取项目标签、网络节点等。

(2) 使用 rename 方法重命名 data1 中的列，使其更具描述性。

(3) 从 data2 中提取项目标签。

(4) 从 data1 中提取所有独特的节点，并合并两个节点列表。

```

# 设定节点颜色和大小
nodes = []
for i in range(len(xc)):
    node = {'name': str(xc[i]), 'symbolSize': 30}
    if str(xc[i]) in project:
        node['itemStyle'] = {
            'color': 'rgb(70, 130, 180)',
            'borderColor': 'rgb(70, 130, 180)',
            'borderWidth': 2,
            'shadowBlur': 10,
            'shadowColor': 'rgba(0, 0, 0, 0.3)'
        }
    else:
        node['itemStyle'] = {
            'color': 'rgb(70, 160, 50)',
            'borderColor': 'rgb(70, 160, 50)',
            'borderWidth': 2,

```

```

        'shadowBlur': 10,
        'shadowColor': 'rgba(0, 0, 0, 0.3)'
    }
    nodes.append(node)

```

- (1) 创建了一个空列表 `nodes` 来存储节点信息。
- (2) 根据节点是否在项目标签中设置不同的颜色和样式。
- (3) `symbolSize` 用于设置节点大小, `itemStyle` 用于设置节点的颜色、边框颜色和其他样式。

```

#调整边的样式
links = [
    {
        'source': str(data.loc[i, '输入']),
        'target': str(data.loc[i, '输出']),
        'value': data.loc[i, '投资比例'],
        'lineStyle': {
            'color': 'rgba(200, 200, 200, 0.6)',
            'width': 1,
            'curveness': 0.3
        }
    }
    for j in range(len(data))
]

for i in range(len(data)):
    links.append(
        {'source': str(data.iloc[i, 1]), 'target': str(data.iloc[i, 0]),
        'value': data.loc[i, '投资比例']}) # 添加边

```

(1) 初始创建 `links` 列表, 用于存储边的信息, 每条边包含 `source` (起点)、`target` (终点)、`value` (投资比例) 和 `lineStyle` (线样式)。

(2) `lineStyle` 设置了边的颜色、宽度和曲率。

(3) 第二个循环用于添加额外的边信息。

```

# 绘制网络图
c = (
    Graph(init_opts=opts.InitOpts(width='800px', height='1600px'))
    .add(
        '',
        nodes,
        links,
        repulsion=2800, # 排斥力度
        linestyle_opts=opts.LineStyleOpts(curve=0.3, width=1), # 设置曲率, 宽度
        edge_symbol=['arrow'], # 设置标记为箭头

```

```

        edge_symbol_size=7, # 设置箭头大小
        label_opts=opts.LabelOpts(is_show=True, font_size=12), # 设置节点标签的显示, 与字体大小
        # 设置边的标签的显示, 位置和格式
        edge_label=opts.LabelOpts(
            is_show=True, position='middle', formatter='{c}%'
        ),
    )
    # 图的全局参数, 包括标题和工具箱的选项
    .set_global_opts(title_opts=opts.TitleOpts(title='关系网络图'),
                    # 设置工具箱的参数, 包括工具箱的方向, 位置和保存图片的功能
                    toolbox_opts=opts.ToolboxOpts(orient='vertical', pos_left='right',
                                                    feature={'saveAsImage': {}}),
                    )
    return c, links

```

- (1) 使用 Graph 类创建一个图对象 c, 并设定初始参数 (如宽度和高度)。
- (2) add 方法添加了节点和边, 同时设定了排斥力、线样式、边标记、标签显示等参数。
- (3) 使用 set_global_opts 方法设置全局选项, 如图表标题和工具箱选项。
- (4) return c,links 返回生成的图对象和边的信息。

3、得到被审计对象的母子公司关系矩阵

```

def gettable(data1):
    # pagerank
    '''
    得到被审计对象的母子公司关系矩阵
    '''
    data = data1
    data = data[['母公司', '子公司']]
    data.reset_index(drop=True).rename(columns={'母公司': '输出', '子公司': '输入'}) # 重命名数据列标签

```

- (1) 输入参数是一个包含公司关系的数据框 data1。
- (2) 提取所需的两列 (母公司 和 子公司), 重命名为更具描述性的 ‘出’ 和 ‘输入’。

```

A = np.array(data)
users = np.unique(A.reshape([-1]))
id_map = {user: i for i, user in enumerate(users)}
id_A = np.vectorize(id_map.get)(A) # 将 str 映射为 int 值

```

- (1) 将数据框转换为 NumPy 数组 A。
- (2) 提取所有唯一的公司名称 users。
- (3) 创建一个映射，将公司名称转化为整数编号 id_map。
- (4) 使用 np.vectorize 将公司名称替换为对应的整数编号 id_A。

```
weights = [1] * len(A[:, -1])
numbers = len(set(A.reshape([-1])))
G = sparse.csr_matrix((weights, (id_A[:, 1], id_A[:, 0])),
shape=(numbers, numbers)) # 得到稀疏矩阵
```

- (1) 为每条边赋予权重 weights，这里设定为 1。
- (2) 计算节点总数 numbers。
- (3) 创建稀疏矩阵 G，表示公司之间的关系图。

```
# 计算 pagerank 值
pr = pagerank_power(G, p=0.85)
print('PageRank 计算完成, 用时: ', time.time() - start)
```

- (1) 使用 pagerank_power 函数计算每个节点的 PageRank 值 pr，参数 p=0.85 表示随机游走的概率为 85%。

```
# 影响力从大到小排序
dc = dict(zip(sorted(list(set(A.reshape([-1])))), pr))
list_dc1 = sorted(dc.items(), key=operator.itemgetter(1))
lst3 = [(u, v) for u, v in list_dc1]
pd1 = pd.DataFrame(lst3).sort_values(by=1,
ascending=False).rename(columns={0: '公司名称', 1: '重要性'})

return pd1
```

- (1) 构建一个稀疏矩阵：从输入的数据中提取出母公司和子公司的关系，然后将数据重命名为"输出"和"输入"列标签。将提取出的母子公司关系构建成一个稀疏矩阵。这里使用了 NumPy 和 SciPy 库来处理矩阵操作。

- (2) 计算 PageRank 值：利用 PageRank 算法计算母子公司的重要性。这里调用了一个名为 pagerank_power 的函数，它可能是自定义的 PageRank 算法的实现，但代码中并未提供该函数的定义，所以具体的实现细节需要查看该函数的源代码。

- (3) 结果排序：将计算得到的重要性按从大到小排序，并以 DataFrame 的形式返回结果。

4、寻找关联节点

主要目的是找到在网络图中边权重大于阈值的节点。

```
def find_associated_nodes(links, threshold=50):
    """
```

寻找关联节点，即图中边大于指定阈值的节点。

Parameters:

- **links:** 包含链接数据的列表。

- **threshold:** 边的阈值，大于该值的边被视为关联节点。

Returns:

- **associated_nodes:** 包含关联节点名称的列表。

```
"""
```

```
associated_nodes = []
```

```
for link in links:
```

```
    if link['value'] > threshold:
```

```
        associated_nodes.append(link['source'])
```

```
        associated_nodes.append(link['target'])
```

```
    # 去重
```

```
associated_nodes = list(set(associated_nodes))
```

```
return associated_nodes
```

定义了一个名为 `find_associated_nodes` 的函数，用于寻找关联节点，即图中边权值大于指定阈值的节点。

(1) 函数说明：该函数接受两个参数，`links` 是一个包含链接数据的列表，`threshold` 是边的阈值，大于该值的边被视为关联节点。

(2) 初始化关联节点列表：在函数开始时，定义了一个空列表 `associated_nodes`，用于存储找到的关联节点。

(3) 遍历链接数据：对输入的链接数据列表进行遍历，每次取出一个链接数据 `link`。

(4) 判断边权值：对于当前的链接数据 `link`，判断其边的权值是否大于指定的阈值 `threshold`。如果是，则将该边的源节点和目标节点添加到关联节点列表中。

(5) 去重：最后，对关联节点列表进行去重操作，确保每个关联节点只出现一次。

(6) 返回结果：返回包含关联节点名称的列表 `associated_nodes`。

```
# 读取数据
```

```
data1 = pd.read_excel('投资信息表.xlsx')
```

```
data2 = pd.read_excel('节点层级表.xlsx')
```

(1) `pd.read_excel` 用于读取 Excel 文件，并加载到 `data1` 和 `data2` 数据框中。

```
start = time.time() # 计算程序运行起始时间
```

```
c, links = networks(data1, data2) # 绘制中投标关系网络图
```

```
print('网络图生成完成，用时：', time.time() - start)
```

(1) `start = time.time()` 记录程序运行开始的时间。

(2) `networks(data1, data2)` 函数生成关系网络图，返回网络图对象 `c` 和链接信息 `links`。

(3) `print` 语句输出生成网络图所用的时间。

```
headers = ['公司名称', '重要性']
```

```
t4 = (
    Table()
        .add(headers, np.array(gettable(data1))).set_global_opts(
            title_opts=opts.ComponentTitleOpts(title='pagerank 信息')
        )) # pagerank 信息表
```

(1) headers 定义表格的列名。

(2) gettable(data1) 调用之前定义的 gettable 函数，计算 PageRank 值并返回一个包含公司名称和重要性的 DataFrame。

(3) 使用 pyecharts.components.Table 生成表格，并设置表格标题。

```
# 统计关联企业
# 调用 find_associated_nodes 函数获取关联节点
associated_nodes = find_associated_nodes(links, threshold=50)
print('关关节点: ', associated_nodes)
```

(1) find_associated_nodes(links, threshold=50) 函数根据边权重阈值（50）找到关联节点。

(2) print 语句输出关联节点的列表。

```
# 可视化
tab = Tab()
tab.add(c, '关系图')
tab.add(t4, 'pagerank 值')
tab.render('社交网络图.html')
```

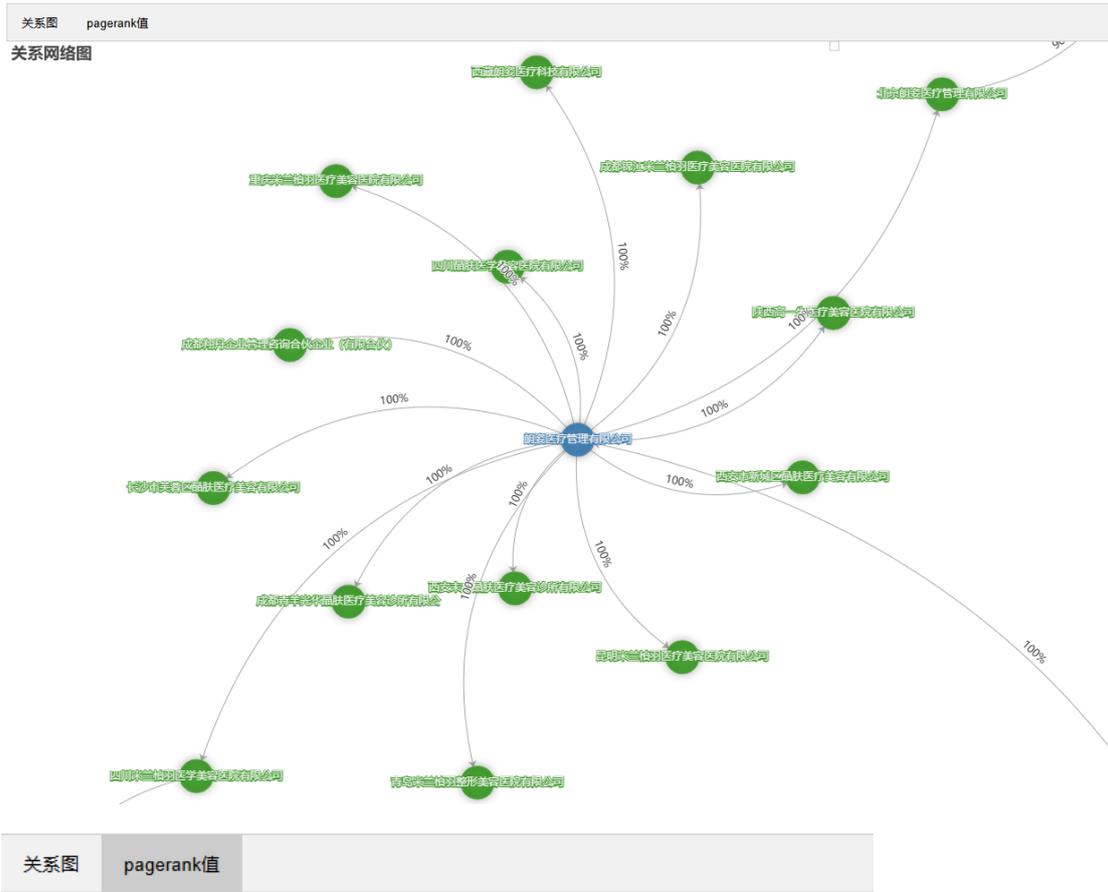
(1) Tab() 创建一个标签页对象。

(2) tab.add(c, '关系图') 将网络图 c 添加到标签页，并命名为“关系图”。

(3) tab.add(t4, 'pagerank 值') 将 PageRank 值表格 t4 添加到标签页，并命名为“pagerank 值”。

(4) tab.render('社交网络图.html') 将整个标签页渲染为 HTML 文件，命名为 社交网络图.html。

社交网络图：



pagerank信息

公司名称	重要性
朗姿股份有限公司	0.22430600348730861
朗姿医疗管理有限公司	0.1490452593377897
上海烜鼎资产管理有限公司-烜鼎长红七号私募证券投资基金	0.03720054678251824
香港中央结算有限公司	0.03720054678251824
申金花	0.03720054678251824
申东日	0.03720054678251824
朗姿股份有限公司-第三期员工持股计划	0.03720054678251824
北京合源融微股权投资中心 (有限合伙)	0.03720054678251824
上海烜鼎资产管理有限公司-烜鼎长红六号私募证券投资基金	0.03720054678251824

(三) 识别关联方

从多个渠道获取企业客户清单数据，并整理成表，然后结合社交网络图，从投资比例角度自动识别客户是否为关联企业，表格样式如下：

1	公司名称
2	朗姿股份有限公司
3	北京朗姿服饰有限公司
4	北京卓可服装有限公司
5	北京莱茵服装有限公司
6	朗姿（韩国）有限公司
7	成都朗姿商业发展有限公司
8	朗姿国际贸易有限公司
9	服装控股有限公司
10	西藏喇叭服饰有限公司
11	朗姿时尚（香港）有限公司
12	山南明懿时装有限公司
13	天津朗姿供应链管理有限公司
14	朗姿医疗管理有限公司
15	北京朗姿医疗管理有限公司
16	昆明韩辰医疗美容医院有限公司
17	成都高新柏羽美容服务有限公司
18	四川米兰柏羽医学美容医院有限公司
19	成都高新米兰柏羽医学美容医院有限公司
20	深圳米兰柏羽医疗美容门诊部
21	西安米兰柏羽医疗美容医院有限公司
22	四川晶肤医学美容医院有限公司
23	重庆晶肤医疗美容有限公司
24	长沙市芙蓉区晶肤医疗美容有限公司
25	西安晶肤医疗美容有限公司
26	成都青羊光华晶肤医疗美容诊所有限公司
27	成都高新锦城晶肤医疗美容诊所有限公司
28	西安市新城区晶肤医疗美容有限公司
29	成都武侯新南晶肤医疗美容诊所有限公司
30	成都成雅晶肤医疗美容门诊部有限公司
31	成都锦江星宏晶肤医疗美容诊所有限公司

输出结果：将包含公司名称和是否为关联方的结果保存到名为“识别结果.xlsx”的Excel文件中。

```
#第二部分：识别客户是否为关联方
```

```
# 读取数据表
```

```
data_table = pd.read_excel('企业客户清单.xlsx')
```

```
# 获取社交网络中的关联节点
```

```
associated_nodes = find_associated_nodes(links, threshold=50)
```

```
# 判断每个公司是否在社交网络中
```

```
data_table['是否为关联方'] = data_table['公司名称'].apply(lambda x: '是' if x in associated_nodes else '否')
```

```
# 输出结果
```

```
df = data_table[['公司名称', '是否为关联方']]
```

```
df.to_excel('识别结果.xlsx')
```

(1) 读取数据表：从名为“企业客户清单.xlsx”的 Excel 文件中读取数据，将数据存储在 data_table 变量中。

(2) 获取社交网络中的关联节点：调用了 find_associated_nodes 函数，获取了社交网络中的关联节点列表。

(3) 判断每个公司是否在社交网络中：对数据表中的每个公司名称应用了一个 lambda 函数，检查该公司是否在关联节点列表中。如果在列表中，则将对应的是否为关联方字段设置为“是”，否则设置为“否”。

这段代码的主要作用是将企业客户清单中的公司与社交网络中的关联节点进行比较，判断客户是否为关联方，并将结果保存到 Excel 文件中。

识别结果：

	A	B	C
		公司名称	是否为关联方
1			
2	0	朗姿股份有限公司	是
3	1	北京朗姿服饰有限公司	是
4	2	北京卓可服装有限公司	是
5	3	北京莱茵服装有限公司	否
6	4	朗姿(韩国)有限公司	否
7	5	成都朗姿商业发展有限公司	是
8	6	朗姿国际贸易有限公司	是
9	7	服装控股有限公司	否
10	8	西藏哗叽服饰有限公司	是
11	9	朗姿时尚(香港)有限公司	否
12	10	山南明懿时装有限公司	否
13	11	天津朗姿供应链管理有限公司	是
14	12	朗姿医疗管理有限公司	是
15	13	北京朗姿医疗管理有限公司	是
16	14	昆明韩辰医疗美容医院有限公司	否
17	15	成都高新柏羽美容服务有限公司	否
18	16	四川米兰柏羽医学美容医院有限公司	是
19	17	成都高新米兰柏羽医学美容医院有限公司	否
20	18	深圳米兰柏羽医疗美容门诊部	否
21	19	西安米兰柏羽医疗美容医院有限公司	否
22	20	四川晶肤医学美容医院有限公司	是
23	21	重庆晶肤医疗美容有限公司	否
24	22	长沙市芙蓉区晶肤医疗美容有限公司	是
25	23	西安晶肤医疗美容有限公司	否
26	24	成都青羊光华晶肤医疗美容诊所有限公司	否
27	25	成都高新锦城晶肤医疗美容诊所有限公司	否
28	26	西安市新城区晶肤医疗美容有限公司	是
29	27	成都武侯新南晶肤医疗美容诊所有限公司	否
30	28	成都成雅晶肤医疗美容门诊部有限公司	否
31	29	成都锦江星宏晶肤医疗美容诊所有限公司	否

项目十一：计算机视觉与应用

任务一 计算机视觉的概念与技术

一、实训要求

请在学习完知识准备下的内容后，完成下列各个任务。

任务分为两种模式：

①补全代码。请查看各任务下的任务描述，结合实际要求与所学内容填写参数，补全代码，需额外注意的内容已在提示中给出。

②自由编程。无案例背景限制，可自行补充并练习。

二、基础图像处理之读取图像

请在学习完知识准备下的内容后，完成下列图像读取任务。所需库是 CV2。

在程序编码补全中根据上述信息补全对应代码，其中数据以数据中心文件列表中的实际数据为准。

填写完毕后，可点击需求实现查看完整代码，确认无误后点击执行并显示结果，待代码运行完毕后即可查看结果。



图 11- 1

【参考答案】：

(1) imread

三、基础图像处理之图像旋转

请在学习完知识准备下的内容后，完成下列图像旋转任务。所需库是 CV2

①请补全代码。

②自由输入旋转角度，查看输出。

在程序编码补全中根据上述信息补全对应代码，其中数据以数据中心文件列表中的实际数据为准。

填写完毕后，可点击需求实现查看完整代码，确认无误后点击执行并显示结果，待代码运行完毕后即可查看结果。

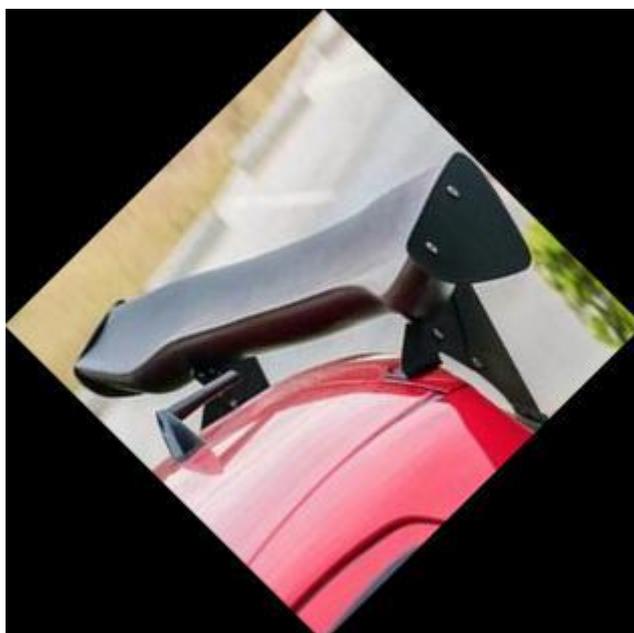


图 11- 2

【参考答案】：

(1) 45

四、基础图像处理之去噪处理

请在学习完知识准备下的内容后，完成下列图像去噪任务，所用知识点是中值滤波。所需库是 CV2。

①请补全代码。

②自由输入噪声强度，查看输出。

在程序编码补全中根据上述信息补全对应代码，其中数据以数据中心文件列表中的实际数据为准。

填写完毕后，可点击需求实现查看完整代码，确认无误后点击执行并显示结果，待代码运行完毕后即可查看结果。



图 11-3



图 11-4

图 11-3 为处理前的图片，图 11-4 为去噪处理后的图片。

【参考答案】：

(1) 2000

任务二 计算机视觉的基本任务

一、汽车零部件缺陷检测（目标检测）

（一）、任务说明

1. 任务背景

在工业生产过程中，由于现有技术、工作条件等因素的不足和局限性，极易影响制成品的质量。其中，表面缺陷是产品质量受到影响的最直观表现。因此，为了保证合格率和可靠的质量，必须进行产品表面缺陷检测。“缺陷”一般可以理解为与正常样品相比的缺失、缺陷或面积。表面缺陷检测是指检测样品表面的划痕、缺陷、异物遮挡、颜色污染、孔洞等缺陷，从而获得被测样品表面缺陷

的类别、轮廓、位置、大小等一系列相关信息。人工缺陷检测曾经是主流方法，但这种方法效率低下；检测结果容易受人为主观因素的影响，不能满足实时检测的要求。它已逐渐被其他方法所取代。而计算机视觉学科可以很好的满足该需求，由机器代替人工进行工业零部件的缺陷检测可以高效且高准确度地实现任务。

2. 任务要求

在生产线上，使用摄像头拍摄零件表面，检测汽车零部件（如车门）表面的缺陷（如划痕、凹痕、污渍），通过目标检测定位缺陷位置。

使用 DAGM 数据集（针对工业表面缺陷检测的小型数据集）作为案例，有 1046 张数据图片。使用 YOLOv5 模型实现对数据集图片内的缺陷部位进行检测并标出目标框，且对缺陷类别进行标注。数据集图片的样式如下所示，并且图中有一处缺陷。

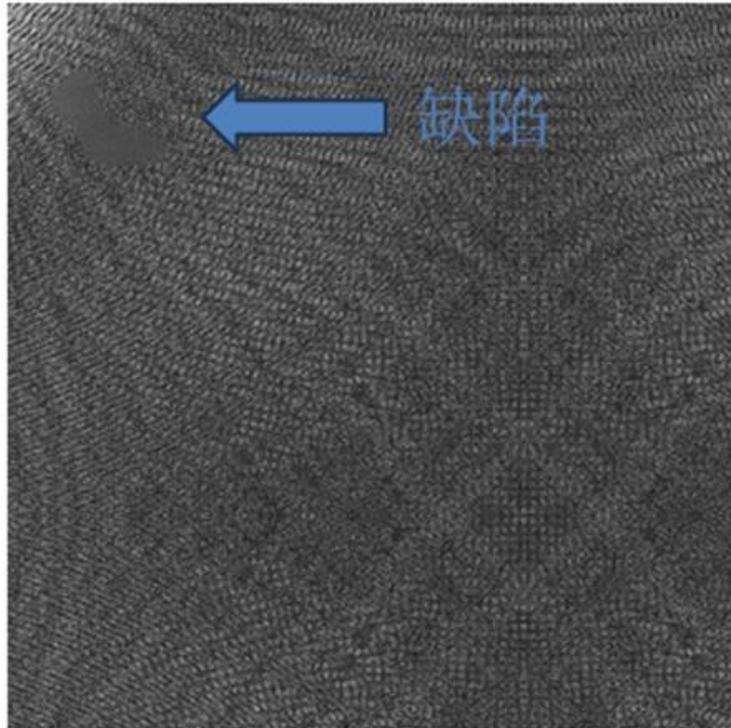


图 11-5

我们希望获得检测后的图片样式如下，图片上有目标框，包括缺陷的类别以及预测值：

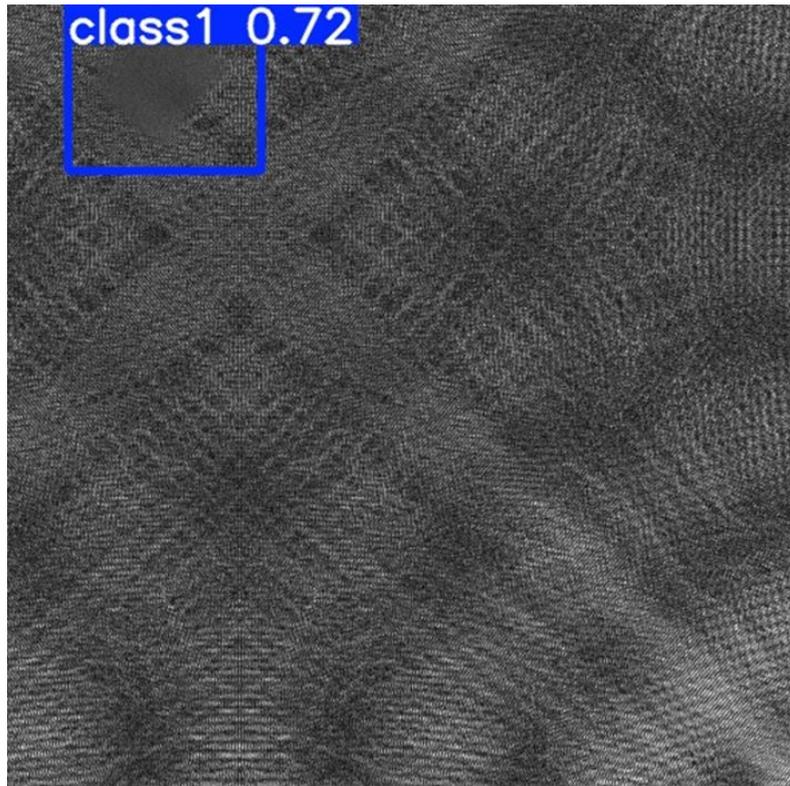


图 11-6

(二)、实现步骤

对汽车零部件进行目标检测。训练模型识别不同类别的缺陷（如划痕、污渍等），通过验证集评估模型性能（mAP、F1 分数等），并可视化检测结果以验证模型有效性。

1. 数据下载与转换

将数据集 DAGM 标注文件转换为 YOLO 训练需要的格式（VOC）。

在数据集的标注文件转换完成后，将数据集的标注文件(.xml)及原始图片分别放到目标项目 data 文件夹下的 Annotations 及 Images 中。

然后执行目标项目内的 makeTxt.py 和 voc_label.py 文件，目的是记录标注文件以及原始图片的图片名称，执行后会在 labels 文件夹内记录对应名称。

根据数据集编写.yaml 配置文件，train, val 分别对应上一步生成的 train.txt, val.txt 文件路径，nc 对应数据集分类总类目数量，names 数组存放每一类名称，在这个数据集中，每一类代表不同的缺陷类型，比如"scratch" 划痕、"stain"污渍、"dent" 凹痕、"foreign"异物等。

该文件存放在 data 文件夹内。

```

dagm.yaml x
1 # Dataset configuration for YOLOv5
2 train: data/train.txt
3
4 val: data/test.txt # Or 'test' if you're using a separate test set
5
6 nc: 10 # Number of classes
7
8 names: ['class1', 'class2', 'class3', 'class4', 'class5', 'class6', 'class7', 'class8', 'class9', 'class10']
9

```

图 11-7

2. 模型训练

在 yolo 的 github 官方项目内下载 yolov5 权重文件及配置文件（以 .pt 和 .yaml 结尾）并放入项目文件内，作为预训练权重。修改 train.py 内的 weights 为先前下载的权重文件（yolov5s.pt），修改 cfg 为先前下载的配置文件（yolov5s.yaml），修改 data 项为先前编写的数据集配置文件（dagm.yaml），执行该文件即可进行训练，训练过程的其余超参数同样在该文件内修改。

```

dagm.yaml x train.py x
545 def parse_opt(known=False):
546     - Tutorial: https://docs.ultralytics.com/yolov5/tutorials/train\_custom\_data
547     ***
548     parser = argparse.ArgumentParser()
549     parser.add_argument("name_or_flags": "--weights", type=str, default=ROOT / "yolov5s.pt", help="initial weights path")
550     parser.add_argument("name_or_flags": "--cfg", type=str, default="", help="yolov5s.yaml path")
551     parser.add_argument("name_or_flags": "--data", type=str, default=ROOT / "data/dagm.yaml", help="dataset.yaml path")
552     parser.add_argument("name_or_flags": "--hyp", type=str, default=ROOT / "data/hyps/hyp.scratch-low.yaml", help="hyperparameters path")
553     parser.add_argument("name_or_flags": "--epochs", type=int, default=100, help="total training epochs")
554     parser.add_argument("name_or_flags": "--batch-size", type=int, default=16, help="total batch size for all GPUs, -1 for autobatch")
555     parser.add_argument("name_or_flags": "--imgsz", "--img", "--img-size", type=int, default=640, help="train, val image size (pixels)")
556     parser.add_argument("name_or_flags": "--rect", action="store_true", help="rectangular training")
557     parser.add_argument("name_or_flags": "--resume", nargs="?", const=True, default=False, help="resume most recent training")
558     parser.add_argument("name_or_flags": "--nosave", action="store_true", help="only save final checkpoint")
559     parser.add_argument("name_or_flags": "--noval", action="store_true", help="only validate final epoch")
560     parser.add_argument("name_or_flags": "--noautoanchor", action="store_true", help="disable AutoAnchor")
561     parser.add_argument("name_or_flags": "--noplots", action="store_true", help="save no plot files")
562     parser.add_argument("name_or_flags": "--evolve", type=int, nargs="?", const=300, help="evolve hyperparameters for x generations")
563     parser.add_argument("name_or_flags": "--evolve_population", type=str, default=ROOT / "data/hyps", help="location for loading population")
564 )
565 parser.add_argument("name_or_flags": "--resume_evolve", type=str, default=None, help="resume evolve from last generation")
566 parser.add_argument("name_or_flags": "--bucket", type=str, default="", help="gsutil bucket")
567 parser.add_argument("name_or_flags": "--cache", type=str, nargs="?", const="ram", help="image --cache ram/disk")
568 parser.add_argument("name_or_flags": "--image_weights", action="store_true", help="use weighted image selection for training")
569 parser.add_argument("name_or_flags": "--device", default="", help="cuda device, i.e. 0 or 0,1,2,3 or cpu")
570 parser.add_argument("name_or_flags": "--multi-scale", action="store_true", help="vary img-size +/- 50%")
571 parser.add_argument("name_or_flags": "--single_cls", action="store_true", help="train multi-class data as single-class")
572 parser.add_argument("name_or_flags": "--optimizer", type=str, choices=["SGD", "Adam", "AdamW"], default="SGD", help="optimizer")
573 parser.add_argument("name_or_flags": "--sync-bn", action="store_true", help="use SyncBatchNorm, only available in DDP mode")
574 parser.add_argument("name_or_flags": "--workers", type=int, default=8, help="max dataloader workers (per RANK in DDP mode)")
575 parser.add_argument("name_or_flags": "--project", default=ROOT / "runs/train", help="save to project/name")
576 parser.add_argument("name_or_flags": "--name", default="exp", help="save to project/name")
577 parser.add_argument("name_or_flags": "--exist-ok", action="store_true", help="existing project/name ok, do not increment")
578 parser.add_argument("name_or_flags": "--quad", action="store_true", help="quad dataloader")
579 parser.add_argument("name_or_flags": "--cos-lr", action="store_true", help="cosine LR scheduler")
580 parser.add_argument("name_or_flags": "--label-smoothing", type=float, default=0.0, help="Label smoothing epsilon")
581 parser.add_argument("name_or_flags": "--patience", type=int, default=100, help="EarlyStopping patience (epochs without improvement)")
582 parser.add_argument("name_or_flags": "--freeze", nargs="+", type=int, default=[0], help="Freeze layers: backbone=10, first3=0 1 2")
583 parser.add_argument("name_or_flags": "--save-period", type=int, default=-1, help="Save checkpoint every x epochs (disabled if < 1)")
584 parser.add_argument("name_or_flags": "--seed", type=int, default=0, help="Global training seed")
585 parser.add_argument("name_or_flags": "--local_rank", type=int, default=-1, help="Automatic DDP Multi-GPU argument, do not modify")

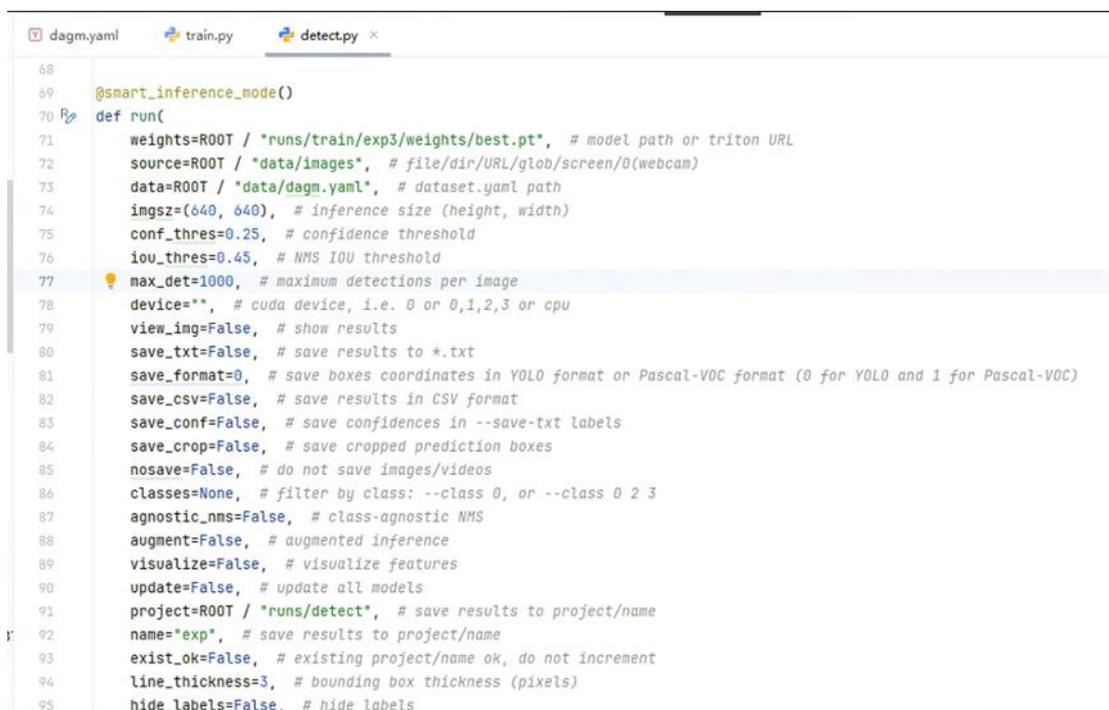
```

图 11-8

3. 模型预测

训练结束后，可以在 runs 文件夹内获得对应的训练结果以及训练过后的权

重，使用训练后的权重进行预测可以在此进行获取。修改 detect.py 文件内超参数，weights 修改为 runs 文件夹内对应权重文件路径，修改 data 配置文件为先前编写的数据集配置文件后执行 detect.py 即可对 images 文件夹内图片进行预测。将检测结果保存为 YOLO 格式。



```
68
69 @smart_inference_mode()
70 def run(
71     weights=ROOT / "runs/train/exp3/weights/best.pt", # model path or triton URL
72     source=ROOT / "data/images", # file/dir/URL/glob/screen/0(webcam)
73     data=ROOT / "data/dagm.yaml", # dataset.yaml path
74     imgsz=(640, 640), # inference size (height, width)
75     conf_thres=0.25, # confidence threshold
76     iou_thres=0.45, # NMS IOU threshold
77     max_det=1000, # maximum detections per image
78     device="", # cuda device, i.e. 0 or 0,1,2,3 or cpu
79     view_img=False, # show results
80     save_txt=False, # save results to *.txt
81     save_format=0, # save boxes coordinates in YOLO format or Pascal-VOC format (0 for YOLO and 1 for Pascal-VOC)
82     save_csv=False, # save results in CSV format
83     save_conf=False, # save confidences in --save-txt labels
84     save_crop=False, # save cropped prediction boxes
85     nosave=False, # do not save images/videos
86     classes=None, # filter by class: --class 0, or --class 0 2 3
87     agnostic_nms=False, # class-agnostic NMS
88     augment=False, # augmented inference
89     visualize=False, # visualize features
90     update=False, # update all models
91     project=ROOT / "runs/detect", # save results to project/name
92     name="exp", # save results to project/name
93     exist_ok=False, # existing project/name ok, do not increment
94     line_thickness=3, # bounding box thickness (pixels)
95     hide_labels=False, # hide labels
```

图 11-9

4. 检测结果可视化

将预测的部分结果进行可视化展示。输出图片 train_batch0.jpg, val_batch0.jpg，每个图片由 4*4 的数据图片组成，分别展示训练集和验证集的批量样本可视化，用于观察模型学习情况；输出图片 val_batch0_pred.jpg，展示验证集预测结果的可视化。

5. 模型评估

输出 results.png，确认模型在各类别上不同训练周期的表现，如边界框回归损失（预测框与真实框的差异）、目标置信度损失（判断是否有目标）、分类损失（类别预测的准确性）、精确率、召回率等。

输出 F1_curve.png，展示模型在不同置信度下的性能表现，帮助选择最优阈值。

输出 confusion_matrix.png，混淆矩阵，展示各类别的分类性能；对角线表

示正确预测的样本数,非对角线表示错误预测的样本数(如 class8 预测为背景)。在右侧加入色条使用颜色深浅表示比例(从 0.0 到 1.0),深蓝色代表高比例,浅蓝色代表低比例。

对比 val_batch0_labels.jpg (真实标签)和 val_batch0_pred.jpg (预测结果),分析误检/漏检原因。

(三)、目标检测实操

根据目标检测的实现步骤,完成代码的编写。

在技术需求转换中根据关键词填写对应参数,其中数据以数据中心下的实际数据为准。

填写完毕后,可点击需求实现查看完整代码,确认无误后点击执行并显示结果,待代码运行完毕后即可查看结果。如下所示:

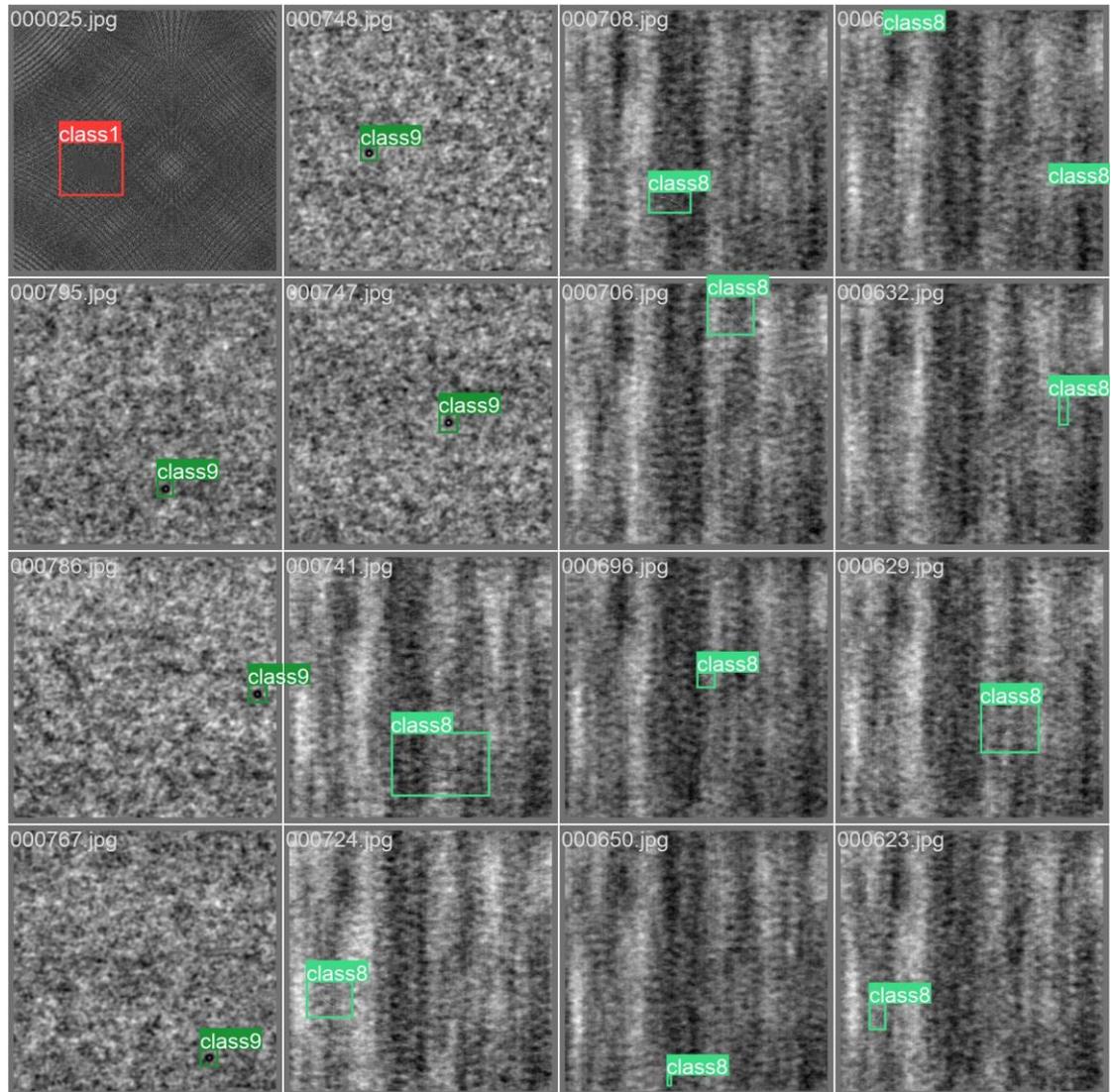


图 11- 10 数据检测-实际

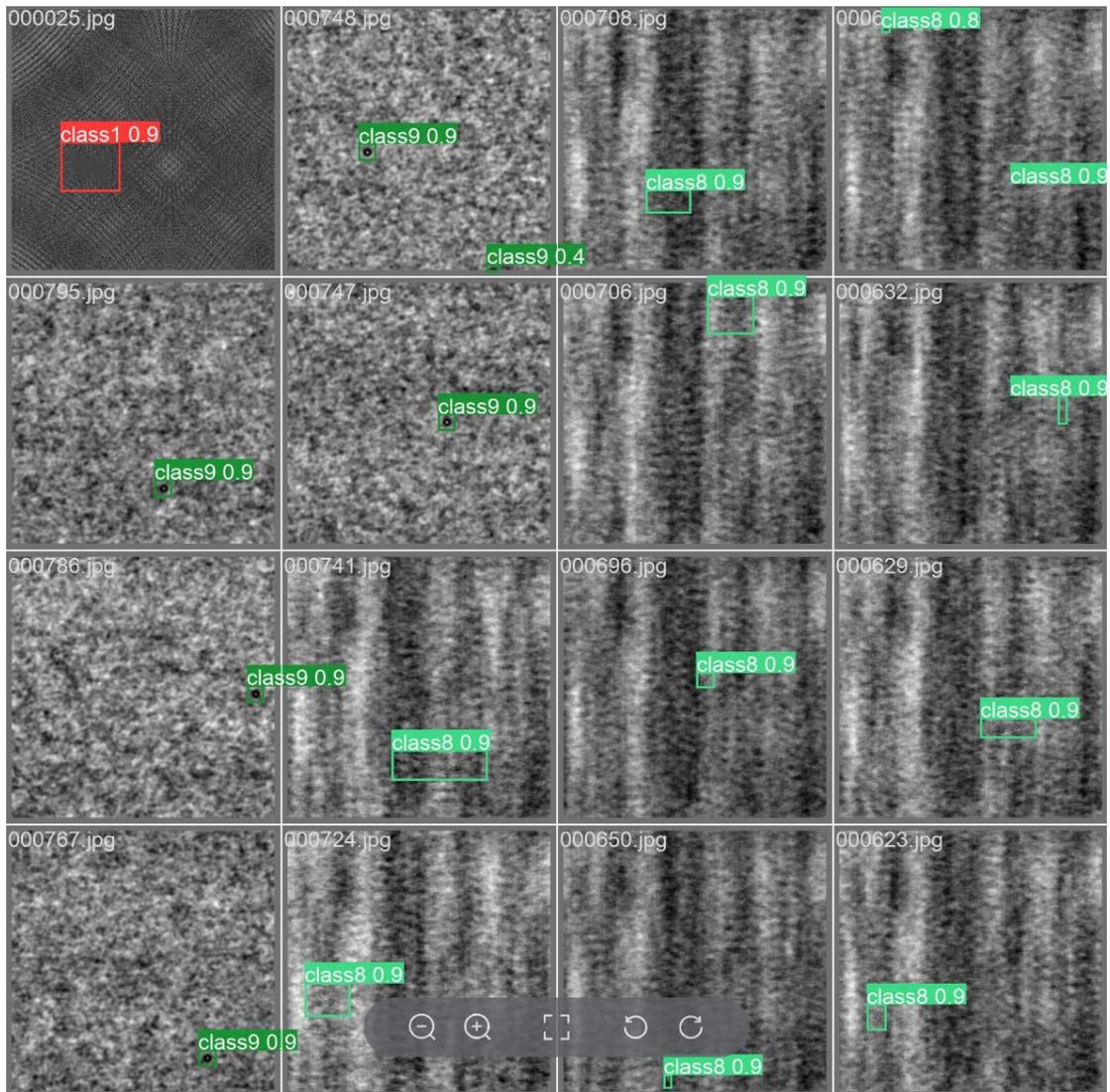


图 11- 11 数据检测-预测

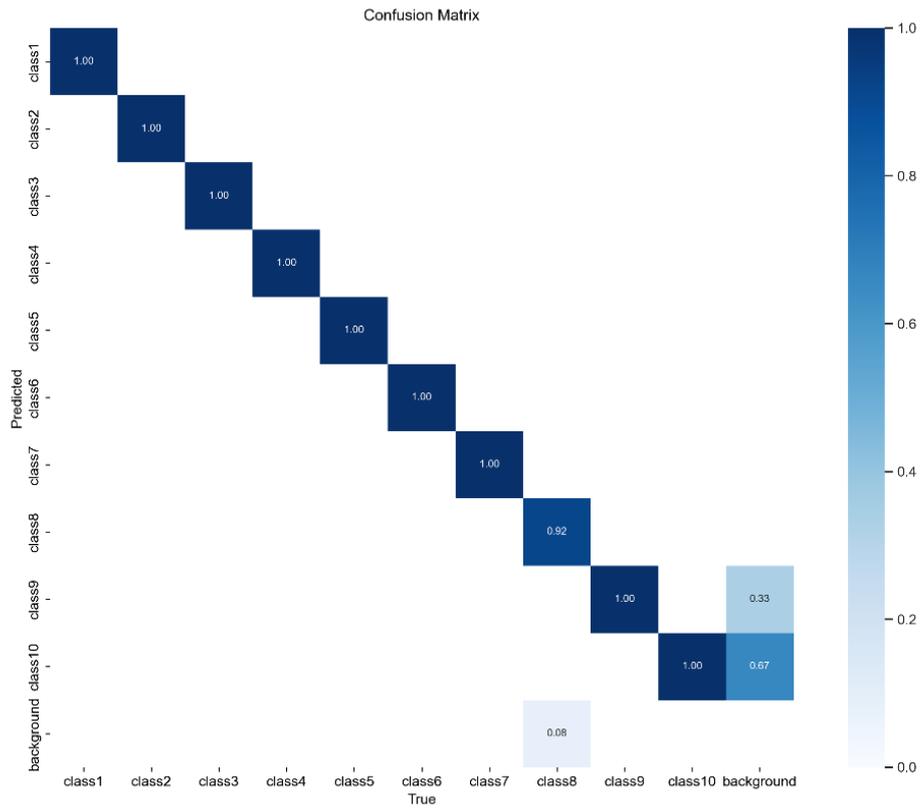


图 11- 12 混淆矩阵

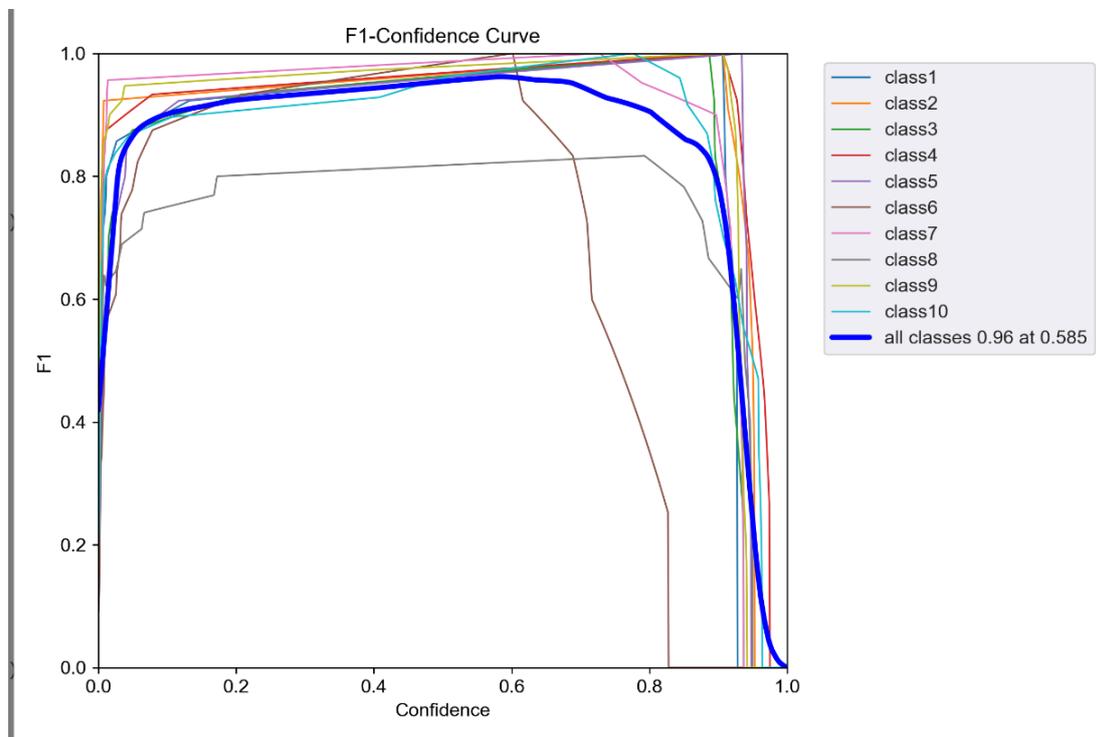


图 11- 13 F1 置信度曲线

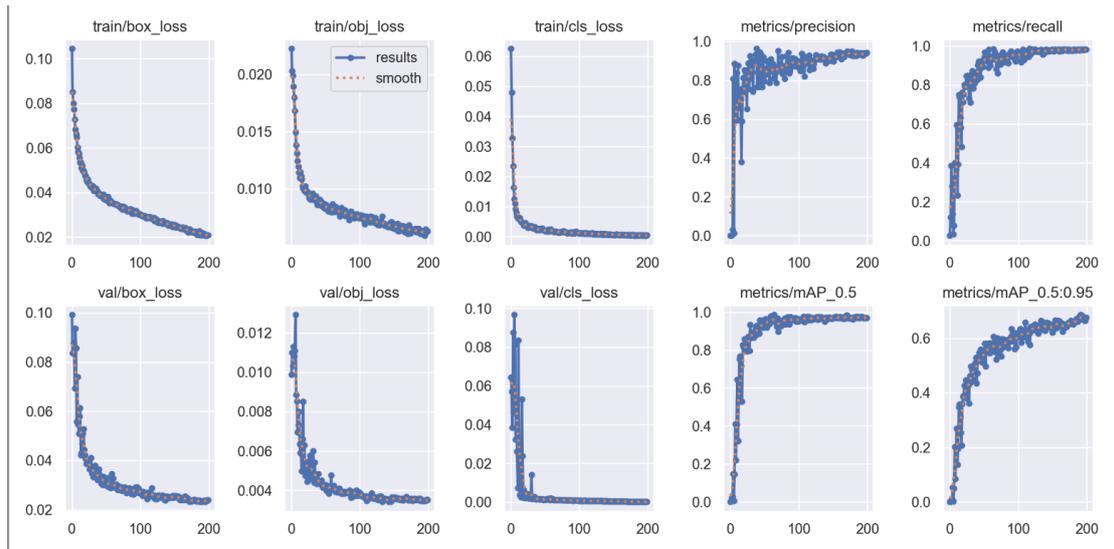


图 11-14 模型损失函数和评估指标变化

参考答案：

序号	关键词	参数
1	数据集分类数量	10
2	weights 权重文件	yolov5s.pt
3	cfg 配置文件	yolov5s.yaml

二、汽车零件瑕疵区域分割（图像分割）

（一）、任务说明

1. 任务背景

图像分割是指将数字图像划分为若干个不同区域的过程，使得每个区域内的像素具有相似的特征，是图像处理到图像分析的关键步骤。在工业领域，图像分割的意义重大，它可以帮助工程师们精确提取目标物体的特征信息，实现对产品外观、尺寸、表面缺陷等方面的检测，从而提高产品的质量和可靠性，减少人力投入，提升生产效率。

2. 任务要求

通过语义分割技术标注车门表面不同瑕疵类型（划痕、锈蚀等），精确分割汽车零件上的瑕疵区域，量化缺陷面积。

使用 U-Net 模型对数据集 NEU-CLS 进行像素级分割。对比分割面积与实际尺寸，评估模型精度。

数据集：NEU Surface Defect Database (NEU-CLS)，包含工业零件表面缺陷分割标注。NEU-CLS 包含 6 种常见的金属表面缺陷，每种缺陷对应一个类别：

Crazing（龟裂）：表面细小裂纹。

Inclusion（夹杂）：金属内部非金属杂质。

Patches（斑块）：局部颜色或纹理异常。

Pitted_surface（麻面）：表面密集小凹坑。

Rolled-in_scale（轧制氧化皮）：轧制过程中残留的氧化层。

Scratches（划痕）：线性表面损伤。

在本次案例中，选取 Inclusion（夹杂）、Patches（斑块）、Scratches（划痕）三种类型的样本，分别为 In_1~In50、Pa_1~Pa50、Sc_1~Sc50。图片样式如下所示：

类别 Inclusion (In_3)：

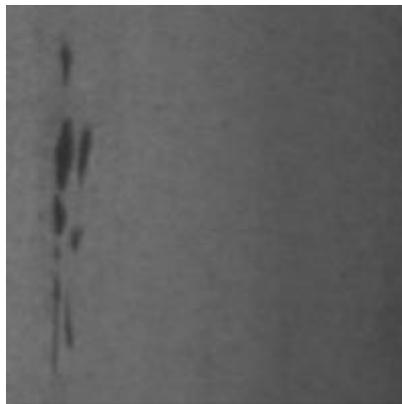


图 11-15

类别 Patches (Pa_10)：

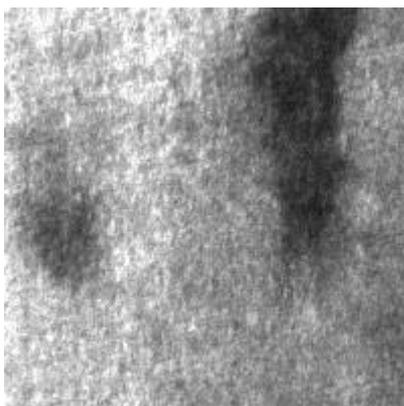


图 11- 16

类别 Scratches (Sc_11) :

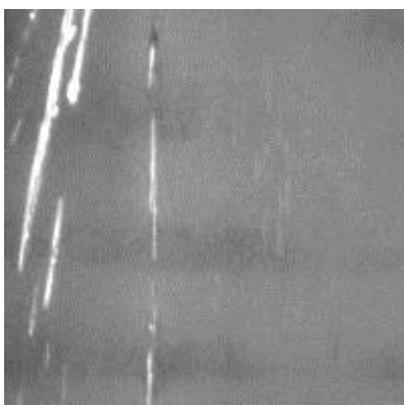


图 11- 17

(二)、实现步骤

1. 数据标注

nue-cls 数据集语义标注并没有多分类，本次任务的实现重新对数据集进行了标注，并且标注需要注意质量以及数量。

标注方法：

终端安装 labelme 标注工具（注意 python 版本，版本过高会导致闪退，推荐 python3.8）：`pip install labelme==3.16.2`

在 anaconda 中运行安装 labelme 的环境，并执行：`labelme` 进行调用。

选择对应文件夹后选择 `create polygons` 进行选点标注，标注完成后标记对应类别自动生成标注文件。

在 View 中勾选 `Auto Save mode`。

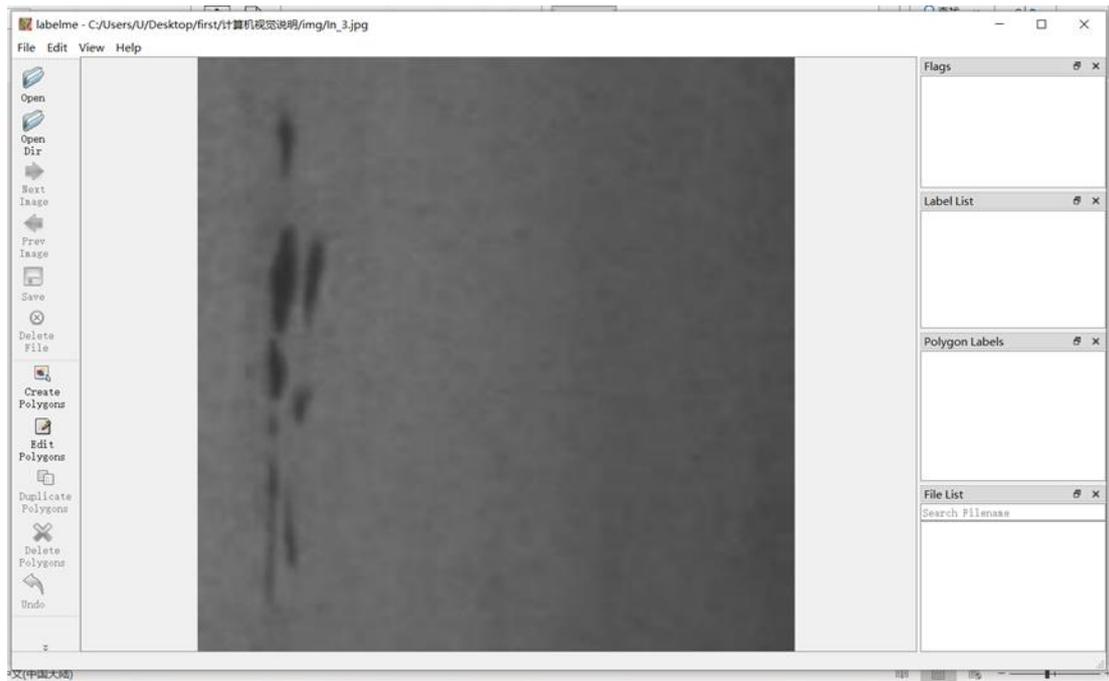


图 11-18

2. 模型训练与预测

训练前需要检查数据的格式是否满足要求,该库要求数据集格式为 VOC 格式,需要准备好的内容有输入图片和标签;输入图片为 .jpg 图片,无需固定大小,传入训练前会自动进行 resize。

将标签文件放在 VOCdevkit 文件夹下的 VOC2007 文件夹下的 SegmentationClass 中,将图片文件放在 VOCdevkit 文件夹下的 VOC2007 文件夹下的 JPEGImages 中,利用 voc_annotation.py 文件生成对应的 txt。注意修改 train.py 的 num_classes 为分类个数+1,比如本次数据集的图片类别有三类,则 num_classes=4。

接下来配置这个基于 U-Net 的语义分割模型,主要步骤:配置参数 → 初始化模型 → 加载数据 → 设置训练策略 → 启动训练循环

配置参数:设置训练超参数(如学习率、批次大小、优化器等);初始学习率为 $Init_lr = 1e-4$,最小学习率为 $Min_lr = Init_lr * 0.01$,优化器类型 $optimizer_type = "adam"$,学习率衰减方式为余弦衰减 $lr_decay_type = 'cos'$,总训练轮次为 $UnFreeze_Epoch = 200$ 。

分布式训练初始化:设置 $distributed = False$,不启用分布式训练,使用单 GPU 或 CPU;

模型初始化：设置 `pretrained=False`，随机初始化权重；`model_path` 加载预训练权重（部分匹配的权重）；

数据加载：读取训练集和验证集的文件列表（VOC 格式数据集）；

训练准备：冻结/解冻模型主干（`freeze_backbone`）以加速初始训练；自适应学习率计算，根据 `batch_size` 自适应调整学习率；优化器和学习率调度器配置；

训练循环：调用 `fit_one_epoch` 进行逐轮训练。

运行 `train.py` 即可开始训练。

在 `predict.py` 指定预测模式，预测图片路径，分类，权重，执行 `predict.py` 即可进行预测；输出 In_3、Pa_10、Sc_11 的图像分割预测结果。

（三）、图像分割实操

根据图像分割的实现步骤，完成代码的编写。

在技术需求转换中根据关键词填写对应参数，其中数据以数据中心下的实际数据为准。

填写完毕后，可点击需求实现查看完整代码，确认无误后点击执行并显示结果，待代码运行完毕后即可查看结果。如下所示：

输出：

In_3、Pa_10、Sc_11的图像分割预测结果如下：

图 11-19

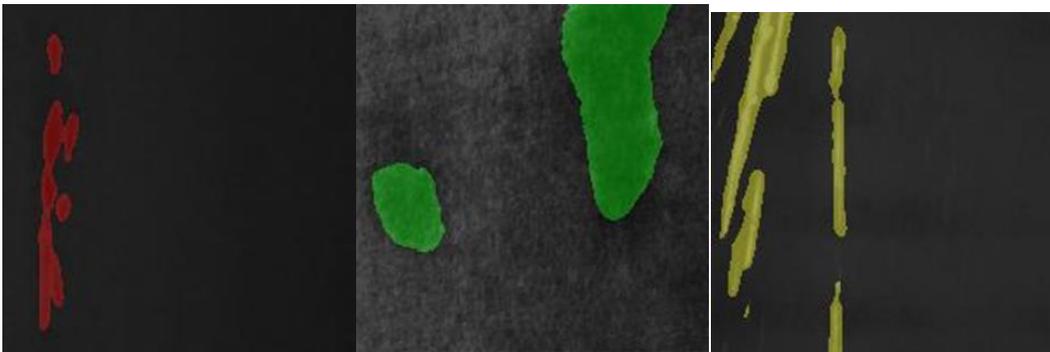


图 11-20

参考答案：

序号	关键词	参数
----	-----	----

1	分类个数(num_classes)	4
2	优化器类型	adam
3	学习率下降的方式	cos

三、汽车零部件缺陷样本生成（图像生成）

（一）、任务说明

1. 任务背景

在机器学习，尤其是深度学习领域，数据集的规模和质量对模型的训练效果具有至关重要的影响。当数据集不足时，通过图像生成技术，可以生成具有不同特征、不同场景和不同光照条件下的图像样本，从而增加数据集的多样性，有助于模型学习到更丰富的特征表示，多样化的图像样本有助于模型学习到更加鲁棒的特征表示，使其在面对不同场景和条件时都能保持稳定的性能。

2. 任务目标

在数据不足情况下，利用 GAN 合成缺陷样本提升检测模型性能；生成车门划痕或锈迹图像的缺陷样本，用于数据扩充，丰富训练数据。

使用 CycleGAN 关于数据集 SDD 生成不同类型的表面缺陷图像。将生成图像加入训练集，分析其对检测精度的提升。

数据集原有的缺陷图像如下所示：

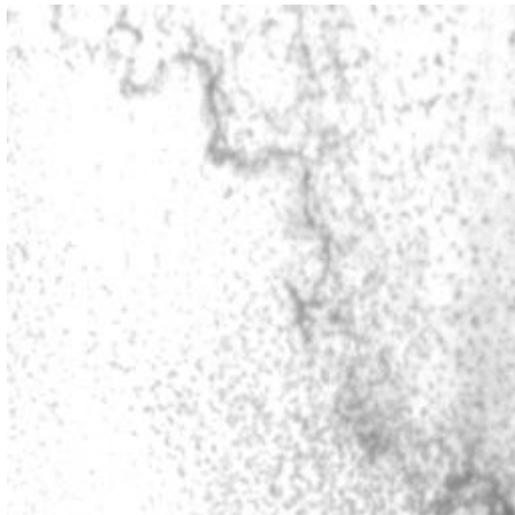


图 11-21

现在想生成一个缺陷样本用于数据扩充：

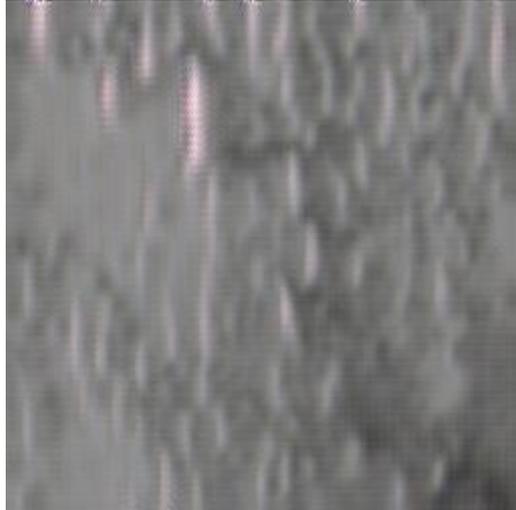


图 11-22

(二)、实现步骤

根据数据集的划分，修改 train.py 和 test.py 的配置文件：

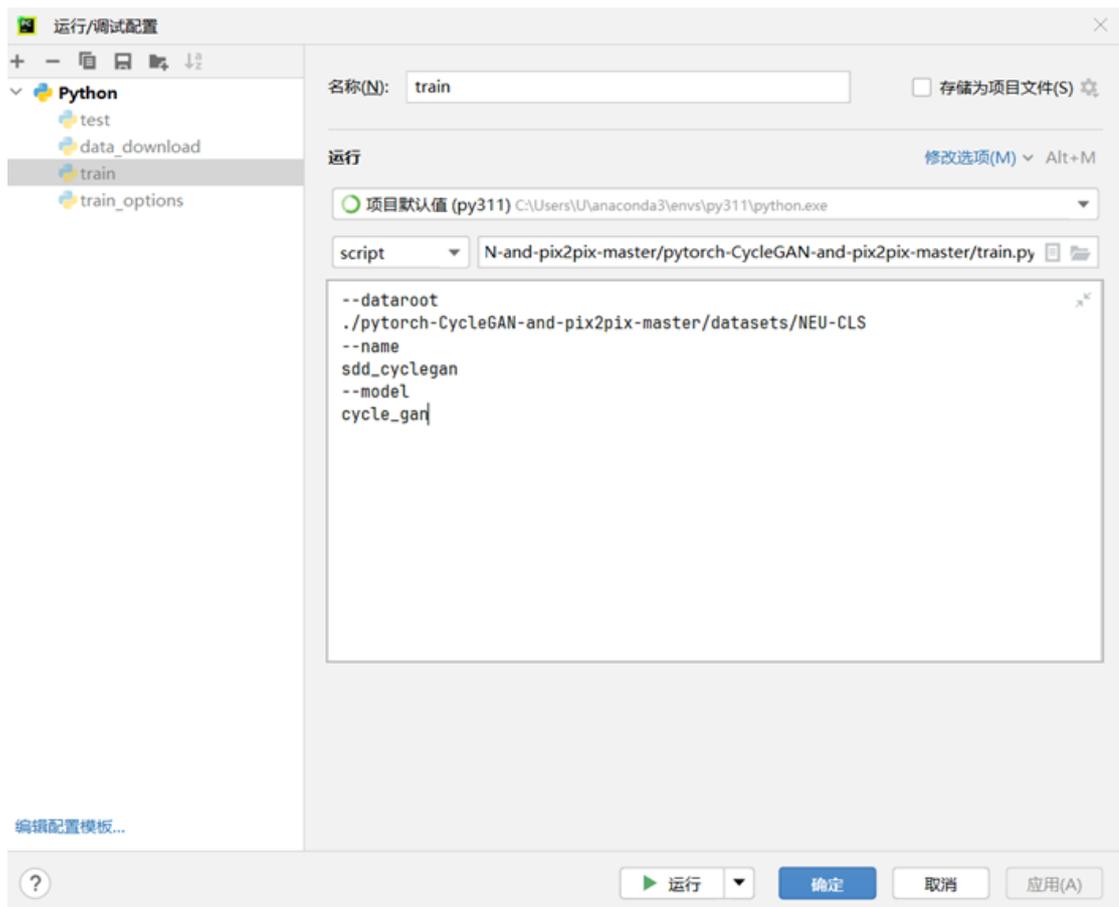


图 11-23

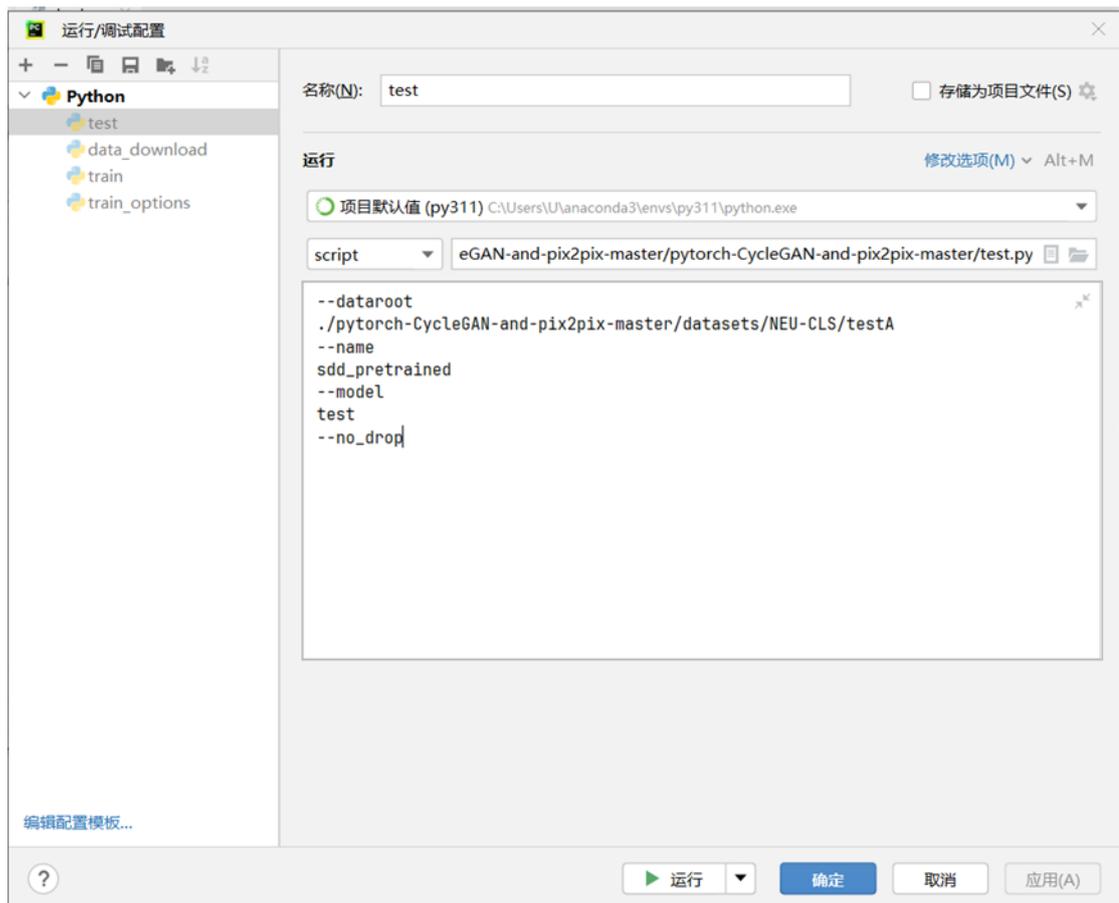


图 11-24

Dataroot：数据集图片根目录，在此目录下将数据集划分为 trainA, trainB, testA, testB

Name：结果保存文件夹名称

Model：指定了训练脚本 train.py 要使用的模型类型

训练使用终端命令：

```
pythonpytorch-CycleGAN-and-pix2pix-master/train.py
--dataroot ./pytorch-CycleGAN-and-pix2pix-master/datasets/NEU-DET
--name sdd_cyclegan --model cycle_gan
```

预测终端命令：

```
python ./pytorch-CycleGAN-and-pix2pix-master/test.py
--dataroot ./pytorch-CycleGAN-and-pix2pix-master/datasets/NEU-DET/testA
--name ./pytorch-CycleGAN-and-pix2pix-master/sdd_pretrained --model
test --no_dropout
```

在 visdom 动态查看训练状态：python -m visdom.server

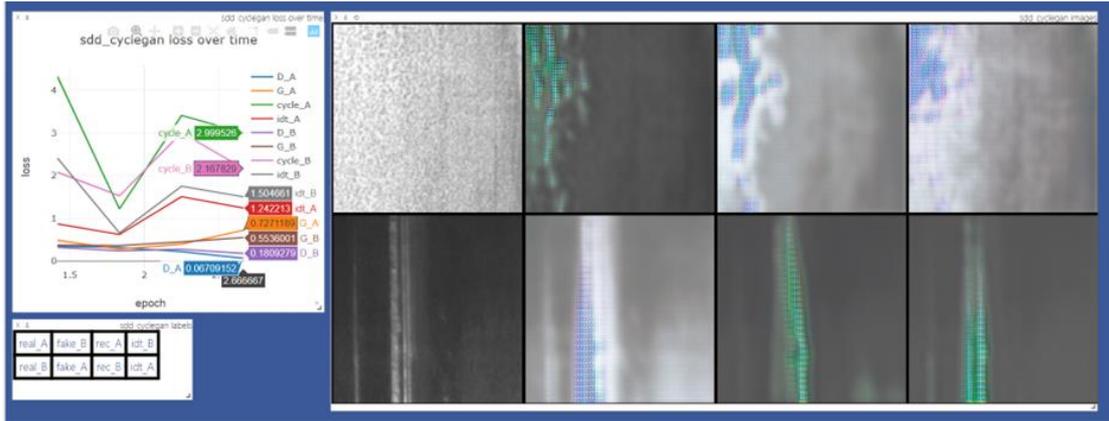


图 11- 25

其余超参数修改在 options 文件夹的 base_options 文件内。

模型训练：

导入 Visualizer 类，用于可视化训练过程中的图像和结果。

通过 TrainOptions 类解析命令行参数或配置文件，获取训练选项（如学习率、批次大小、数据集路径等）；

使用 create_dataset 根据选项创建数据集加载器，使用 create_model 根据选项创建模型 (pix2pix)；create_dataset 和 create_model 是工厂函数，根据 opt 中的参数动态选择数据集/模型类；

进行训练循环：`for epoch in range(opt.epoch_count, opt.n_epochs + opt.n_epochs_decay + 1)`

定期将训练结果（如生成的图像）显示到 Visdom 或保存到 HTML 文件；随机输出一张生成的图像。

（三）、图像生成实操

根据图像生成的实现步骤，完成代码的编写。

在技术需求转换中根据关键词填写对应参数，其中数据以数据中心下的实际数据为准。

填写完毕后，可点击需求实现查看完整代码，确认无误后点击执行并显示结果，待代码运行完毕后即可查看结果。如下所示：

输出：
原图片及生成的图片如下所示：

图 11- 26

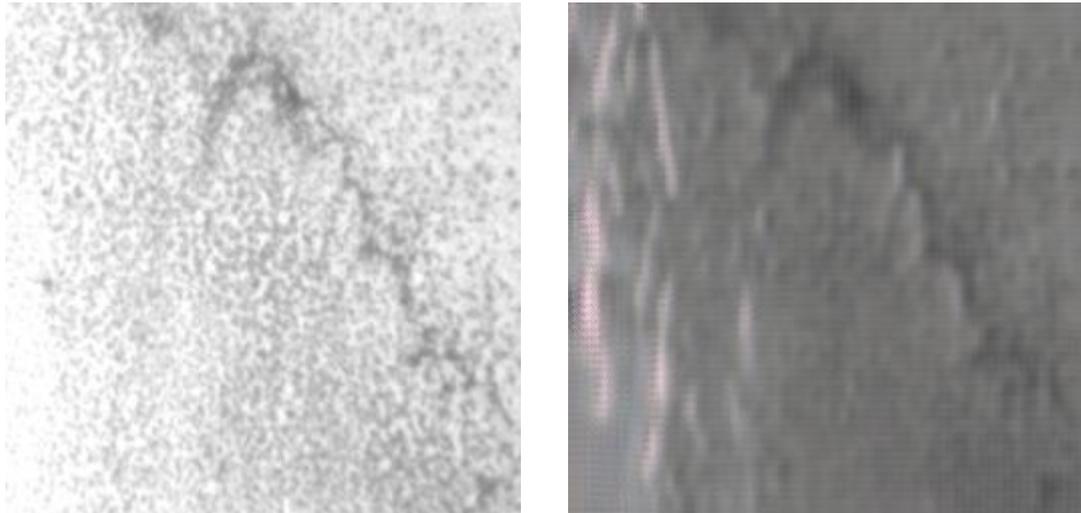


图 11-27

参考答案：

序号	关键词	参数
1	解析训练配置参数的类	TrainOptions()
2	创建数据集的函数	create_dataset

任务三 计算机视觉的常见任务

一、案例背景

在现代工业制造体系中，焊接工艺是连接各类金属构件的核心环节，广泛应用于汽车制造、航空航天、电子设备组装等关键领域。焊点质量直接决定产品结构强度与安全性，如汽车车身焊接若存在缺陷，可能在行驶中引发部件脱落；航空发动机焊接瑕疵会威胁飞行安全。但传统人工质检方式依赖经验，效率低、漏检误判率高，难以适配大规模、高精度生产需求，因此构建智能焊接质量检测系统迫在眉睫。

本项目围绕焊接质量全流程检测展开，从“焊点合格性判定 - 缺陷定位与分类 - 裂缝精细分割”三个递进任务，打造覆盖“宏观筛选 - 缺陷识别 - 精准量化”的完整质检方案。

二、划分焊点图像

(一)、任务说明

在工业焊接生产里，人工判定焊点质量耗时费力还易受主观影响。本任务借助 GDxray 数据集，将焊点图像划分成合格与缺陷两类，经随机旋转、缩放、裁剪增强数据后，加载预训练 ResNet50 模型并修改最后一层构建二分类模型，以二元交叉熵为损失函数训练。过程中观察训练曲线（loss 和 accuracy），训练后在测试集预测并输出混淆矩阵与准确率，实现自动判别焊点是否合格，助力企业提升生产效率、节约成本，推动工业生产智能化质检升级。

(二)、实现步骤

第一次运行会自动下载预训练权重。

训练时可以注释掉预测部分的代码，这样可以将训练和预测分开执行，预测时同样可以注释掉执行训练部分的语句。

数据集增强部分实现：

```
# 数据集增强
image_transforms = {
    'train': transforms.Compose([
        transforms.RandomResizedCrop(size=256, scale=(0.8, 1.0)),
        transforms.RandomRotation(degrees=15),
        transforms.RandomHorizontalFlip(),
        transforms.CenterCrop(size=224),
        transforms.ToTensor(),
        transforms.Normalize(mean=[0.485, 0.456, 0.406],
                             std=[0.229, 0.224, 0.225])
    ]),
    'valid': transforms.Compose([
        transforms.Resize(size=256),
        transforms.CenterCrop(size=224),
        transforms.ToTensor(),
        transforms.Normalize(mean=[0.485, 0.456, 0.406],
                             std=[0.229, 0.224, 0.225])
    ])
}
```

图 11- 28

修改网络结构符合自己的数据集，这里是 6 分类：

```

# 适应自己的数据集
fc_inputs = resnet50.fc.in_features
resnet50.fc = nn.Sequential(
    nn.Linear(fc_inputs, out_features=256),
    nn.ReLU(),
    nn.Dropout(0.4),
    nn.Linear(in_features=256, out_features=6),
    nn.LogSoftmax(dim=1)
)

```

图 11-29

指定预测时使用的权重地址：

```

trained_model = torch.load('models/'+data_model_100.pt')

```

图 11-30

1. `transforms.RandomResizedCrop(size=256, scale=(0.8, 1.0))`：

作用：随机裁剪并缩放图像。

详细说明：首先从图像中随机选取一个区域，并将该区域缩放至 256x256 像素（`size=256`）。`scale=(0.8, 1.0)` 表示裁剪区域的比例可以是原图大小的 80% 到 100% 之间的任意大小。这是数据增强的一种方式，用来增加训练的多样性，提高模型的泛化能力。

2. `transforms.RandomRotation(degrees=15)`：

作用：随机旋转图像。

详细说明：这会将图像随机旋转一个角度，旋转角度范围为 -15 到 +15 度。旋转图像是数据增强的一种方式，有助于提高模型的鲁棒性，使其在面对不同角度的输入时仍能做出准确预测。

3. `transforms.RandomHorizontalFlip()`：

作用：随机水平翻转图像。

详细说明：每次训练时，有 50% 的概率会将图像水平翻转。水平翻转是常见的数据增强方式，特别适用于图像中的对象左右对称的情况。

4. `transforms.CenterCrop(size=224)`：

作用：裁剪图像的中央部分。

详细说明：将图像的中央部分裁剪为 224x224 像素大小。裁剪操作通常用于标准化图像大小，确保输入到网络的图像具有相同的尺寸。

（三）、具体操作

使用预训练的 ResNet50 模型对多分类模型进行训练，对测试集进行预测，输出混淆矩阵和准确率。

在技术需求转换中根据关键词填写对应参数，其中数据以数据中心下的实际数据为准。

填写完毕后，可点击需求实现查看完整代码，确认无误后点击执行并显示结果，待代码运行完毕后即可查看结果。如下所示：

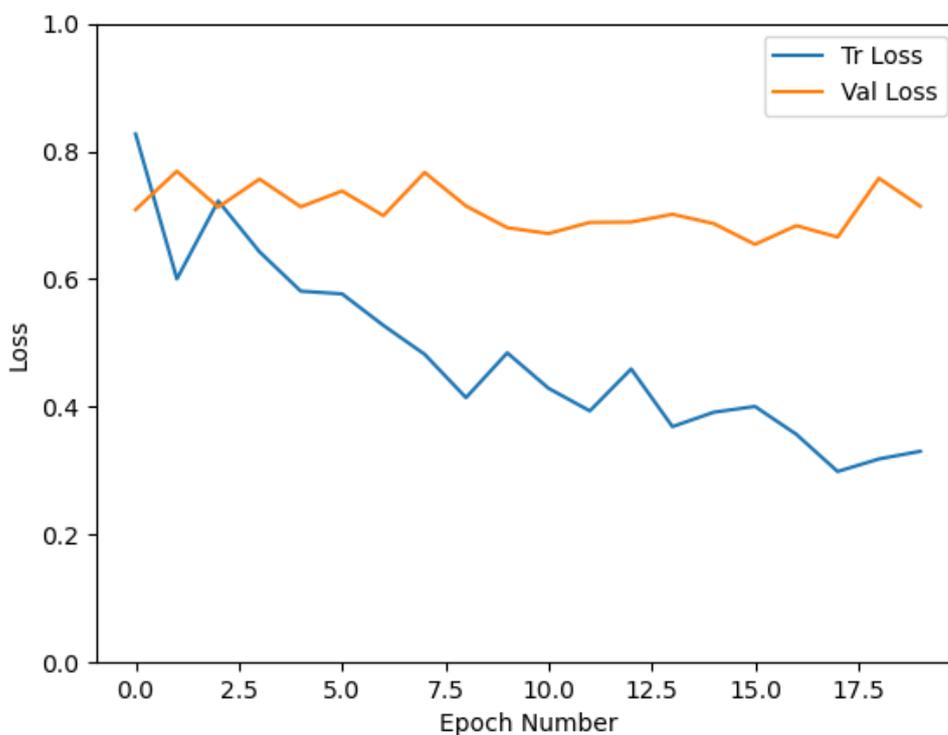


图 11- 31

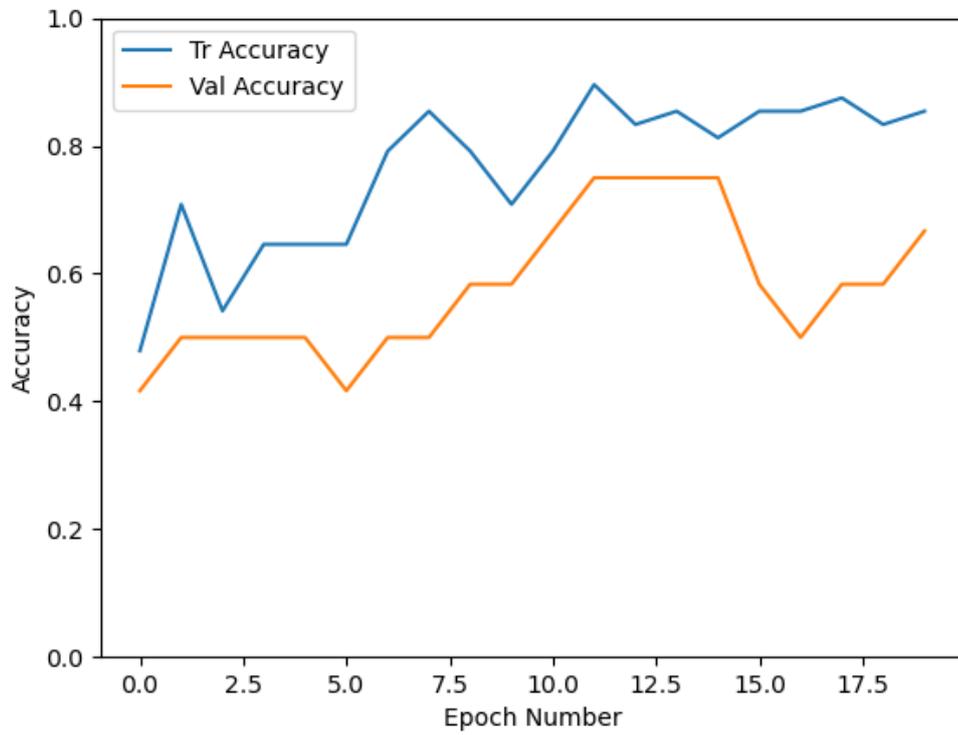


图 11- 32

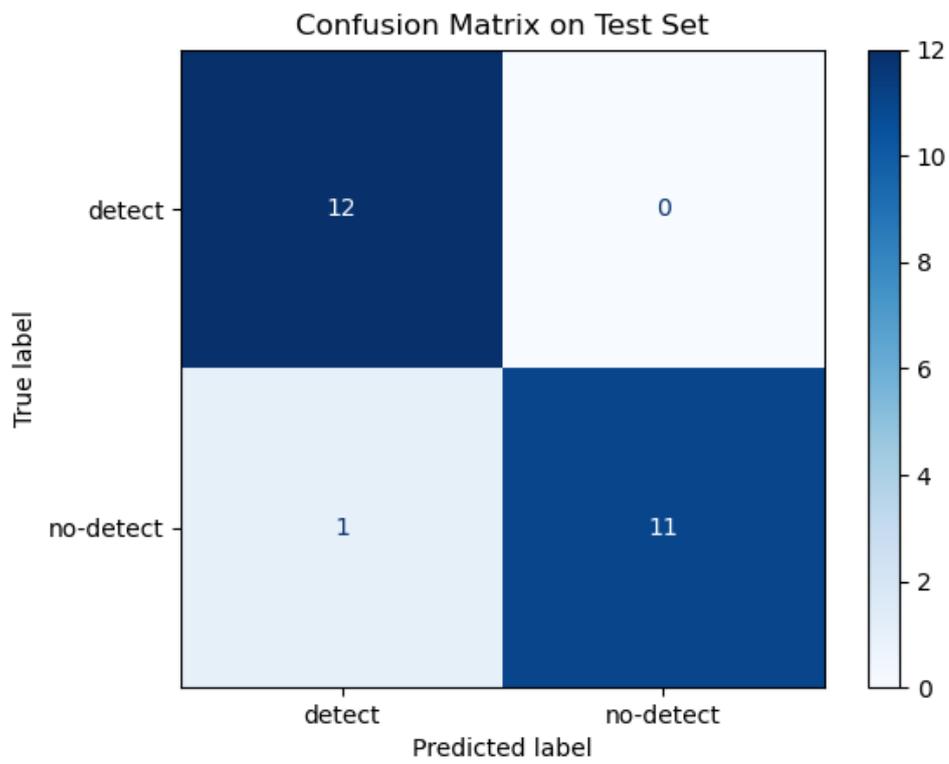


图 11- 33

参考答案:

序号	关键词	参数
1	批量大小	32
2	类别数量	6
3	训练轮数	12

三、定位焊点缺陷的具体不位置与类别

(一)、任务说明

在工业焊接质量把控环节，为精准定位焊点缺陷的具体位置与类别，开展本目标检测任务。选用 Kaggle 的 “Steel Defects” 数据集（因无专用焊点缺陷数据集，以此钢铁表面缺陷数据集作为替代入门级数据），先进行数据准备，用 LabelImg 工具标注缺陷位置，生成 YOLO 格式标注文件，明确缺陷类别与坐标；接着利用 YOLOv5 脚本，加载标注数据训练模型，过程中调整学习率、置信度阈值等超参数；完成训练后，加载权重对测试图像推理，输出并展示缺陷位置和类别，通过分析结果，助力工业场景中高效识别焊点缺陷，为焊接质量评估与生产优化提供技术支撑，推动焊接质检向更智能、精准方向发展。

(二)、实现步骤

本任务修改超参数同样在 train.py 和 detect.py 内进行，训练前请准备好预训练权重文件以及配置文件，预测前请准备好测试图片以及权重文件。

配置文件示例:

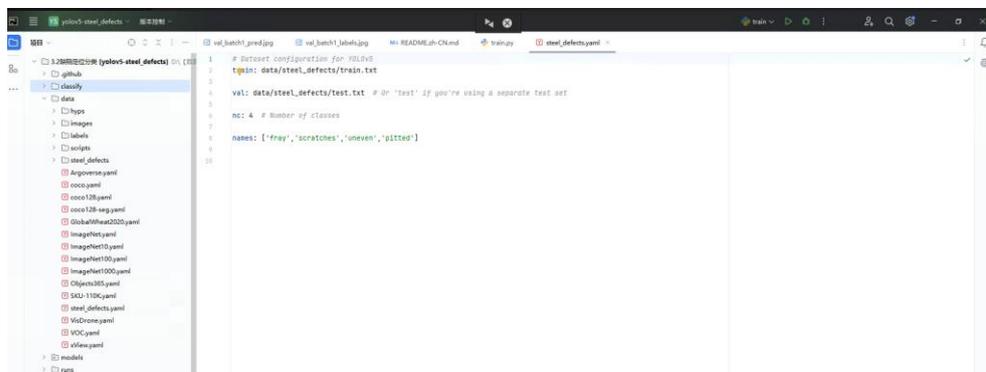


图 11- 34

参考 LabelImg（目标检测标注工具）的安装与使用教程-CSDN 博客

Pycharm 终端: `pip install labeling`

Anaconda 运行对应环境后执行: `labelimg`



图 11-35

左侧配置中选择 yolo 的标注格式, 对缺陷标注, 画框, 标记目标的 label, 然后切换到下一张继续标注。

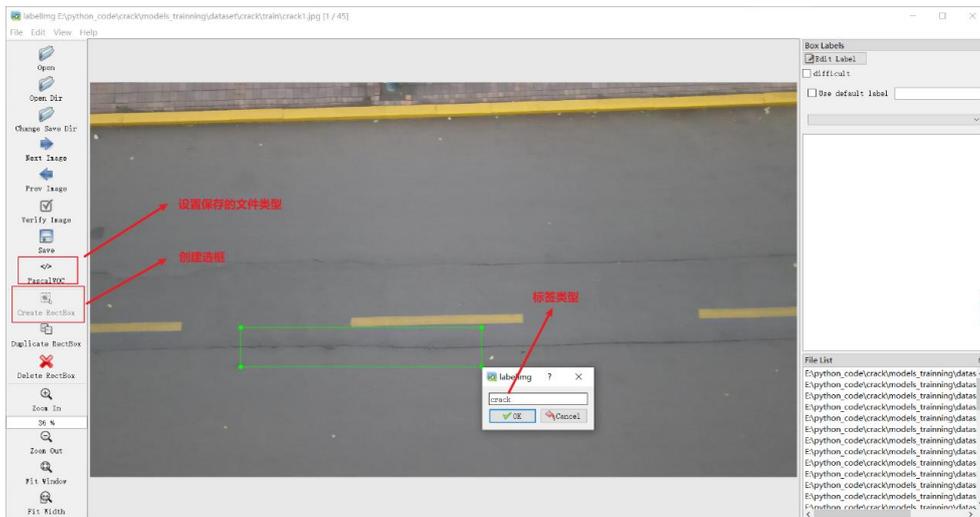


图 11-36

标注文件格式如下:

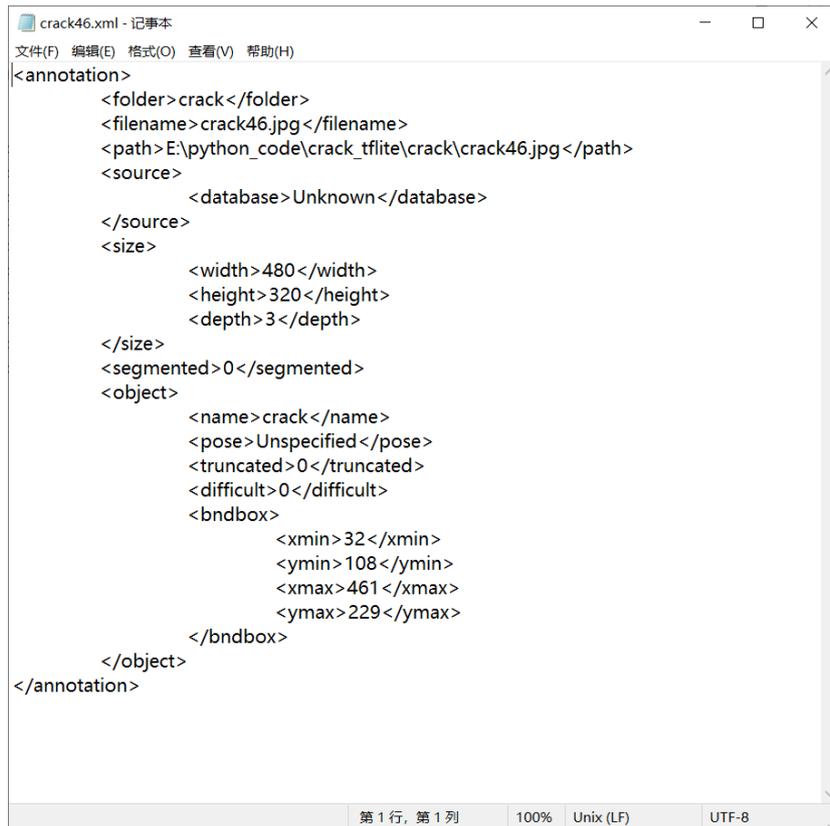


图 11- 37

在标注完成后将图片和标注文件对应放入项目内 data 文件夹中，执行目标项目内的 makeTxt.py 和 voc_label.py 文件（参考：Yolov5 训练自己的数据集（VOC）_如何用 voc 数据集训练 yolov5-CSDN 博客），目的是记录标注文件以及原始图片的图片名称，执行后会在 labels 文件夹内记录对应名称。

在配置文件内标注记录 train.txt, val.txt 的具体路径以及图片的分类。执行 trian.py 和 dectect.py 进行训练和预测。

```
# Dataset configuration for YOLOv5
train: data/steel_defects/train.txt

val: data/steel_defects/test.txt # Or 'test' if you're using a separate test set

nc: 4 # Number of classes

names: ['fray', 'scratches', 'uneven', 'pitted']
```

图 11- 38

（三）、具体操作

找到对应数据集，并使用标注工具自行进行标注（包括类别和坐标）对模型进行训练。训练后使用训练权重对测试图像进行预测。

在技术需求转换中根据关键词填写对应参数，其中数据以数据中心下的实际

数据为准。

填写完毕后，可点击需求实现查看完整代码，确认无误后点击执行并显示结果，待代码运行完毕后即可查看结果。如下所示：

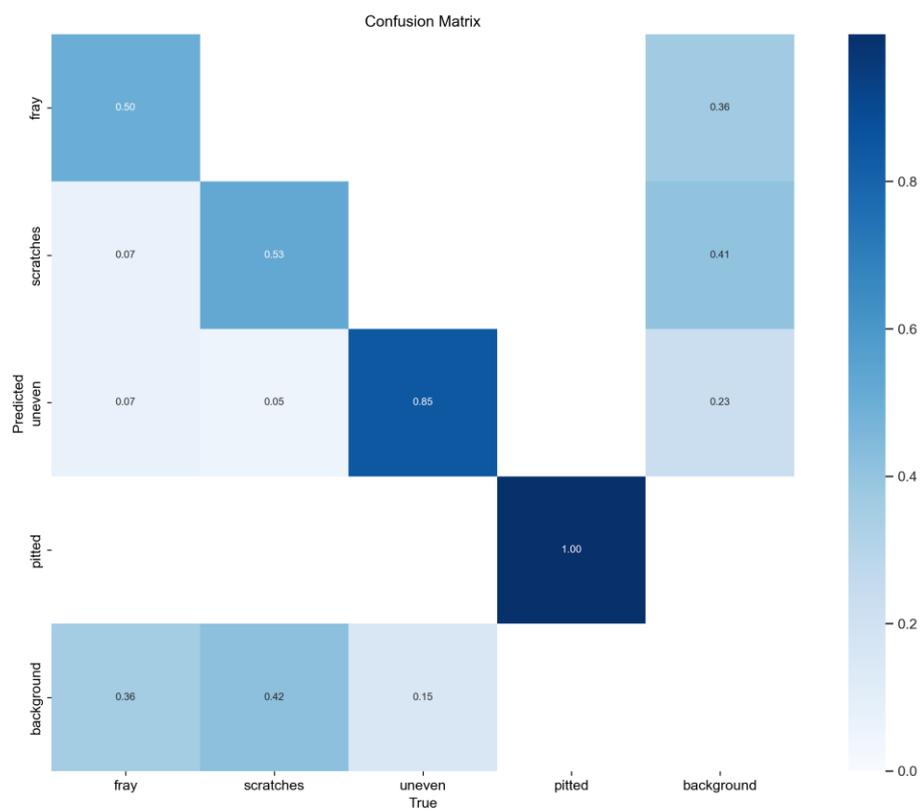


图 11-39

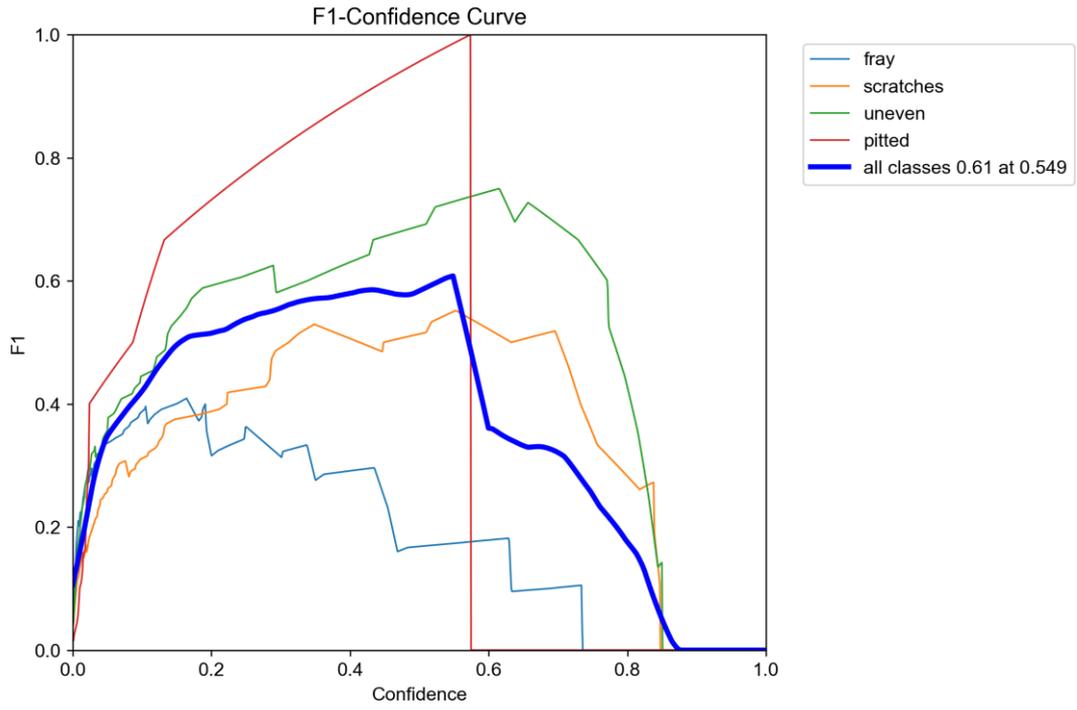


图 11-40

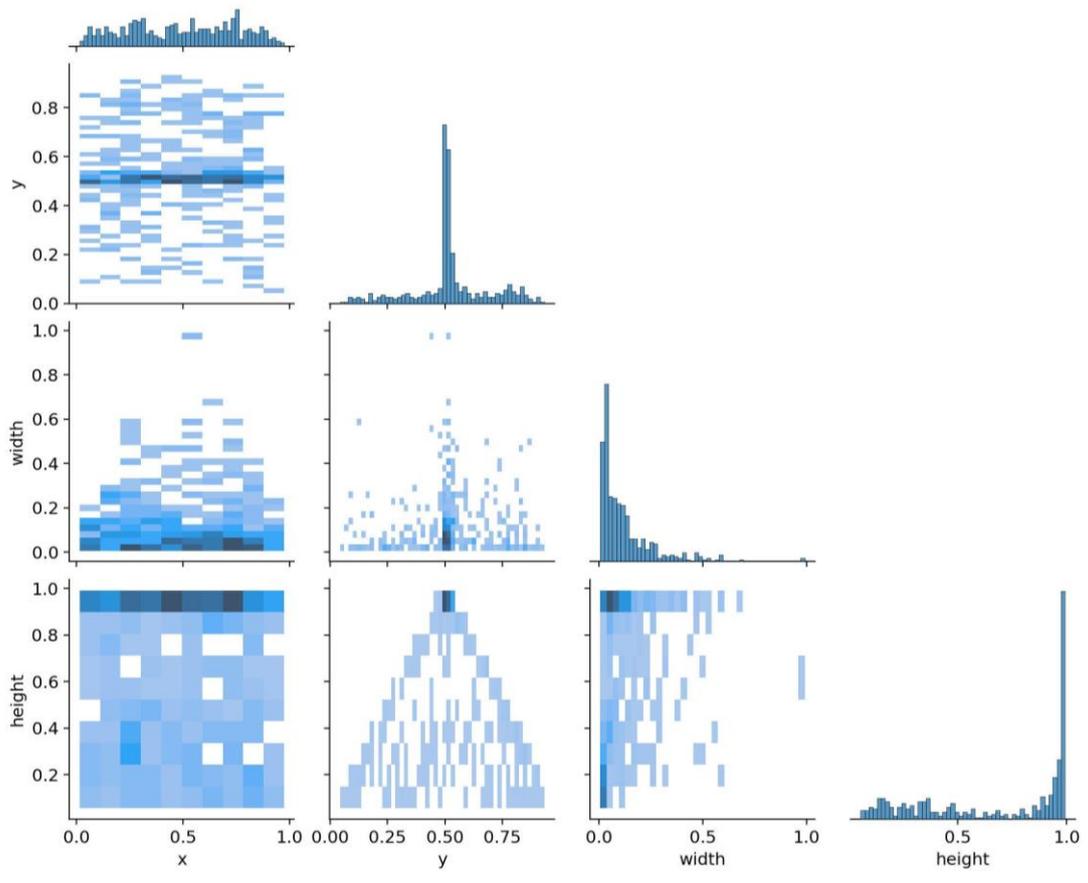


图 11-41

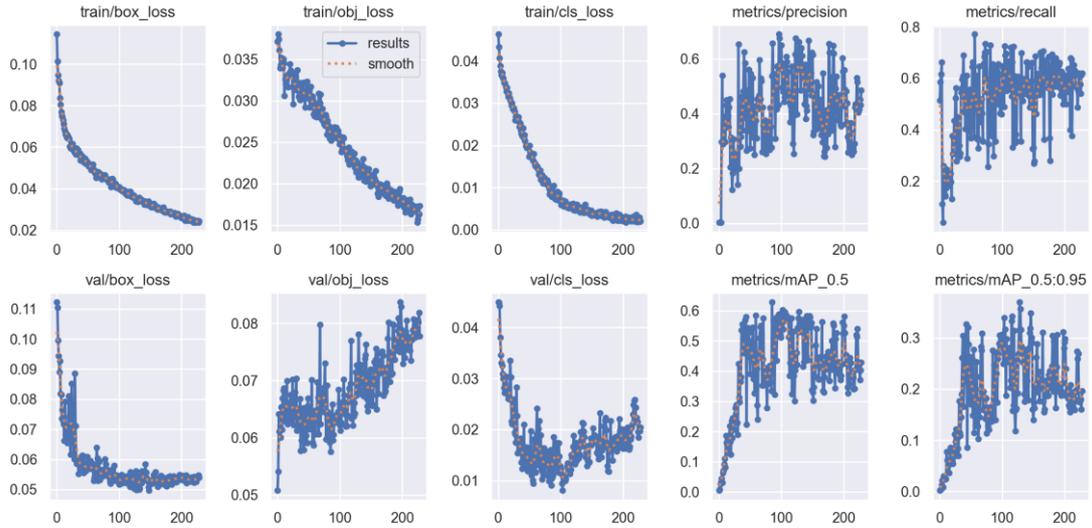


图 11-42

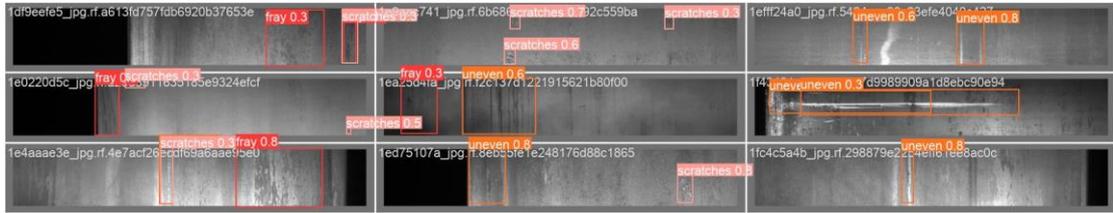


图 11-43

参考答案：

序号	关键词	参数
1	迭代次数	300~500
2	批量大小	16
3	种群大小	50

四、缺陷裂缝分割

(一)、任务说明

在工业焊接质量检测场景中，为精准把控焊点缺陷情况，开展缺陷裂缝分割任务。目标聚焦于利用语义分割技术，精准分割焊点上的裂缝区域并量化缺陷面积。数据准备时，可手动为焊点裂缝标注分割掩膜生成二值图像，若有 DAGM 等现成工业缺陷数据集（含标注好的缺陷区域与类别）则可直接取用，省去手动标注流程；模型构建选择经典的 U-Net 网络执行语义分割，以交叉熵损失（BCE）

为损失函数，引导模型学习裂缝与非裂缝区域的差异；模型评估环节，通过观察分割结果，计算 IoU（交并比）衡量分割区域与真实区域的重叠度，计算面积误差评估缺陷面积量化的精准性，助力工业生产中更高效、准确地识别和分析焊点裂缝缺陷，为焊接质量管控与生产优化提供技术支持。

（二）、实现步骤

交叉熵损失：

```
def CE_Loss(inputs, target, cls_weights, num_classes=21):
    n, c, h, w = inputs.size()
    nt, ht, wt = target.size()
    if h != ht and w != wt:
        inputs = F.interpolate(inputs, size=(ht, wt), mode="bilinear", align_corners=True)

    temp_inputs = inputs.transpose(1, 2).transpose(2, 3).contiguous().view(-1, c)
    temp_target = target.view(-1)

    CE_loss = nn.CrossEntropyLoss(weight=cls_weights, ignore_index=num_classes)(temp_inputs, temp_target)
    return CE_loss
```

图 11-44

Anaconda 终端 pip install labelme 下载标注插件

执行 labelme 命令打开插件可视化窗口，选择要进行标注的图片文件夹，对缺陷位置进行打点标注并标记类别

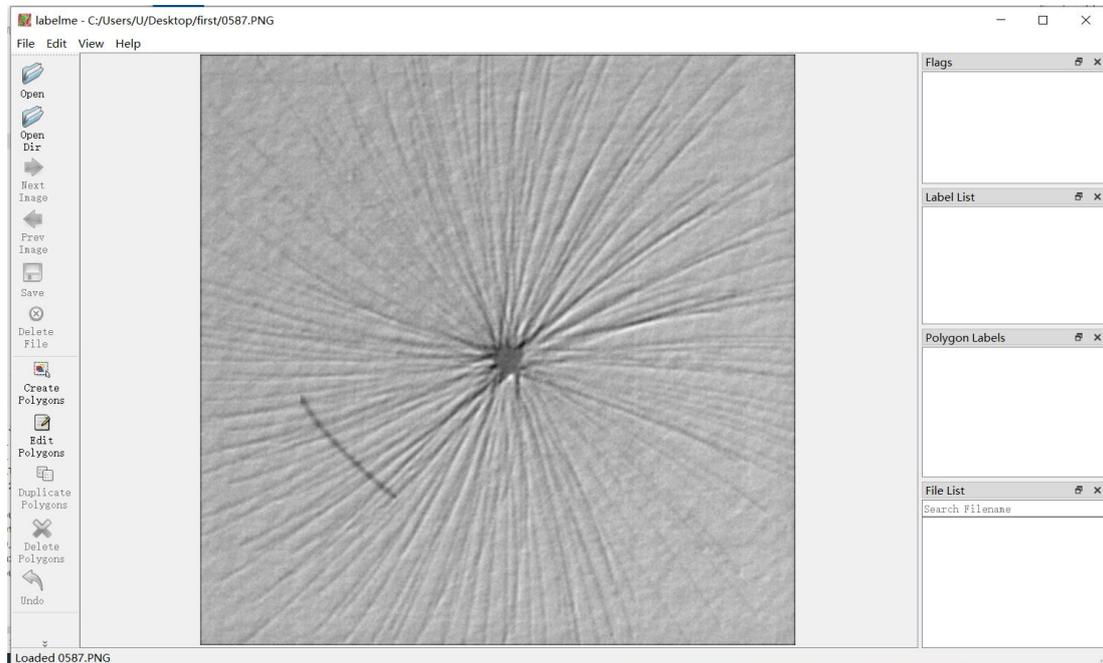


图 11-45

注意：填写超参数以及类别时，请阅读灰色标注再填写，避免参数错误。

训练前，在 unet.py 文件内修改权重以及分类的类别数目，别漏加”_background_”。将原始图片和标注文件放入 dataset 文件夹内（jepgimages 放原图，before 放原图和标注完的 json 文件）后修改并执行 json_to_dataset 文件将 json 文件转换为 png 文件。将转换后的标注文件以及原图放入 VOC2007 文件夹的 imagesets 以及 segmentclass 中，执行 vocannotations 对训练集进行划分，修改 train.py 里面的超参数即可训练。同样的，执行 predict.py 以进行预测。



```
13 具体错误为: Too many dimensions: 3 > 2
14 安装方式为命令行 pip install labelme==3.16.7
15 2、此处生成的标签图是8位彩色图，与视频中看起来的数据集格式不太一样。
16 虽然看起来是彩图，但实际上只有8位，此时每个像素点的值就是这个像素点所属的种类。
17 所以其实和视频中VOC数据集的格式一样。因此这样制作出来的数据集是可以正常使用的。也是正常的。
18 '''
19 > if __name__ == '__main__':
20     jpgs_path = "datasets/JPEGImages"
21     pngs_path = "datasets/SegmentationClass"
22     classes = ["1", "4", "6", "_background_"]
23
24     count = os.listdir("./datasets/before/")
25     for i in range(0, len(count)):
26         path = os.path.join("./datasets/before", count[i])
27
28         if os.path.isfile(path) and path.endswith('json'):
29             data = json.load(open(path))
30
31             if data['imageData']:
32                 imageData = data['imageData']
33             else:
34                 imagePath = os.path.join(os.path.dirname(path), data['imagePath'])
35                 with open(imagePath, 'rb') as f:
36                     imageData = f.read()
37                     imageData = base64.b64decode(imageData).decode('utf-8')
38
39             img = utils.img_b64_to_arr(imageData)
40             label_name_to_value = {'_background_': 0}
```

图 11-46

修改 get_miou.py 内参数并执行，可以对测试集进行 miou 计算，结果如下

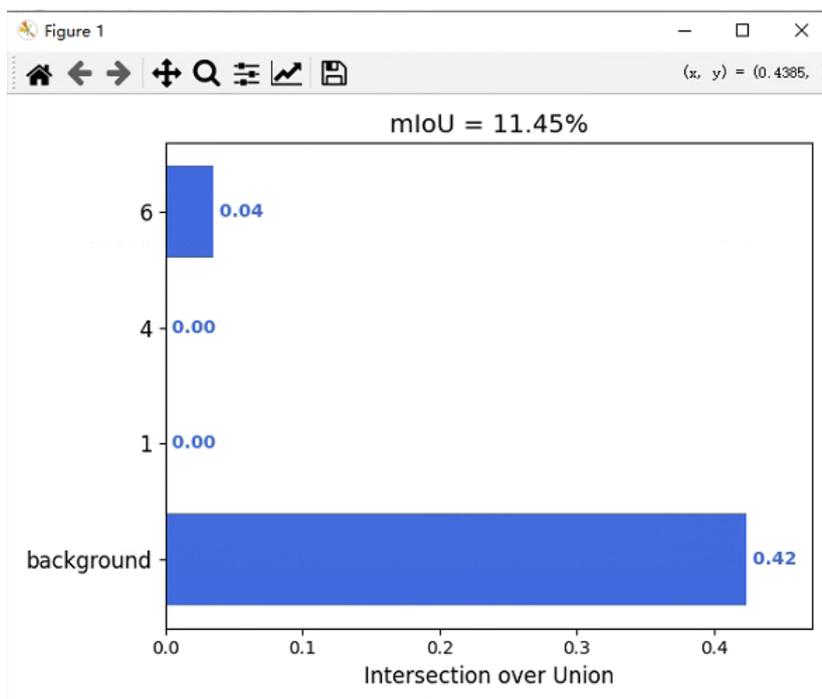


图 11-47

(三)、具体操作

对数据集进行标注，使用 U-NET 网络对数据集图片进行语义分割，得到 IoU 以及面积误差。

在技术需求转换中根据关键词填写对应参数，其中数据以数据中心下的实际数据为准。

填写完毕后，可点击需求实现查看完整代码，确认无误后点击执行并显示结果，待代码运行完毕后即可查看结果。如下所示：

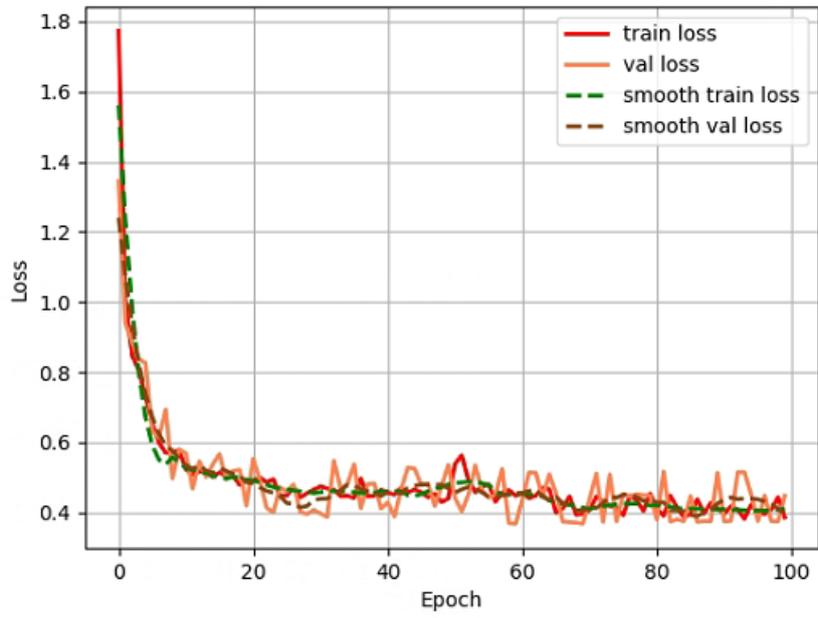


图 11- 48

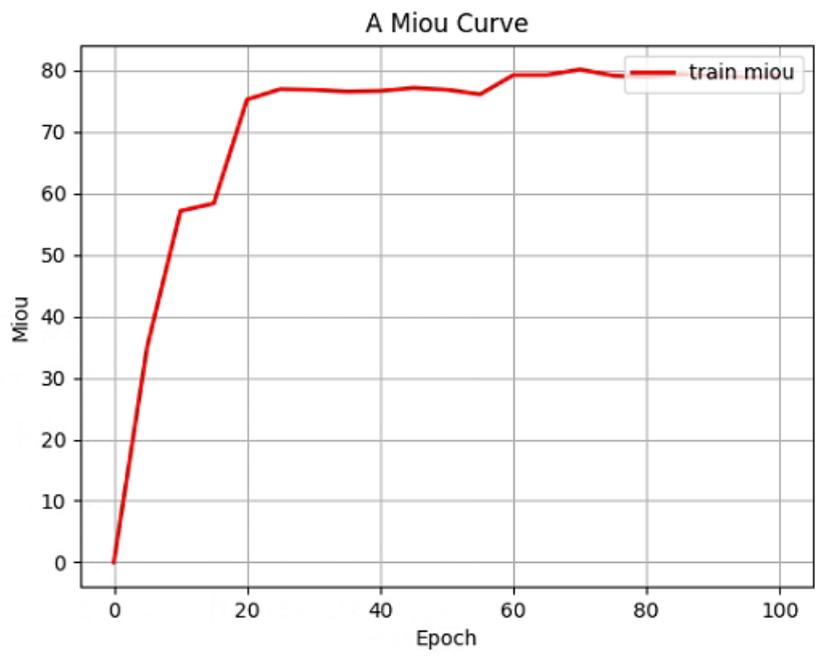


图 11- 49

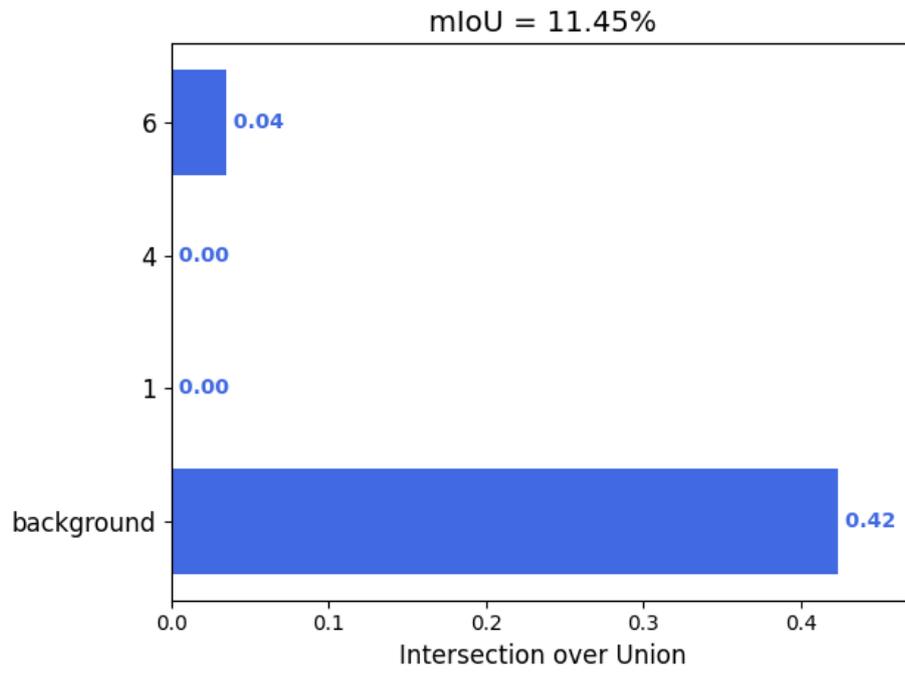


图 11- 50

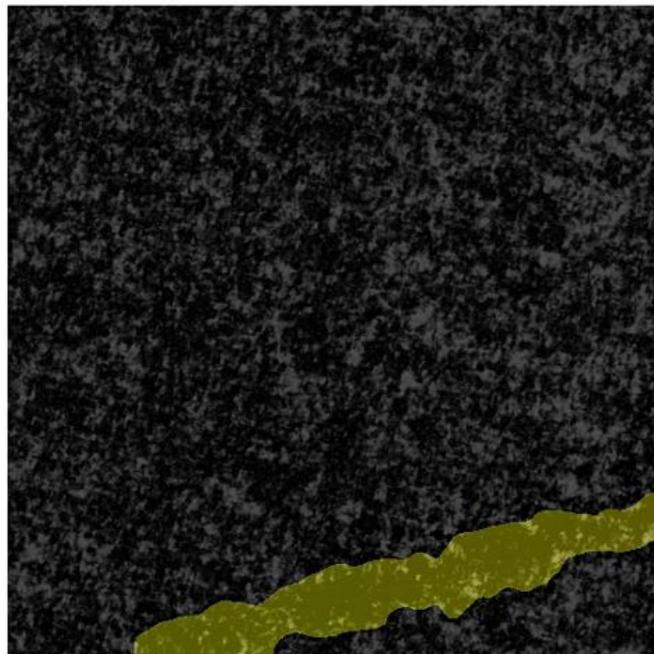


图 11- 51

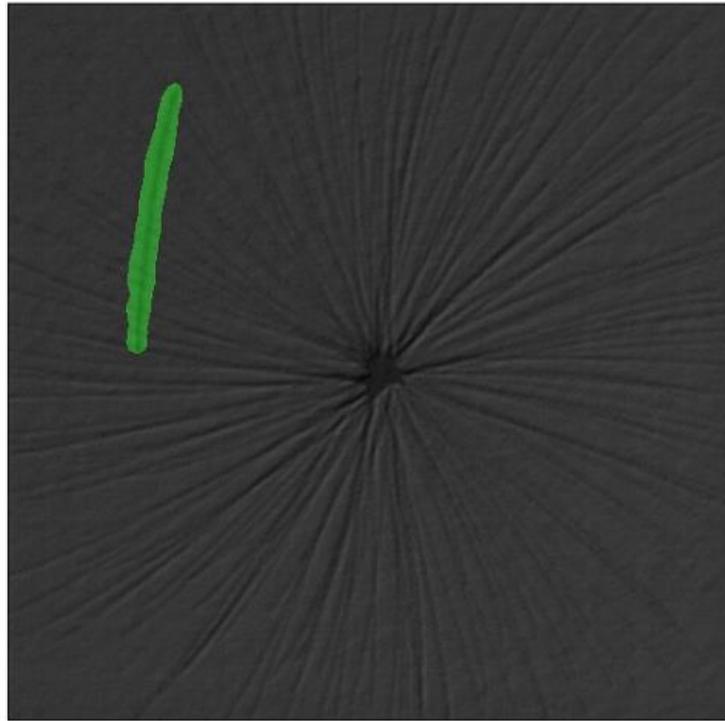


图 11- 52

参考答案:

序号	关键词	参数
1	模型当前开始的训练世代	0
2	模型冻结训练的时代	50
3	是否进行冻结训练	True

项目十二：自然语言处理与应用

任务一 自然语言处理的概念

一、理论知识实训

1. 自然语言处理(NLP)的核心目标是：

- A: 提高计算速度
- B: 优化数据存储
- C: 实现人机语言交互
- D: 增强图像识别能力

标准答案：C

2. 下列哪项不属于 NLP 的两个主要方面？

- A: 自然语言理解
- B: 自然语言生成
- C: 文本情感分析
- D: 语音合成

标准答案：D

3. "bank"一词在 NLP 中面临的挑战是：

- A: 词法歧义
- B: 句法歧义
- C: 语义漂移
- D: 数据稀疏

标准答案：A

4. 解决同义词问题("happy"和"joyful")的最佳方法是：

- A: 规则词典
- B: 词嵌入技术
- C: 句法分析树
- D: 字符级编码

标准答案：B

5. 网络用语"yyds"的流行反映了 NLP 的哪个挑战？

- A: 词法歧义
- B: 句法复杂性
- C: 数据噪声
- D: 语言动态性

标准答案: D

6. BERT 模型主要解决 NLP 中的什么问题?

- A: 数据存储
- B: 计算效率
- C: 语言歧义
- D: 硬件兼容

标准答案: C

7. 下列哪项是 NLP 的典型应用?

- A: 机器翻译
- B: 图像分类
- C: 视频渲染
- D: 3D 建模

标准答案: A

8. 语义漂移现象的本质是:

- A: 数据错误
- B: 语言灵活性
- C: 算法缺陷
- D: 硬件限制

标准答案: B

9. 处理"元宇宙"等新词的最佳方法是:

- A: 固定词典
- B: 语法规则
- C: 图像辅助
- D: 更新语料库

标准答案: D

10. 从 Siri(2011)到 ChatGPT(2023)的发展体现了:

- A: 硬件进步

- B: 数据量减少
- C: 方法论的变革
- D: 算法简化

标准答案: C

二、提示词工程与评价指标

(一)、提示词工程与本章概述

1. 实训目标

在智能制造和汽车相关的问答系统中，结合提示词工程与知识库检索的技术，已经成为提升系统性能和智能化水平的关键。随着人工智能技术的快速发展，尤其是自然语言处理（NLP）领域的突破，基于大规模语言模型的问答系统已成为现代智能制造和汽车领域中不可或缺的组成部分。为了更好地解决这一领域中专业性强、信息庞杂、实时性要求高等问题，提示词工程与知识库检索的结合发挥了重要作用。

2. 知识库检索的作用

知识库检索技术通过在一个大型的结构化或非结构化数据库中寻找相关信息，能够为问答系统提供准确的背景信息和参考资料。智能制造和汽车领域中，企业常常拥有大量的技术文档、操作手册、服务记录等，这些信息包含了非常宝贵的行业知识。结合知识库检索，可以在用户提问时，迅速从知识库中找到相关的标准答案或技术支持文档，再结合提示词工程输出更加符合用户需求的回答。

例如，针对“当前某车型的发动机出现异响如何诊断”的问题，系统不仅可以从训练数据中生成回答，还可以检索到车辆维修手册中的相关故障排查步骤，进而结合提示词生成具体的诊断流程和注意事项。

3. 提示词工程与知识库检索的结合

在实际应用中，提示词工程与知识库检索的结合，能够显著提升问答系统的准确性和可靠性。具体来说，提示词工程有助于准确引导系统从知识库中检索出相关信息，而知识库检索则为系统提供了更加精准、可靠的背景信息，避免了依赖模型自生成内容可能导致的错误或模糊回答。

流程化问题解决：通过设计针对性的提示词，结合知识库中已有的解决方案，可以帮助用户快速解决问题，避免繁琐的查找过程。例如，在智能制造领域，操作人员可以根据提示词得到设备维修的步骤、常见故障代码及处理方法。

增强专业性和实用性：对于技术性较强的领域(如汽车维修、自动化系统等)，知识库中的内容通常包含大量专业术语和复杂的技术流程。通过提示词的引导，可以帮助模型识别并准确生成专业化的回答，而不仅仅是停留在通用语言层面。

4. 平台中的应用

在本章与下一章中，我们将引导学生学习如何结合提示词工程和知识库检索技术，构建一个汽车领域的问答系统。这不仅是对语言模型应用能力的提升，也涉及到如何高效地管理和利用大量行业知识。

接下来的内容将深入探讨知识库的构建方法，如何设计有效的提示词以及如何整合这些技术形成一个完整的问答系统。我们将逐步介绍如何使用现有的文档数据、技术标准和行业经验，构建一个能够应对真实世界问题的智能问答系统，并利用提示词和检索技术来提升系统的响应速度与准确性。

本章作为入门篇，将为学生们展示提示词工程和知识库检索在汽车领域中的巨大潜力。通过学习这些技术，学生将能够掌握如何利用现代 AI 技术解决实际问题，为今后的研究和应用打下坚实的基础。

5. 任务描述

围绕提示词工程、知识库构建和知识库检索三个主题展开，重点考察以下内容：

1) 提示词工程：优化提示词以提升大语言模型的生成效果，减少翻译幻觉，并提高模型对复杂或模糊问题的处理能力。

2) 知识库构建：探索有效的索引方法、去除冗余信息的技术，以及多语言统一表示的实现方式。

3) 知识库检索：聚焦于提升检索效率、优化语义匹配，以及解决查询与知识库内容语义不一致的问题。

每道题目附有答案和解析，可深入考察相关技术原理及应用场景，适合初学者巩固学习。

练习题及答案汇总：

1. 提示词工程的主要目标是什么？

A. 训练一个全新的大语言模型

- B. 优化计算资源的消耗，降低模型运行成本
- C. 通过精心设计输入文本来引导和控制大模型的输出
- D. 加速大模型生成文本的速度

标准答案：C

2. 在构建 Few-shot 学习提示词时，为了减少模型的“幻觉”现象，应优先关注？

- A: 提示词的数量与长度
- B: 提示词中示例的多样性与相关性
- C: 提示词中的语法一致性
- D: 提示词中符号使用的一致性

标准答案：B

3. 下列哪种方法最适合通过提示词工程提高模型对模糊问题的回答质量？

- A: 使用具体、详细的示例
- B: 提供多种可能答案的示例
- C: 引入更多复杂的语法结构
- D: 将提示词转为模型训练时的语种输入

标准答案：A

4. 在知识库构建的知识表示过程中，下列方法中哪种适合多语种信息的统一表示？

- A: Bag of Words (BoW)
- B: Word2Vec
- C: Multilingual Embedding
- D: TF-IDF

标准答案：C

5. 在构建知识库时，为了提高语义理解能力并减少歧义，常使用的技术是？

- A: 特征工程
- B: 偏差修正算法
- C: 语义嵌入
- D: 扩展词典

标准答案：C

5. 在基于语义匹配的知识库检索中，最常用的深度学习模型是？

A: RNN

B: LSTM

C: Transformer

D: KNN

标准答案：C

7. 在知识库检索任务中，以下哪种方法最能缓解查询词语义与知识库内容语义不一致的问题？

A: 提高索引的深度

B: 使用上下文增强的语义匹配模型

C: 引入更多训练样本

D: 简化查询的表达方式

标准答案：B

8. 在知识库检索过程中，若采用基于向量空间模型的检索，最常用的计算文档与查询相似度的度量方法是？

A: 曼哈顿距离

B: 汉明距离

C: 欧几里得距离

D: 余弦相似度

标准答案：D

9. 在进行知识库检索时，使用 BERT 进行查询和文档的语义匹配时，最可能面临的挑战是？

A: 计算资源消耗过大

B: 查询的上下文无法准确传递

C: 查询和文档之间的语法差异太大

D: 查询语句难确定

标准答案：A

10. 在大规模知识库构建过程中，若希望通过自动化的方式有效整合来自不

同数据源（如文本、结构化数据、图像等）的异构信息，并确保知识库的扩展性和高效性，哪种方法最为合适？

A: 基于自监督学习的多模态预训练模型（Self-supervised Multimodal Pretraining Models）

B: 混合式信息抽取框架（Hybrid Information Extraction Framework）

C: 层次化知识库聚合策略（Hierarchical Knowledge Base Aggregation Strategy）

D: 基于神经符号方法的知识图谱推理（Neural-Symbolic Approach for Knowledge Graph Inference）

标准答案：A

（二）、评价指标与提示词效果

1. 实训目标

本实训的目的是计算输入答案（myanswer）与一组参考答案（retrieved_answers）之间的相似度。为此，我们采用基于 BERT 预训练模型的文本相似度评分方法——BERTScore。通过该方法，不仅能够衡量文本之间的表面匹配度，还能捕捉文本的语义相似性。本节实训对比的示例，分别来自于没有使用提示词工程和知识库检索问答辅助的模型输出，以及应用了提示词工程和知识库检索问答辅助的模型输出。

2. 主要步骤

1) 输入数据预处理：首先，输入文本和参考答案需要经过分词处理。由于任务是中文文本，因此我们使用 jieba 分词工具对文本进行分词。分词后，文本会以词语之间空格分隔的格式进行处理，以便输入到 BERT 模型中进行相似度计算。

2) BERTScore 计算：使用 bert-score 库中的 score 函数计算每个参考答案与输入答案之间的 Precision (P)、Recall (R) 和 F1 分数。这三个分数分别衡量预测文本与参考文本之间的精确度、召回率以及综合表现。

3) 输出最相似答案：基于计算得到的 F1 分数，我们找到与输入答案最相似的参考答案，作为最佳匹配答案，并输出其相似度得分。

3. 函数与代码实现

- `clean_and_tokenize_zh(text)`: 这是一个用于中文分词的辅助函数，利用 `jieba` 库将文本分词并返回空格分隔的词语列表。

- `evaluate_answer_similarity(myanswer, retrieved_answers)`: 这是主函数，用于计算输入答案与参考答案列表的相似度。该函数会返回最相似的参考答案、其相似度得分，以及所有参考答案的得分。

输入:

- `myanswer`: 输入答案（例如，用户的回答或机器生成的答案）。
- `retrieved_answers`: 参考答案列表，通常是来自知识库或从先前的查询中检索到的多个答案。

4. 结果说明

在执行代码时，您会看到输出包含以下内容:

- 输入答案: 显示您输入的答案文本。
- 最相似答案: 显示与输入答案最相似的参考答案。
- 最高相似度得分: 显示最相似答案的相似度得分（基于 `BERTScore` 的 F1 分数）。

通过这种方式，可以快速判断输入结果的质量和与参考答案的相关性，接下来由您依据注释的提示补充完整代码并运行代码。

5. 编程实现

【底码与参考答案中的代码仅供参考，RAG 技术体验可在产业案例-AI 模型体验-任务四中进行】

```

import jieba
import torch
from bert_score import score

# 中文分词函数, 使用jieba进行分词
def clean_and_tokenize_zh(text):
    # 使用jieba进行中文分词, 返回分词后的词语列表
    return ' '.join(jieba.cut(text))

# 计算BERTScore相似度并输出最相似的参考答案
def evaluate_answer_similarity(myanswer, retrieved_answers):
    """
    根据输入答案和参考答案列表计算BERTScore相似度得分, 并输出最相似

    Args:
    - myanswer (str): 输入的答案。
    - retrieved_answers (list): 参考答案列表。

    Returns:
    - best_answer (str): 最相似的答案。
    - best_score (float): 最相似答案的得分。
    - scores (list): 所有答案的相似度得分。
    """
    # 分词并计算BERTScore
    myanswer = clean_and_tokenize_zh(myanswer)
    retrieved_answers = [clean_and_tokenize_zh(ans) for ans in

    # 计算BERTScore
    P, R, F1 = score([myanswer] * len(retrieved_answers), retr

    # 计算每个参考答案的F1分数
    f1_scores = F1.tolist()

    # 找到得分最高的答案
    best_score = max(f1_scores)
    best_answer = retrieved_answers[f1_scores.index(best_score

    return best_answer, best_score, f1_scores

# 测试数据
myanswer = "不同的车型和品牌可能有不同的操作步骤, 但大多数现代汽车

retrieved_answers = [
    '在中央显示屏主界面上边缘向下滑动屏幕, 进入车辆功能界面。',
    '您可以通过按下车辆上的多功能方向盘上的按钮或者触摸中央显示屏来',
    '可以通过从中央显示屏主界面上边缘向下滑动屏幕的方式进入车辆功能',
    '您可以通过从中央显示屏主界面上边缘向下滑动屏幕的方式进入车辆功
]

# 调用BERTScore计算函数
best_answer, best_score, scores = evaluate_answer_similarity(m

# 输出结果
print(f"输入答案: {myanswer}")
print(f"最相似答案: {best_answer}")
print(f"最高相似度得分: {best_score:.4f}")

```

```

[ ]:
# 测试数据
myanswer = "从中央显示屏主界面上边缘向下滑动屏幕, 进入车辆功能界面。
retrieved_answers = [
    '在中央显示屏主界面上边缘向下滑动屏幕, 进入车辆功能界面。',
    '您可以通过按下车辆上的多功能方向盘上的按钮或者触摸中央显示屏来',
    '可以通过从中央显示屏主界面上边缘向下滑动屏幕的方式进入车辆功能',
    '您可以通过从中央显示屏主界面上边缘向下滑动屏幕的方式进入车辆功
]

# 调用BERTScore计算函数
best_answer, best_score, scores = evaluate_answer_similarity(m

# 输出结果
print(f"输入答案: {myanswer}")
print(f"最相似答案: {best_answer}")
print(f"最高相似度得分: {best_score:.4f}")

```

图 12- 1

步骤分析:

单选题 (单选题 5.26 分)

1. 以下关于代码中 `evaluate answer similarity` 函数中计算和输出的内容, 哪一项描述是正确的?

A. `score` 函数返回的 Precision(P) 和 Recall(R) 分数可以单独用于选择最相似的参考答案, 而不需要考虑 F1 分数。

B. `score` 函数计算的 F1 分数是基于输入答案与所有参考答案中每个词汇的匹配情况, 最终得分越高表示输入答案和参考答案之间的语义相似度越高。

C. 输入答案与参考答案之间的相似度计算是基于词频 (TF) 和逆文档频率 (IDF) 方法, 不使用 BERT 模型进行语义比较。

D. `clean and tokenize zh` 函数将文本中的每个字符进行分割, 这与标准的中文分词方式不同, 可能会影响相似度计算的结果。

正确答案: B

解析:

A: 错误。虽然 Precision 和 Recall 分别衡量不同方面的相似度, 但最终还是通过 F1 分数 (Precision 与 Recall 的综合) 来选择最相似的答案。单独使用 Precision 或 Recall 可能无法准确反映文本的整体相似度。B: 正确。BERTScore 的 F1 分数通过比较输入答案和参考答案中每个词汇的匹配情况 (基于 BERT 模型的语义嵌入) 来计算, 最终得分越高表示语义相似度越高。C: 错误。代码中使用的是 BERT 模型来计算文本相似度, 而不是基于传统的 TF-IDF 方法。BERTScore 是基于 BERT 模型的语义嵌入来评估文本相似度。D: 错误。`clean and tokenize zh` 函数是使用 `jieba` 进行中文分词, 它会将文本切分为词语而非单个字符, 这样有助于更准确地处理中文文本并进行相似度计算。

任务二 自然语言处理的基本任务

一、理论知识实训

1. 中文分词任务 "自然语言处理很有趣" 的正确切分是:

A: ["自", "然", "语"]

B: ["自然语言处理", "很", "有趣"]

C: ["自然", "语言"]

D: ["自然", "语", "言"]

标准答案: B

2. 词性标注任务中"run"在句子"I run daily"中的正确标注是:

A: 名词

B: 形容词

C: 动词

D: 副词

标准答案: C

3. 命名实体识别(NER)的主要功能是:

A: 分析句子情感

B: 修正语法错误

C: 生成同义词

D: 识别专有名词

标准答案: D

4. 句法分析任务"The cat chased the mouse"中"the mouse"的成分是:

A: 宾语

B: 定语

C: 状语

D: 主语

标准答案: A

5. 文本分类任务"这款手机电池续航差"最可能的标签是:

A: 价格投诉

B: 负面评价

C: 功能咨询

D: 广告推广

标准答案: B

6. 自然语言推理 (NLI) 任务中, "前提: 他在吃饭" 与 "假设: 他在用餐" 的关系是:

A: 矛盾

B: 中立

C: 蕴含

D: 无关

标准答案: C

7. 阅读理解任务中, 文章提到 "会议在巴黎举行", 问题 "会议地点?" 的答案是:

A: 伦敦

B: 线上

C: 未提及

D: 巴黎

标准答案: D

8. 生成类任务中, 输入 "人工智能" 后最可能的扩展是:

A: "正在改变世界"

B: "长度 5 个字符"

C: "A. I. 缩写"

D: "12345"

标准答案: A

9. 对话系统中用户问 "明天会下雨吗?", 系统的最佳回应是:

A: "我不喜欢下雨"

B: "明天多云转小雨"

C: "今天是晴天"

D: "你猜"

标准答案: B

10. 机器翻译任务中 "How old are you?" 的正确中文翻译是: A: "你好"

B: "你几岁了"

C: "你多大了"

D: "你老了吗"

标准答案: C

二、汽车问答数据库预处理

(一)、PDF 文本提取

1. 实训目标

- 了解 PDF 文件的文本提取技术: 掌握如何使用 pdfplumber 和 PyPDF2 等工具从 PDF 文件中提取文本内容。

- 实现数据预处理和可视化展示: 通过可视化步骤, 理解数据处理过程中的每个阶段, 帮助理解数据提取后的效果。

- 掌握函数实现与代码编写: 通过编写和调试代码, 帮助理解如何实现文件下载、文本提取及数据处理。

- 理解数据流与处理过程: 通过逐步展示每个处理步骤的数据变化, 培养学生的程序调试和数据分析能力。

2. 主要步骤

- PDF 文件下载: 使用 FstWeb 类的 download_public 方法从指定路径下载 PDF 文件。此步骤涉及到文件的获取和路径传递。

- 文本提取: 使用 pdfplumber 库从 PDF 文件中提取文本内容。每一页的文本被提取出来, 并储存到一个列表中。

- 可视化处理效果: 在每个步骤之后使用 visualize 函数可视化数据的处理效果。visualize 函数将输出当前处理步骤的名称、文本内容预览, 并显示处理后的数据大小。

- 数据存储: 提取到的文本内容被存储在 data 列表中, 用于后续的处理。

3. 函数与代码实现

- visualize(step_name, content_preview=None)

作用: 可视化每个处理步骤的效果, 包括处理步骤的名称、预览内容, 以及处理后的数据量。

输入: step_name: 当前步骤的名称; content_preview: 可选, 当前步骤的

数据预览。

输出：打印步骤名称、数据预览、数据大小等信息。

- `extract_pdf_text(pdf_path)`

作用：提取 PDF 文件中的文本内容。使用 `pdfplumber` 库对 PDF 文件的每一页进行文本提取。

输入：`pdf_path`：PDF 文件路径。

输出：`all_content`：包含 PDF 文件所有页面文本内容的列表。

- `FstWeb().download_public('文件名称.xlsx')`

作用：下载指定名称的 PDF 文件。

输出：返回 PDF 文件路径，用于后续的文本提取。

- `pdfplumber.open(pdf_path)`

作用：打开 PDF 文件并为每一页提取文本。

4. 结果说明

- 提取的文本：在 `extract_pdf_text` 函数中，使用 `pdfplumber` 从每一页 PDF 文件中提取文本，并将结果存储在 `all_content` 列表中。通过 `visualize` 函数输出文本预览，帮助检查提取效果。

- 数据预览：`visualize` 函数打印每个步骤的数据预览，便于理解每个阶段的输出内容。

- 数据量：`visualize` 还会显示处理后的数据量，帮助判断数据处理的规模。

5. 编程实现

【底码与参考答案中的代码仅供参考，RAG 技术体验可在产业案例-AI 模型体验-任务四中进行】

[1]:

```
from expand import FstWeb
import pdfplumber
from PyPDF2 import PdfReader
import json
import matplotlib.pyplot as plt
def visualize(step_name, content_preview=None):
    """ 可视化每一步的处理效果 """
    print(f"Step: {step_name}")
    if content_preview:
        print("Content preview:")
        for text in content_preview[:5]: # Preview first 5 it
            print(text)
    print(f>Data size after {step_name}: {len(data)}\n")
def extract_pdf_text(pdf_path):
    """ 提取PDF文本 """
    all_content = []
    with pdfplumber.open(pdf_path) as pdf:
        for page in pdf.pages:
            text = page.extract_text()
            if text:
                all_content.append(text)
    visualize("Text Extraction", all_content) # 可视化提取
    return all_content
pdf_path = await FstWeb().download_public('train_a.pdf')
data = [] # 存储所有提取的数据

# 1. 提取PDF文本
pdf_text = extract_pdf_text(pdf_path)
```

图 12-2

步骤分析:

单选题(单选题 5.26 分)

1. 在 extract pdf texti 函数中, text=page.extracttext0 的作用是什么?

- A. 提取当前页面的所有图像
- B. 提取当前页面的文本内容
- C. 提取当前页面的元数据
- D. 获取当前页面的文本格式

正确答案: B

单选题（单选题 5.26 分）

2. 在代码中，`text=page.extract_text()` 提取文本时，有时可能会遇到页码格式异常或 PDF 文件损坏等问题，导致 `extract_text()` 返回 `None`。如果想确保程序能够健壮地处理这种情况，并且只将有效的文本内容存储到 `all content` 列表中，应该如何修改代码？

- A. 在 `extract text0` 后加入判断条件，忽略 `None` 值并跳过该页的处理
- B. 修改代码，使得如果 `extract text()` 返回 `None` 时，直接将 `None` 添加 `all-content` 中
- C. 使用 `try-except` 块来捕捉错误，忽略 `None` 值并继续执行
- D. 增加一个日志记录系统，记录每个返回 `None` 的页面，并终止程序执行

正确答案：A

（二）、数据清洗与过滤

1. 实训目标

本实训旨在让学生掌握如何对提取的文本进行清洗和预处理。通过统计字数变化并进行可视化，学生将学习如何评估数据清洗对文本的影响，掌握常见的文本预处理技巧，并提高数据分析能力。

2. 主要步骤

• 提取 PDF 文本

学生将使用 `pdfplumber` 库从指定的 PDF 文件中提取每一页的文本内容，并统计每一页的字数。

• 数据清洗

在这一阶段，学生将学习如何对提取的文本进行清洗。具体操作包括去除不必要的字符（如换行符、制表符等）、根据不同的分隔符对文本进行分割，并根据字数限制筛选合适的文本片段。通过这一过程，学生可以提高数据质量，为后续分析提供更为干净的文本数据。

• 字数对比可视化

最后，学生将使用 `matplotlib` 库绘制图表，比较清洗前后每页的字数变化。通过对比图，学生能够直观了解数据清洗的效果，进一步加深对数据清洗过程和其重要性的理解。

3. 函数与代码实现

- 文本提取函数

该函数的功能是从 PDF 文件中提取每一页的文本，并计算每页的字数。文本内容将被存储在一个列表中，同时记录每一页的字数，供后续分析使用。

- 数据清洗函数

该函数负责对提取的文本进行清洗。清洗的内容包括去除无关字符、根据符号分割文本、并根据字数限制筛选符合条件的句子。最终返回清洗后的文本和更新后的字数。

- 字数对比可视化函数

通过该函数，学生可以生成一个折线图，展示清洗前后每页字数的变化。这将帮助学生直观理解数据清洗对文本长度的影响，并评估清洗效果。

4. 结果说明

- 数据清洗效果

通过清洗过程，学生能够去除文本中的无关信息，如多余的空格、换行符、特殊符号等。清洗后的文本质量较高，且每页字数会有一定程度的减少或优化。

- 字数变化可视化

通过可视化展示清洗前后的字数对比，学生可以直观地看到数据清洗对文本长度的影响，从而理解数据清洗在实际应用中的重要性。通过这种方式，学生不仅能够掌握数据处理技巧，还能提高分析和呈现数据结果的能力。

5. 编程实现

【底码与参考答案中的代码仅供参考，RAG 技术体验可在产业案例-AI 模型体验-任务四中进行】

[2]:

```
import pdfplumber
import matplotlib.pyplot as plt
from expand import FstWeb
# 提取PDF文本并统计每页的字数
def ExtractPDFText(pdf_path):
    with pdfplumber.open(pdf_path) as pdf:
        text_data = []
        page_lengths = []
        for page in pdf.pages:
            text = page.extract_text()
            text_data.append(text)
            page_lengths.append(len(text.split())) # 统计每页

    return text_data, page_lengths

def Datafilter(text_data, max_seq=1024):
    cleaned_data = []
    page_lengths_after = []

    # 处理每页的文本
    for page_text in text_data:
        cleaned_text = []

        # 以符号“■”、“*”或制表符“\t”为分隔符，分割文本
        if "■" in page_text:
            sentences = page_text.split("■")
        elif "*" in page_text:
            sentences = page_text.split("*")
        elif "\t" in page_text:
            sentences = page_text.split("\t")
        else:
            sentences = page_text.split("。") # 默认使用句号“。”

        # 清理每个句子
        for sentence in sentences:
            sentence = sentence.strip().replace("\n", "").replace("\r", "")
            if len(sentence) > 5 and len(sentence) <= max_seq:
                cleaned_text.append(sentence)

        # 合并清洗后的句子并添加到结果中
        cleaned_data.append("。".join(cleaned_text))
        page_lengths_after.append(len("。".join(cleaned_text).split()))

    return cleaned_data, page_lengths_after

# 可视化字数对比
def plot_word_count_comparison(page_lengths_before, page_lengths_after):
    plt.figure(figsize=(10, 6))
    plt.plot(range(len(page_lengths_before)), page_lengths_before, 'b', marker='o')
    plt.plot(range(len(page_lengths_after)), page_lengths_after, 'r', marker='o')
    plt.title('Word Count Comparison Before and After Cleaning')
    plt.xlabel('Page Number')
    plt.ylabel('Word Count')
    plt.legend()
    plt.show()

# 示例
pdf_path = await FstWeb().download_public('train_a.pdf')

# 步骤 1: 提取原始文本和页数
text_data, page_lengths_before = ExtractPDFText(pdf_path)

# 步骤 2: 数据清洗
cleaned_data, page_lengths_after = Datafilter(text_data, max_seq=1024)

# 步骤 3: 对比清洗前后的字数
plot_word_count_comparison(page_lengths_before, page_lengths_after)
```

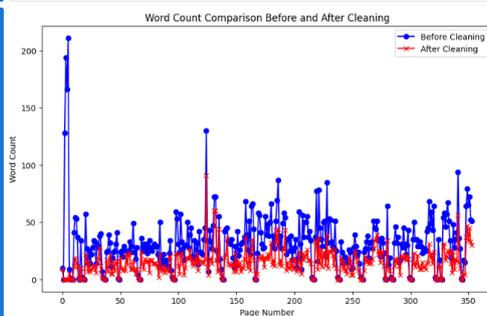


图 12-3

（三）、滑窗法分段提取

1. 实训目标

本任务的主要目标是帮助学生理解并实现滑窗法 (Sliding Window)，并将其应用于文本数据分段提取。通过该任务，学生将掌握以下技能：

- 滑窗法的原理与实现；
- 如何在文本处理中应用滑动窗口技术；
- 使用 Python 进行文本操作和数据结构管理。

2. 主要步骤

- 初始化参数：确定窗口大小、步长等基本参数，并准备好存储结果的空列表 data。

- 文本拼接与滑动：通过一个 while 循环来遍历句子，并根据滑窗法判断是否将当前句子与前一个句子拼接。如果拼接后的文本长度超过窗口大小，则将其加入 data 列表。

- 更新窗口：当拼接后的文本超过窗口大小时，移除窗口左侧的部分，以保证窗口大小不变。

- 返回处理结果：函数返回根据滑窗法提取的文本段落，并对前 5 个句子进行可视化展示，帮助学生理解文本分段的效果。

3. 函数与代码实现

- 参数：

sentences: 输入的句子列表，函数通过遍历这些句子来进行分段。

kernel: 窗口大小，表示每段的最大长度（以字符数计算）。默认值为 512。

stride: 步长，表示每次滑动的步幅，默认值为 1。

data: 存储滑窗分段的列表。如果未传入，函数内会初始化为空列表。

返回值：返回一个新的列表 data，包含根据滑窗法分割后的文本段。

- 功能：

利用滑窗法逐步拼接句子，直到拼接后的文本长度超过指定的窗口大小 (kernel)。当文本长度超过窗口大小时，将其加入到 data 列表，并更新窗口左端的文本。通过步长 stride 控制滑动步伐，即每次向后移动多少个句子。

4. 结果说明

- 通过该函数，学生将能够将长文本按照滑窗法的方式进行分段，使得每段的长度符合一定的限制。

- 通过可视化函数，学生能够看到滑窗法在实际文本中的效果，并理解如何根据不同的窗口大小和步长参数，调整文本分段的方式。

5. 编程实现

【底码与参考答案中的代码仅供参考，RAG 技术体验可在产业案例-AI 模型体验-任务四中进行】

```
[2]:
import matplotlib.pyplot as plt
import matplotlib.patches as patches

def SlidingWindow(sentences, kernel=512, stride=1, data=None):
    """ 滑动法分段提取 """
    if data is None:
        data = [] # 如果没有传入 data, 则在函数内定义

    sz = len(sentences)
    cur = ""
    fast = 0
    slow = 0

    while fast < sz:
        sentence = sentences[fast]
        # 判断当前句子拼接后的长度是否超过窗口大小
        if len(cur + sentence) > kernel and (cur + sentence) r
            data.append(cur + sentence + " ") # 将当前句子和上
            # 更新窗口左侧, 保持窗口大小
            cur = cur[len(sentences[slow] + " "):] # 删除滑动
            slow += 1
        cur += sentence + " "
        fast += stride # 步长控制滑动步伐

    # 可视化前5个句子的效果
    visualize("Sliding Window", sentences[:5])
    return data

def visualize(title, sentences):
    """ 可视化滑动法分段提取效果 """

# 示例数据
sentences = [
    "机器学习是一种人工智能技术。",
    "它允许计算机通过数据学习并做出决策。",
    "深度学习是机器学习的一种方法。",
    "它通过模拟人脑结构进行学习。",
    "在很多领域, 深度学习取得了很好的成果。",
    "例如图像识别、自然语言处理等领域。",
    "这些技术的应用正在逐渐改变我们的生活。",
    "未来, 我们会看到更多的创新和突破。"
]

# 使用滑动法提取文本
data = SlidingWindow(sentences, kernel=100, stride=1)

# 打印提取的文本段
print("\n滑动提取的文本段的过程:")
for i, segment in enumerate(data):
    print(f"段 {i + 1}: {segment}")
```

滑动提取的文本段的过程:

段 1: 机器学习是一种人工智能技术。 它允许计算机通过数据学习并做出决策。 深度学习是机器学习的一种方法。 它通过模拟人脑结构进行学习。 在很多领域, 深度学习取得了很好的成果。 例如图像识别、自然语言处理等领域。

段 2: 它允许计算机通过数据学习并做出决策。 深度学习是机器学习的一种方法。 它通过模拟人脑结构进行学习。 在很多领域, 深度学习取得了很好的成果。 例如图像识别、自然语言处理等领域。 这些技术的应用正在逐渐改变我们的生活。

段 3: 深度学习是机器学习的一种方法。 它通过模拟人脑结构进行学习。 在很多领域, 深度学习取得了很好的成果。 例如图像识别、自然语言处理等领域。 这些技术的应用正在逐渐改变我们的生活。 未来, 我们会看到更多的创新和突破。

图 12-4

步骤分析:

单选题

1. 滑动法(Sliding Window)常用于在序列数据中提取子序列或进行特征提取。假设你有一个长度为 10 的序列, [1,2,3,4,5,6,7,8,9,10], 并使用大小为 4 的

滑动窗口对该序列进行处理。每次窗口向右滑动一个元素，且窗口内的元素不会重复。下列哪一项是滑窗法处理后的结果？

- A. [1, 2, 3, 4], [2, 3, 4, 5], [3, 4, 5, 6], [4, 5, 6, 7], [5, 6, 7, 8], [6, 7, 8, 9], [7, 8, 9, 10]
- B. [1, 2, 3, 4], [2, 3, 4, 5], [3, 4, 5, 6], [45, 6, 7], [5, 6, 7, 8], [6, 7, 8, 9], [7, 8, 9, 10], [8, 9, 10]
- C. [1, 2, 3], [2, 3, 4], [3, 4, 5], [4, 5, 6], [5, 6, 7], [6, 7, 8], [7, 8, 9], [8, 9, 10]
- D. [1, 2, 3, 4], [1, 2, 3, 4], [1, 2, 3, 4], [1, 2, 3, 4], [1, 2, 3, 4], [1, 2, 3, 4], [1, 2, 3, 4]

正确答案：A

解析：滑窗法中，窗口的大小为4，因此每次窗口都会包含4个连续的元素，且窗口右移1个元素。选项A正确地列出了每个窗口包含的4个元素，并且窗口向右滑动一个元素，直到遍历完整个序列。

（四）、页面解析与滑窗法结合

1. 实训目标

本任务旨在帮助学生理解如何在PDF文档解析过程中提取页头、根据文本大小与格式划分文本块，并结合滑窗法进行数据清洗和分段。通过本任务，学生将掌握以下技能：

- 如何从PDF文档中提取特定格式的内容（如页头和正文）；
- 如何使用滑窗法结合文本的大小信息对正文进行分段处理；

2. 主要步骤

• 提取页头：首先从PDF每一页中提取出页头（一级标题），通过查找页中的特定关键词如“目录”或根据行的垂直位置来确定标题。

• 解析正文：逐页读取文本，按顺序遍历每个词，通过文本的大小、内容的格式来判断是否需要分段。特别地，根据字体大小来判断哪些内容属于同一段落。

• 应用滑窗法：将提取到的文本通过滑窗法进行分段，确保每个段落的长度符合预设的最大字数限制。

• 可视化：最后通过可视化方法，展示页面解析和滑窗法结合后的效果，帮助学生理解分段后的数据结构。

3. 函数与代码实现

- `GetHeader` 函数

功能：提取页面的页头。页头的定义是位于页面顶部的文本，通常为一级标题。通过读取每一页中的文本，判断文本的位置与内容来识别标题。

步骤：

使用 `extract_words` 方法获取页面上的所有单词；

排除包含“目录”或“.....”的文本（这些不是标题）；

根据文本在页面中的垂直位置（`top` 值）判断是否为一级标题。

- `ParseBlock` 函数

功能：解析每一页的文本内容，并将其分成段落。文本段落的划分基于字体大小、文本内容和特定的关键词（如“警告！”、“注意！”等）。使用滑窗法确保每段文本长度符合最大字数限制。

步骤：

从每一页提取文本，并判断是否为标题。如果是标题则跳过该页；

遍历页面中的每个词，检查其字体大小（`size`），根据相同大小的连续文本归为同一段落；

根据滑窗法将长文本分段，每段不超过 `max_seq` 的字数限制。

使用 `Datafilter` 函数对每个段落进行处理，过滤和清洗数据。

4. 结果说明

- 通过该函数，学生能够提取 PDF 中的一级标题，并使用滑窗法对正文进行分段。每个段落的长度将不超过指定的 `max_seq` 字数限制，确保分段后的文本适合进一步处理（如输入到机器学习模型中）。

- 可视化函数将展示页头提取和滑窗法分段的效果，帮助学生理解如何通过格式化信息（如字体大小）来进行段落划分。

- 最终结果是一个分段后的文本数据集，每个段落的最大字数由 `max_seq` 控制，便于后续的分析和处理。

5. 编程实现

【底码与参考答案中的代码仅供参考，RAG 技术体验可在产业案例-AI 模型体验-任务四中进行】

```

import pdfplumber
from PyPDF2 import PdfReader
import json
import matplotlib.pyplot as plt
from expand import FstWeb
def visualize(step_name, content_preview=None):
    """ 可视化每一步的处理效果 """
    print(f"Step: {step_name}")
    if content_preview:
        print("Content preview:")
        for text in content_preview[:5]: # Preview first 5 it
            print(text)
    print(f>Data size after {step_name}: {len(data)}\n")
def SlidingWindow(sentences, kernel=512, stride=1):
    """ 滑动法分段提取 """
    sz = len(sentences)
    cur = ""
    fast = 0
    slow = 0
    while fast < len(sentences):
        sentence = sentences[fast]
        if len(cur + sentence) > kernel and (cur + sentence) r
            data.append(cur + sentence + ". ")
            cur = cur[len(sentences[slow] + ". "):]
            slow = slow + 1
            cur = cur + sentence + ". "
            fast = fast + 1
    visualize("Sliding Window", sentences[:5]) # 可视化前5个句
def Datafilter(line, header, pageid, max_seq=1024):
    """ 数据清洗与过滤 """
    sz = len(line)
    if sz < 6:
        return
    if sz > max_seq:
        if "■" in line:
            sentences = line.split("■")
        elif "·" in line:
            sentences = line.split("·")
        elif "\t" in line:
            sentences = line.split("\t")
        else:
            sentences = line.split(". ")
        for subsentence in sentences:
            subsentence = subsentence.replace("\n", "")
            if len(subsentence) < max_seq and len(subsentence)
                subsentence = subsentence.replace(", ", "").rep
                if subsentence not in data:
                    data.append(subsentence)
    else:
        line = line.replace("\n", "").replace(", ", "").replace
        if line not in data:
            data.append(line)
    visualize("Data Cleaning & Filtering") # 可视化数据过滤结果

```

```

def GetHeader(page):
    """ 提取页头即一级标题 """
    try:
        lines = page.extract_words()[::-1]
    except:
        return None
    if len(lines) > 0:
        for line in lines:
            if "目录" in line["text"] or "....." in line[
                return None
            if 17 < line["top"] < 20:
                return line["text"]
        return lines[0]["text"]
    return None
def ParseBlock(pdf_path, max_seq=1024):
    """ 页面解析与滑动法结合 """
    with pdfplumber.open(pdf_path) as pdf:
        for i, p in enumerate(pdf.pages):
            header = GetHeader(p)
            if header is None:
                continue
            texts = p.extract_words(use_text_flow=True, extra
            squence = ""
            lastsize = 0
            for idx, line in enumerate(texts):
                if idx < 1:
                    continue
                if idx == 1:
                    if line["text"].isdigit():
                        continue
                    cursize = line["size"]
                    text = line["text"]
                    if text == "□" or text == "·":
                        continue
                    elif text in ["警告!", "注意!", "说明!"]:
                        if len(squence) > 0:
                            Datafilter(squence, header, i, max_seq
                            squence = ""
                        elif format(lastsize, ".5f") == format(cursize
                            squence = squence + text if len(squence) >
                        else:
                            lastsize = cursize
                            if len(squence) < 15 and len(squence) > 0:
                                squence = squence + text
                            else:
                                if len(squence) > 0:
                                    Datafilter(squence, header, i, max
                                    squence = text
                                if len(squence) > 0:
                                    Datafilter(squence, header, i, max_seq
                                visualize("Page Parsing & Sliding Window") # 可视化页面解
                                pdf_path = await FstWeb().download_public('train_a.pdf')
                                data = []
                                # 页面解析与滑动法结合
                                ParseBlock(pdf_path, max_seq=1024)
                                ParseBlock(pdf_path, max_seq=512)

```

图 12-5

步骤分析:

单选题（单选题 5.26 分）

1. 在 ParseBlock 函数中，如何决定一个文本行是否属于同一段落？

A. 在当前行的文本大小与上一行相同且文本内容不为空时，当前行会被认为是与上一段落相关联，继续加入该段落。

B. 只有当前行的 size 与上一行的 size 不一致时，才会开始新的段落。

C. 在当前行包含特定的符号（如“□”或“·”）时，无论字体大小如何，都会结束当前段落并开始新的段落。

D. 如果当前行的文本内容为“警告！”“注意！”或“说明！”，则会自动开始新的段落，而不依赖字体大小。

正确答案：A

（五）、数据处理与保存

1. 实训目标

本实训旨在帮助学生熟练掌握如何从 PDF 文件中提取文本，并通过数据清洗、滑窗法分段提取、规则解析等步骤进行文本处理的全部流程。学生将学会如何：

- 提取 PDF 文本内容；
- 清洗和过滤无关数据；
- 利用滑窗法进行长文本分段处理；
- 提取和处理页面标题和内容；
- 最终将处理后的数据保存为 JSON 格式的输出文件。

通过本实训，学生将深刻理解 PDF 文本提取、数据清洗、自然语言处理和文本保存的基础流程，掌握如何使用 Python 中的相关库和技术实现这些任务。

2. 主要步骤

- 文本提取：从 PDF 文件中提取每页的文本内容。
- 数据清洗与过滤：清理无用数据，剔除不符合要求的文本，保留有效的内容。
- 滑窗法分段提取：通过滑窗技术将长文本按预设的大小分段处理，以适应后续的分析 and 处理。
- 页面解析与滑窗法结合：结合滑窗法与页面内容的解析，提取页面中的有效信息。

- 解析所有页面并应用滑窗法：遍历所有页面，提取文本内容并应用滑窗法分段。

- 数据保存与输出：将所有处理后的数据保存为 JSON 格式，以便后续使用。

3. 函数与代码实现

- `extract_pdf_text(pdf_path)`

功能：从指定的 PDF 文件中提取所有页面的文本内容。

- `visualize(step_name, content_preview=None)`

功能：可视化处理步骤的执行情况，并显示当前处理的文本内容预览。

- `SlidingWindow(sentences, kernel=512, stride=1)`

功能：应用滑窗法将长文本按指定窗口大小分段，并避免重复内容。

- `Datafilter(line, header, pageid, max_seq=1024)`

功能：清洗文本数据，剔除不符合条件的内容，保留有效的句子。

- `ParseBlock(pdf_path, max_seq=1024)`

功能：结合滑窗法进行页面解析，提取有效文本内容。

- `save_output(output_file="地址")`

功能：将处理后的文本数据保存到 JSONL 格式的文件中。

4. 结果说明

数据保存：所有清洗和处理过的数据将保存为 JSONL 格式的文件，每行包含一个 ID 和对应的文本内容，便于后续使用或分析。

5. 编程实现

【底码与参考答案中的代码仅供参考，RAG 技术体验可在产业案例-AI 模型体验-任务四中进行】

```

[4]:
import pdfplumber
from PyPDF2 import PdfReader
from expand import Fstlib
import json
import matplotlib.pyplot as plt
# 提取PDF文本并统计每页的字数
def extract_pdf_text(pdf_path):
    """ 提取PDF文本 """
    all_content = []
    with pdfplumber.open(pdf_path) as pdf:
        for page in pdf.pages:
            text = page.extract_text()
            if text:
                all_content.append(text)
    visualize("Text Extraction", all_content) # 可视化提取的文本
    return all_content
def visualize(step_name, content_preview=None):
    """ 可视化每一步的处理效果 """
    print(f"Step: {step_name}")
    if content_preview:
        print("Content preview")
        for text in content_preview[:5]: # Preview first 5 items
            print(text)
    print(f"Data size after {step_name}: {len(data)}\n")
def SlidingWindow(sentences, kernel=512, stride=1):
    """ 滑动窗口分段提取 """
    sz = len(sentences)
    cur = ""
    fast = 0
    slow = 0
    while fast < len(sentences):
        sentence = sentences[fast]
        if len(cur + sentence) > kernel and (cur + sentence) not in data:
            data.append(cur + sentence + "\n")
            cur = cur[len(sentences[slow]):]
            slow = slow + 1
            cur = cur + sentence + "\n"
            fast = fast + 1
    visualize("Sliding Window", sentences[:5]) # 可视化前5个句子的效果
def Datafilter(line, header, pageid, max_seq=1024):
    """ 数据清洗与过滤 """
    sz = len(line)
    if sz < 6:
        return
    if sz > max_seq:
        if "■" in line:
            sentences = line.split("■")
        elif "*" in line:
            sentences = line.split("*")
        elif "\t" in line:
            sentences = line.split("\t")
        else:
            sentences = line.split("\n")
        for subsentence in sentences:
            subsentence = subsentence.replace("\n", "")
            if len(subsentence) < max_seq and len(subsentence) > 5:
                subsentence = subsentence.replace(",", "").replace("\n", "").replace("\t", "")
                if subsentence not in data:
                    data.append(subsentence)
    else:
        line = line.replace("\n", "").replace(",", "").replace("\t", "")
    if line not in data:
        data.append(line)
    visualize("Data Cleaning & Filtering") # 可视化数据过滤结果

```

```

def GetHeader(page):
    """ 提取页头第一级标题 """
    try:
        lines = page.extract_words()[1:]
    except:
        return None
    if len(lines) > 0:
        for line in lines:
            if "目录" in line["text"] or "....." in line["text"]:
                return None
            if 17 < line["top"] < 20:
                return line["text"]
        return lines[0]["text"]
    return None
def ParseBlock(pdf_path, max_seq=1024):
    """ 页面解析与滑动窗口 """
    with pdfplumber.open(pdf_path) as pdf:
        for i, p in enumerate(pdf.pages):
            header = GetHeader(p)
            if header is None:
                continue
            texts = p.extract_words(use_text_flow=True, extra_attrs=["size"])
            squence = ""
            lastsize = 0
            for idx, line in enumerate(texts):
                if idx < 1:
                    continue
                if idx == 1:
                    if line["text"].isdigit():
                        continue
                    cursize = line["size"]
                    text = line["text"]
                    if text == "□" or text == "*":
                        continue
                    elif text in ["警告!", "注意!", "说明!"]:
                        if len(squence) > 0:
                            Datafilter(squence, header, i, max_seq)
                            squence = ""
                    elif format(lastsize, ".5f") == format(cursize, ".5f"):
                        squence = squence + text if len(squence) > 0 else text
                    else:
                        lastsize = cursize
                        if len(squence) < 15 and len(squence) > 0:
                            squence = squence + text
                        else:
                            if len(squence) > 0:
                                Datafilter(squence, header, i, max_seq)
                                squence = text
            if len(squence) > 0:
                Datafilter(squence, header, i, max_seq)
            visualize("Page Parsing & Sliding Window") # 可视化页面解析和滑动窗口的结合
def ParseOnePageWithRule(pdf_path, max_seq=512, min_len=6):
    """ 规则解析一页PDF """
    for idx, page in enumerate(PdfReader(pdf_path).pages):
        page_content = ""
        text = page.extract_text()
        words = text.split("\n")
        for word in words:
            text = word.strip().strip("\n")
            if "....." in text or "目录" in text:
                continue
            if len(text) < 1:
                continue
            if text.isdigit():
                continue
            page_content = page_content + text
            if len(page_content) < min_len:
                continue
            if len(page_content) < max_seq:
                if page_content not in data:
                    data.append(page_content)
        else:
            sentences = page_content.split("\n")
            cur = ""
            for sentence in sentences:
                if len(cur + sentence) > max_seq and (cur + sentence) not in data:
                    data.append(cur + sentence)
                cur = sentence
            else:
                cur = cur + sentence
            visualize("Parse One Page With Rule") # 可视化规则解析的结果

```

```

def ParseAllPage(pdf_path, max_seq=512, min_len=6):
    """ 解析所有页面并应用滑动法 """
    all_content = ""
    for idx, page in enumerate(PdfReader(pdf_path).pages):
        page_content = ""
        text = page.extract_text()
        words = text.split("\n")
        for word in words:
            text = word.strip().strip("\n")
            if "....." in text or "目录" in text:
                continue
            if len(text) < 1:
                continue
            if text.isdigit():
                continue
            page_content = page_content + text
        if len(page_content) < min_len:
            continue
        all_content = all_content + page_content
    sentences = all_content.split("。")
    SlidingWindow(sentences, kernel=max_seq)

    visualize("Parse All Pages") # 可视化所有页面解析后的结果

def save_output(output_file="output.json1"):
    """ 保存数据到文件 """
    with open(output_file, "w", encoding='utf-8') as out:
        for index, line in enumerate(data):
            line = line.strip("\n")
            json_object = json.dumps({"id": index + 1, "text": line}, ensure_ascii=False)
            out.write(json_object)
            out.write("\n")

    visualize("Data Saving", data[:5]) # 可视化前5条数据的保存效果

# 主程序入口
if __name__ == "__main__":
    pdf_path = await FstWeb().download_public('train_a.pdf')
    data = [] # 存储所有提取的数据

    # 1. 提取PDF文本
    pdf_text = extract_pdf_text(pdf_path)

    # 2. 数据清洗与过滤
    Datafilter("sample line", "header", 1, max_seq=1024)

    # 3. 滑动法分块提取
    SlidingWindow(pdf_text[0].split("。"), kernel=512)

    # 4. 页面解析与滑动法结合
    ParseBlock(pdf_path, max_seq=1024)
    ParseBlock(pdf_path, max_seq=512)

    # 5. 解析所有页面并应用滑动法
    ParseAllPage(pdf_path, max_seq=256)
    ParseAllPage(pdf_path, max_seq=512)
    ParseOnePageWithRule(pdf_path, max_seq = 256)
    ParseOnePageWithRule(pdf_path, max_seq = 512)

    # 6. 数据保存与输出
    save_output(output_file="processed_data4.json1")

    print("Data processing complete.")

```

```

Step: Text Extraction
Content preview:
欢迎
感谢您选择了具有优良安全性、舒适性、动力性和经济性的汽车。首次使用前请
仔细、完整地阅读本手册内容，将有助于您更好地了解和使用车辆。
本手册中的所有资料均为出版时的最新资料，但本公司将对产品进行不断的改进和优化，您所购的车
辆可能与本手册中的描述有所不同，请以实际
接收的车辆为准。
如您有任何问题，或需要预约服务，请拨打电话 联系我们。您也可以开车前往中心。在抵达之前，请
您注意驾车
安全。
© 汽车销售有限公司
目录
前言 设置尾门开启角度

```

图 12-6

任务三 自然语言处理的应用

一、理论知识实训

1. 医疗领域中，从电子病历提取“每日服用阿司匹林 100mg”属于 NLP 的哪种应用？

- A: 医学文献分类
- B: 临床记录分析
- C: 症状匹配
- D: 基因序列预测

标准答案：B

2. 金融领域合同智能审查主要识别：

- A: 语法错误
- B: 格式排版问题
- C: 风险条款
- D: 签字人身份

标准答案：C

3. 教育领域的 e-rater 系统主要用于：

- A: 监考防作弊
- B: 课件制作
- C: 课堂考勤
- D: 作文自动评分

标准答案：D

4. 电商评论分析“电池续航差”高频出现，有助于：

- A: 改进产品设计
- B: 降低商品价格
- C: 增加广告投放
- D: 缩短物流时间

标准答案：A

5. Zoom 实时字幕功能主要服务：

- A: 外语学习者

- B: 听障人士
- C: 视频编辑人员
- D: 网络管理员

标准答案: B

6. Google Lens 的 AR 翻译属于 NLP 与哪种技术的结合?

- A: 区块链
- B: 量子计算
- C: 增强现实
- D: 虚拟货币

标准答案: C

7. 车载系统"调低空调并打开导航"需要处理:

- A: 纯文本指令
- B: 图像识别
- C: 生物特征
- D: 多模态输入

标准答案: D

8. Sudowrite 剧本生成系统属于 NLP 的:

- A: 内容生成应用
- B: 语法检查工具
- C: 语音转写系统
- D: 情感分析平台

标准答案: A

9. 农户描述"叶片褐色斑点"通过 NLP 可识别为:

- A: 土壤肥力
- B: 天气变化
- C: 稻瘟病
- D: 虫害天敌

标准答案: C

10. 客服录音转质检主要检查:

- A: 通话时长
- B: 合规性遗漏
- C: 方言口音
- D: 背景噪音

标准答案: B

二、自然语言处理的应用

(一)、问答知识库构建与检索

1. 实训目标

本次实训任务的目标是通过构建一个基于知识库的汽车问答检索系统,帮助学生掌握以下技能:

- 使用 SentenceTransformer 进行文本嵌入并实现多语言文本向量化。
- 构建和存储基于文本向量的知识库。
- 根据输入文本计算相似度,检索出最相似的文本。
- 掌握如何使用 PyTorch 进行 GPU 加速处理,并将模型和计算迁移到 GPU。

通过本任务,学生将学习到如何使用 SentenceTransformer 和 PyTorch 的功能,构建一个简单的知识库,并能够根据输入查询文本与知识库中的条目进行相似度计算与检索。

2. 主要步骤

- 设备选择:检查是否可以使用 GPU 进行计算(如果有可用的 CUDA 设备,则使用 GPU,否则使用 CPU)。

- 加载数据与知识库:

若知识库文件已经存在,则加载知识库。

若知识库文件不存在,则从数据文件中加载数据,并为每一条数据生成文本嵌入向量,构建知识库并保存。

- 文本相似度计算与检索:

定义一个检索函数 `retrieve_similar_sentences`,根据输入文本生成嵌入向量,计算与知识库中所有条目的相似度。

根据计算出的相似度,返回与输入文本最相似的条目。

展示结果:将检索到的文本输出。

3. 结果说明

输出结果：对于给定的输入文本（如：“如何进入车辆功能界面？”），程序将会查找与其相似的知识库中的文本，并输出最相似的一个（或多个）条目。

成功加载与保存知识库：如果程序首次运行并成功创建了知识库，或者加载了已存在的知识库，控制台将显示：

知识库已加载

或者 知识库已保存到 ...

4. 编程实现

【底码与参考答案中的代码仅供参考，RAG 技术体验可在产业案例-AI 模型体验-任务四中进行】

```

import json
import os
import torch
from sentence_transformers import SentenceTransformer, util
from modelscope import AutoModelForCausalLM, AutoTokenizer

# 检查是否有可用的GPU
device = 'cuda' if torch.cuda.is_available() else 'cpu'
print(f"Using device: {device}")

# 数据集路径
data_path = r"E:\First-work\人工智能平台\newoutput.jsonl"
# 知识库保存路径
knowledge_base_path = r'E:\First-work\人工智能平台\knowledge_base.jsonl'

# 使用 SentenceTransformer 模型进行文本嵌入
model = SentenceTransformer('paraphrase-multilingual-MiniLM-L12-v2')

# 加载知识库
if os.path.exists(knowledge_base_path):
    # 如果存在已保存的知识库，直接加载
    knowledge_base = torch.load(knowledge_base_path, map_location=device)
    print("知识库已加载")
else:
    # 如果知识库文件不存在，则构建知识库
    print("构建知识库中...")
    data = []
    with open(data_path, 'r', encoding='utf-8') as f:
        for line in f:
            data.append(json.loads(line))

    knowledge_base = []
    for entry in data:
        ja_text = entry["text"]
        # 向量化target字段，并迁移到 GPU
        ja_embedding = model.encode(ja_text, convert_to_tensor=True)
        knowledge_base.append((ja_embedding, entry))

# 保存知识库到本地
torch.save(knowledge_base, knowledge_base_path)
print(f"知识库已保存到 {knowledge_base_path}")

def retrieve_similar_sentences(input_text, top_k=1, similarity_threshold=0.8):
    # 向量化文本，并迁移到 GPU
    input_embedding = model.encode(input_text, convert_to_tensor=True)

    # 计算相似度
    similarities = []
    for ja_embedding, entry in knowledge_base:
        similarity = util.pytorch_cos_sim(input_embedding, ja_embedding)
        if similarity >= similarity_threshold:
            similarities.append((similarity, entry))

    # 按相似度排序，取出前k条
    similarities = sorted(similarities, key=lambda x: x[0], reverse=True)

    # 输出相似的条目中的target和targett字段作为参考
    references = [{"text": entry["text"]} for _, entry in similarities[:top_k]]
    return references

input_text = '如何进入车辆功能界面?'
similar_sentences = retrieve_similar_sentences(input_text, top_k=1)
for ref in similar_sentences:
    print(f"-----{ref['text']}-----")

```

图 12-7

步骤分析：

1. 以下哪一项最能描述 SentenceTransformer 模型的主要作用，以及其如何影响知识库构建和相似度检索过程？

A. SentenceTransformer 使用 BERT 的词嵌入表示来计算每个单词的词向量，并基于词向量进行相似度匹配，用于优化单词级别的翻译精度。

B. SentenceTransformer 利用预训练的多语言模型生成句子级别的嵌入表示，通过比较句子向量之间的余弦相似度来检索与输入句子相似的文本。该方法避免了基于传统 TF-IDF 的检索方式。

C. SentenceTransformer 通过 LSTM 模型对输入文本进行逐字处理，生成每个字符的嵌入向量并计算序列相似度，用于构建基于字符级别的知识库。

D. SentenceTransformer 采用 Transformer 架构对每个输入句子的每个词进行单独编码，生成词向量后通过 K-means 聚类算法构建知识库，并使用聚类结果进行相似度检索。

正确答案：B

（二）、测试问题生成与综合应用

1. 实训目标

本次实训的目标是通过处理和检索数据，为后续构建一个基于知识库的问答系统做准备。具体来说，学生将学习如何：

从给定的 JSONL 文件中读取数据，并进行数据的随机选择和格式化；

构建并加载一个基于 SentenceTransformer 模型的文本嵌入知识库；

实现检索相似句子的功能，并结合检索到的信息生成回答；

通过与用户交互，构建一个基于知识库检索的提示词工程。

2. 主要步骤

数据加载与格式化：

从指定路径加载一个 JSONL 格式的数据文件。

随机选择 10 条数据，并提取 question 和 answer_1 到 answer_4 字段，进行格式化处理。

构建知识库：

使用 SentenceTransformer 模型对数据集中的文本进行嵌入，生成文本的向量表示。

如果存在已保存的知识库，则加载之；如果不存在，则新建并保存该知识库。

相似句子检索：

实现一个检索函数，通过计算输入文本与知识库中文本的相似度，返回最相似的句子。

基于相似度对句子进行排序，并根据设定的相似度阈值进行筛选。

提示词工程：

通过用户输入的问题进行检索，输出最相关的参考信息。

合并检索到的内容和用户的输入问题，构建用于大语言模型的生成提示(prompt)。

3. 函数代码与实现

- 数据加载和格式化：

通过 `json.load()` 函数读取 JSONL 文件内容并解析为 Python 对象。使用 `random.sample()` 从数据中随机选择 10 个问题，并提取其中的 `question` 与 `answer_1` 到 `answer_4`。

- 知识库构建：

使用 SentenceTransformer 模型将文本转换为向量 (embedding)，存储在 GPU 或 CPU 上。如果知识库文件已存在，直接加载，否则重新构建并保存该知识库。

- 相似句子检索：

`retrieve_similar_sentences()` 函数接收用户输入的文本，利用 SentenceTransformer 模型将文本转换为向量，并与知识库中的所有文本进行相似度计算。

按相似度对结果进行排序，返回最相似的条目。

- 提示词工程：

用户输入问题后，程序首先通过 `retrieve_similar_sentences()` 函数检索相关信息。然后，构建一个包含用户问题和参考信息的 `prompt`，并将在后续输入

到大语言模型中。

4. 结果说明

- 数据格式化结果:

该步骤将展示每个问题 (question) 及其对应的 4 个答案 (answer_1 到 answer_4)。例如:

- 知识库构建结果:

如果程序成功加载了现有的知识库, 则输出“知识库已加载”。

如果需要构建新的知识库, 则会在控制台输出“构建知识库中...”并保存到指定路径。

- 检索结果:

用户输入问题后, 程序会显示与之相关的参考信息。

5. 编程实现

【底码与参考答案中的代码仅供参考, RAG 技术体验可在产业案例-AI 模型体验-任务四中进行】

```

import json

# 定义输入和输出文件路径
input_file = r"E:\First-work\QA-system\Tianchi-LLM-QA-main\data

with open(input_file, 'r', encoding='utf-8') as f:
    try:
        data = json.load(f) # 读取整个文件并解析为一个 JSON 对象
        # for item in data:
        #     print(item) # 打印每个元素
        # print(data[0])
    except json.JSONDecodeError as e:
        print(f"Error decoding JSON: {e}")
import random

# 从 data 随机选择 10 个元素, 并仅保留 'question' 和 'answer_1'
selected_questions = random.sample(data, 10)

# 构造新的列表, 只保留 'question' 和 'answer_1' 到 'answer_4'
formatted_questions = [
    {
        'question': item['question'],
        'answer_1': item['answer_1'],
        'answer_2': item['answer_2'],
        'answer_3': item['answer_3'],
        'answer_4': item['answer_4']
    }
    for item in selected_questions
]

# 输出结果
for i, question in enumerate(formatted_questions, 1):
    print(f"Question {i}: {question['question']}")
    # print(f"Answer 1: {question['answer_1']}")
    # print(f"Answer 2: {question['answer_2']}")
    # print(f"Answer 3: {question['answer_3']}")

```

图 12-8

步骤分析:

1. 在上述代码中, retrieve_similar_sentences() 函数计算输入文本与知识库中的句子相似度时, 以下哪一项描述是正确的?

A. 输入文本的向量化是通过 SentenceTransformer 模型进行的，该模型生成的是文本的嵌入向量(embedding)。只有当两个文本的向量距离小于设定的阈值时，它们才会被认为是相似的，并进行进一步的检索。

B. 相似度计算使用的是余弦相似度，计算公式通过 `util.pytorch cos sim`(函数实现)。该函数返回的是两个向量之间的相似度得分，而不是它们的距离。相似度得分高于设定的阈值时，才会被认为是相关的。

C. 如果输入文本的相似度得分小于设定的阈值，函数会自动返回空列表，而不是继续返回与阈值相近的句子。

D. 在相似句子检索时，知识库中的每个文本都必须经过重新计算嵌入向量，而不是使用已经存储的嵌入向量进行比较。因此，模型每次调用时都需要重新训练。

正确答案：B

项目十三：人工智能开放平台应用

任务一 大模型的使用方法

一、理论知识实训

1. 大模型的核心技术架构通常是：

- A: 卷积神经网络
- B: 循环神经网络
- C: Transformer
- D: 决策树

标准答案：C

2. 以下哪项不属于大模型的典型能力？

- A: 文本生成
- B: 代码补全
- C: 多模态理解
- D: 操控机器人

标准答案：D

3. 提示工程 (Prompt Engineering) 的主要目的是：

- A: 减少模型参数
- B: 引导模型输出
- C: 加速模型训练
- D: 降低计算成本

标准答案：B

4. 设置 temperature=0.2 时，模型输出最可能的特点是：

- A: 确定性高
- B: 随机性强
- C: 长度更长
- D: 包含更多 emoji

标准答案：A

5. 多模态大模型区别于语言大模型的关键能力是：

- A: 参数规模更大
- B: 训练速度更快
- C: 多模态数据处理
- D: 仅支持英文

标准答案: C

6. 使用 Chain-of-Thought 提示技术的场景是:

- A: 图像风格迁移
- B: 数据清洗
- C: 语音降噪
- D: 数学计算

标准答案: D

7. 调用大模型 API 时, max_tokens=50 表示:

- A: 限制输入长度
- B: 限制输出长度
- C: 设置请求超时
- D: 控制 API 费用

标准答案: B

8. 以下哪种提示词设计最符合明确指令原则?

- A: "写首诗"
- B: "总结文章"
- C: "用 3 点向老人解释 5G"
- D: "随便聊聊"

标准答案: C

9. 本地部署轻量化大模型的工具是:

- A: Llama.cpp
- B: TensorFlow Lite
- C: PyTorch Mobile
- D: Kubernetes

标准答案: A

10. 大模型产生"幻觉" (Hallucination)是指:

A: 计算错误

B: 响应延迟

C: 内存溢出

D: 虚构事实

标准答案: D

二、大模型部署与调用方法

在平台中内置一说明文档,如下图所示,请阅读完毕后,完成后续流程图绘制实训任务。

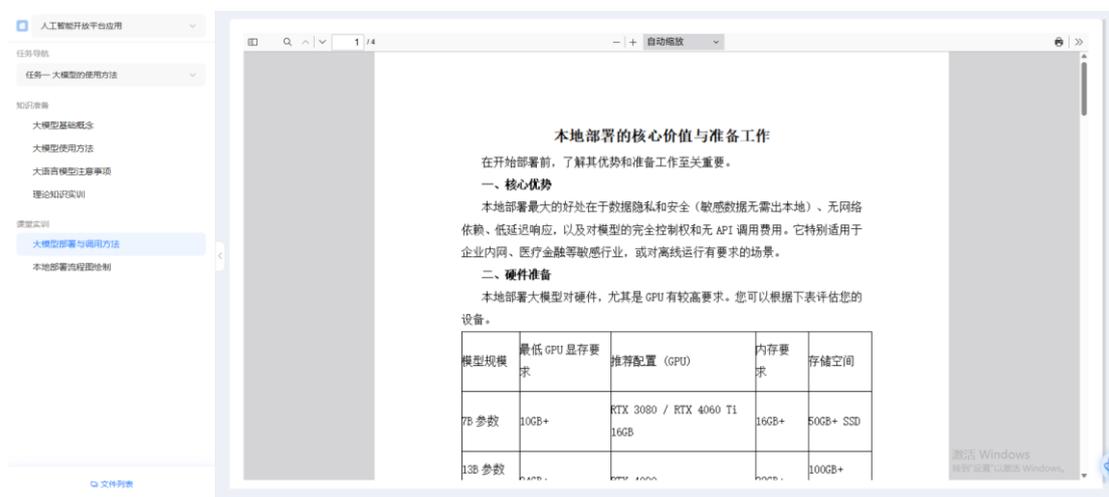


图 13-1

三、本地部署流程图绘制

请在学习了大语言模型本地部署的方法后,完成使用 Ollama 方法部署的流程图绘制,并额外增加一个方框,填写使用 Ollama 运行 qwen2.5-7b 模型的命令。

参考答案如下所示:

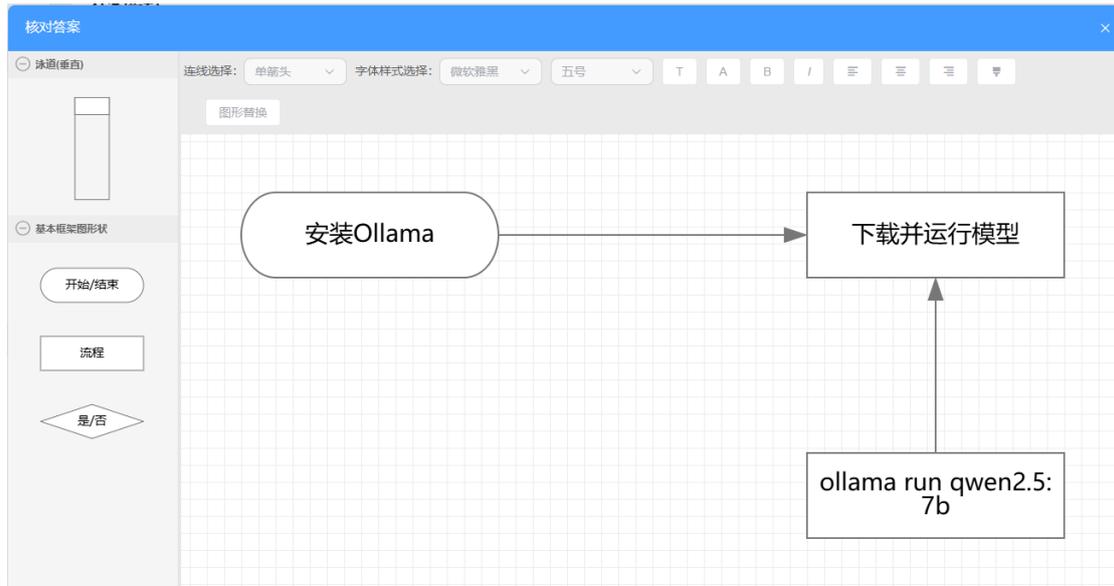


图 13-2

任务二 大语言模型优化相关的前言方法

一、知识准备-理论知识实训

1. 混合专家系统 (MoE) 的核心优势是:

- A: 参数规模最大
- B: 计算效率高
- C: 训练速度最快
- D: 无需微调

标准答案: B

2. 指令微调 (Instruction Tuning) 解决的主要问题是:

- A: 模型体积过大
- B: 数据存储不足
- C: 指令遵循能力差
- D: 硬件兼容性差

标准答案: C

3. RLHF 训练过程中, 奖励模型 (Reward Model) 的作用是:

- A: 生成对抗样本
- B: 压缩模型参数
- C: 加速推理过程
- D: 预测答案质量得分

标准答案: D

4. 模型量化(Quantization)的主要目的是:

- A: 减少资源消耗
- B: 提高模型准确率
- C: 增加数据多样性
- D: 增强对抗攻击防御

标准答案: A

5. 知识蒸馏(Knowledge Distillation)中"软标签"是指:

- A: 人工标注的类别
- B: 概率分布输出
- C: 低精度参数
- D: 加密训练数据

标准答案: B

6. 稀疏注意力(Sparse Attention)通过什么方式提升效率?

- A: 删除停用词
- B: 减少训练轮次
- C: 限制注意力范围
- D: 丢弃长文本

标准答案: C

7. 多任务联合训练(Multi-task Learning)的典型代表模型是:

- A: GPT-3
- B: BERT
- C: LLaMA
- D: T5

标准答案: D

8. 动态量化(Dynamic Quantization)相比静态量化的优势是:

- A: 处理变长输入
- B: 计算更简单
- C: 精度损失更小
- D: 无需校准数据

标准答案: A

9. 对抗训练(Adversarial Training)主要提升模型的:

- A: 生成速度
- B: 多模态能力
- C: 抗干扰能力
- D: 参数规模

标准答案: C

10. 递归结构(Recursive Models)特别适合处理:

- A: 图像分类
- B: 长文本生成
- C: 实时语音识别
- D: 结构化数据预测

标准答案: B

二、提示词工程与检索增强生成技术应用

在平台中内置一说明文档,如下图所示,请阅读完毕后,完成后续实训任务。

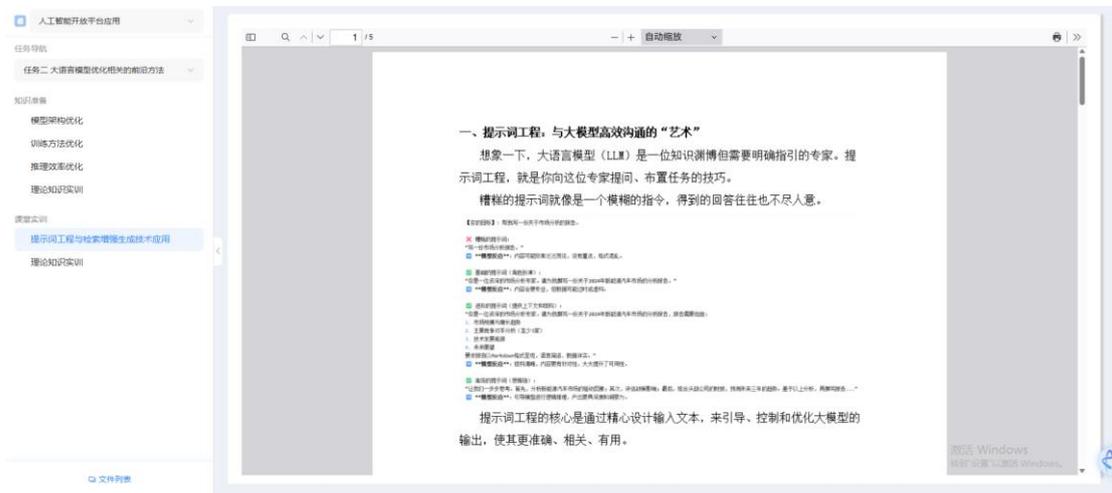


图 13-3

三、课堂实训-理论知识实训

1. 提示词工程的主要目标是什么？

- A. 训练一个全新的大语言模型
- B. 优化计算资源的消耗，降低模型运行成本
- C. 通过精心设计输入文本来引导和控制大模型的输出
- D. 加速大模型生成文本的速度

标准答案：C

2. 检索增强生成（RAG）技术主要是为了解决大模型的哪个核心问题？

- A. 计算速度过慢
- B. 训练成本过高
- C. 缺乏最新或特定领域知识，可能导致“幻觉”
- D. 无法进行多轮对话

标准答案：C

3. 在 RAG 的工作流程中，“检索”这一步通常是如何进行的？

- A. 将用户问题直接发送给大模型，让其自行搜索答案
- B. 使用关键词匹配或向量检索技术，从外部知识库中查找相关文档片段
- C. 要求用户自己提供相关的参考资料
- D. 在互联网上进行实时搜索并汇总结果

标准答案：B

4. 以下哪一项是结合了提示词工程和 RAG 的“增强版提示词”的典型结构？

- A. 直接使用用户的原始问题
- B. 系统角色指令 + 检索到的背景信息 + 用户问题+ 回答格式要求
- C. 仅包含检索到的背景信息
- D. 一个充满示例的复杂表格

标准答案：B

5. 在一个企业知识库问答系统中，用户提问：“我们公司最新的差旅政策有什么变化？”系统最有可能基于什么来生成答案？

- A. 大模型内部训练数据中关于差旅政策的通用知识

- B. 检索到的该公司内部最新的《差旅政策规定》PDF 文档内容
- C. 互联网上其他公司的最新差旅政策
- D. 随机生成一段看似合理的政策描述

标准答案：B

6. 提示词工程中，“角色扮演”（例如：“你是一位资深的法律顾问”）最主要的作用是？

- A. 让模型回答得更快
- B. 让回答更幽默风趣
- C. 约束模型的回答风格和专业领域，使其更符合预期
- D. 减少提示词的字符数量

标准答案：C

7. RAG 系统中的“外部知识库”通常不包括以下哪项内容？

- A. 公司内部的项目文档和产品手册
- B. 大语言模型本身在训练时学习到的参数和知识
- C. 从权威网站爬取的行业报告
- D. 结构化的数据库表格内容

标准答案：B

8. 如果 RAG 系统检索到的背景信息中完全不包含用户问题的答案，一个设计良好的提示词模板应该指示模型怎么做？

- A. 根据自身知识自由发挥，尽力提供一个看似合理的答案
- B. 明确告知用户“根据现有资料无法回答该问题”
- C. 引导用户换一个问题提问
- D. 保持沉默，不输出任何内容

标准答案：B

9. 在 RAG 流程中，将检索到的文档片段与用户问题组合起来，其主要目的是什么？

- A. 增加输入文本的长度，使模型更重视
- B. 为模型提供生成答案所需的事实依据和上下文
- C. 测试模型的文本理解能力

D. 加密用户的问题，保护隐私

标准答案：B

10. 提示词工程和 RAG 技术之间的关系，最恰当的比喻是？

A. 竞争对手：两者目标不同，相互排斥

B. 方向盘与 GPS 导航：提示词工程控制方向（如何答），RAG 提供路径和实时路况（依据什么答）

C. 油门与刹车：一个负责加速，一个负责减速

D. 发动机与车轮：一个是动力来源，一个是执行部件

标准答案：B

任务三 AI 开放平台应用

一、理论知识实训

1. 以下哪项是 AI 开放平台的核心能力之一？

A: 数据库管理

B: 自然语言处理（NLP）

C: 区块链技术

D: 网络安全

标准答案：B

2. 通过 AI 开放平台调用 API 实现“作文批改”功能，主要依赖哪种技术？

A: 自然语言处理（NLP）

B: 计算机视觉（CV）

C: 语音合成（TTS）

D: 增强现实（AR）

标准答案：A

3. 以下哪个平台属于垂直领域的 AI 开放平台？

A: OpenAI

B: Hugging Face

C: 腾讯觅影

D: 百度 AI 开放平台

标准答案: C

4. 与传统开发相比, 使用 AI 开放平台的主要优势是什么?

A: 数据安全性更高

B: 功能定制更灵活

C: 无需网络连接

D: 快速集成且成本低

标准答案: D

5. 工业领域中, AI 开放平台“缺陷检测”功能主要依赖哪种技术?

A: 语音识别

B: 计算机视觉 (CV)

C: 自然语言处理 (NLP)

D: 推荐算法

标准答案: B

6. 以下哪项是国产 AI 开放平台的代表?

A: Google Cloud AI

B: IBM Watson Health

C: OpenAI

D: 科大讯飞开放平台

标准答案: D

7. 在教育领域, 通过拍照识别手写数学公式并解题, 需要调用哪些 API 组合?

A: 语音合成 + 知识图谱

B: 推荐算法 + AR

C: OCR + 解题 API

D: 异常检测 + 时序预测

标准答案: C

8. 医疗领域中, “肺结节标记”功能通常由哪种技术实现?

- A: 计算机视觉 (CV)
- B: 语音技术
- C: 自然语言处理 (NLP)
- D: 推荐算法

标准答案: A

9. 以下哪个案例体现了生成式 AI 的应用?

- A: 作文批改
- B: 药物分子结构设计
- C: 商品推荐
- D: 手写公式识别

标准答案: B

10. 零售领域中, “虚拟试妆” 功能主要结合了哪两种技术?

- A: NLP + TTS
- B: 边缘计算 + 推荐算法
- C: CV + AR
- D: 语音识别 + 知识图谱

标准答案: C

二、streamlit 和本地部署网页平台

Streamlit 是一种开源的 Python 框架, 专为快速构建交互式数据应用和人工智能 (AI) 模型的展示平台而设计。通过 Streamlit, 开发者可以轻松地将机器学习模型和 AI 服务部署为用户友好的网页应用, 而无需掌握复杂的前端开发技术。其简单高效的特性, 使其成为快速搭建本地部署网页平台的理想选择。

技术特点

1. 快速开发: Streamlit 使用简单的 Python 脚本即可生成交互式网页。开发者只需专注于 Python 代码逻辑, 无需额外学习 HTML、CSS 或 JavaScript, 即可实现复杂的界面布局和交互功能。

2. 实时更新: Streamlit 支持热加载功能。当代码发生更改时, 平台可以自动刷新界面, 方便开发者快速迭代和调试。

3. 强大的交互功能: Streamlit 提供多种内置的交互组件, 例如按钮、滑块、下拉菜单、文件上传等, 用户可以通过这些组件与后台的 AI 模型交互, 实现模型输入输出的直观展示。

4. 兼容性强: Streamlit 可以无缝集成各种机器学习和深度学习框架 (如 PyTorch、TensorFlow), 并支持与外部 API 的交互。此外, 应用可以在本地运行, 也可以轻松部署到云端平台。

应用场景

1. AI 模型演示: 利用 Streamlit, 开发者可以为机器学习模型创建直观的演示页面。例如, 用户可以上传图片并实时查看分类结果, 或输入文本并获取自然语言处理模型的输出。

2. 数据可视化与分析: Streamlit 支持与主流数据分析库 (如 Pandas、Matplotlib、Plotly) 集成, 开发者可以通过交互式图表展示分析结果, 为 AI 项目的研究与推广提供直观支持。

3. 本地部署平台: 对于需要保证数据隐私的应用场景, Streamlit 可用于本地部署的网页平台。用户在局域网内访问平台, 与本地的 AI 模型交互, 无需将数据上传到云端。

示例 workflow

1. 构建交互界面: 使用 Streamlit 的简单 API 编写界面代码, 例如文本输入框 (`st.text_input`)、文件上传组件 (`st.file_uploader`) 等。

2. 加载和运行模型: 将训练好的 AI 模型加载到平台中, 使用用户的输入进行实时推理。

3. 展示结果: 利用 Streamlit 提供的可视化组件 (如表格、图像、动态图表等), 直观展示模型的输出结果。

4. 启动应用: 通过 `streamlit run app.py` 启动本地服务器, 用户即可在浏览器中访问 AI 平台。

三、可交互式 AI 开放网页平台

“扣子平台”是字节跳动推出的一个面向开发者和企业的技术平台, 旨在为第三方应用提供智能化服务和解决方案。通过扣子平台, 开发者可以接入字节跳动的人工智能能力, 包括自然语言处理、计算机视觉、语音识别、推荐算法等强

大的技术。平台提供一系列 API 和 SDK 工具，帮助开发者快速集成 AI 技术，优化用户体验并提升业务效率。扣子平台不仅支持文本、语音和图像等多模态的智能处理，还具备高效的数据分析能力，能够为企业提供个性化的推荐、广告投放、用户行为分析等功能，广泛应用于内容创作、社交互动、电商推荐等多个领域，推动数字化转型和智能化发展。

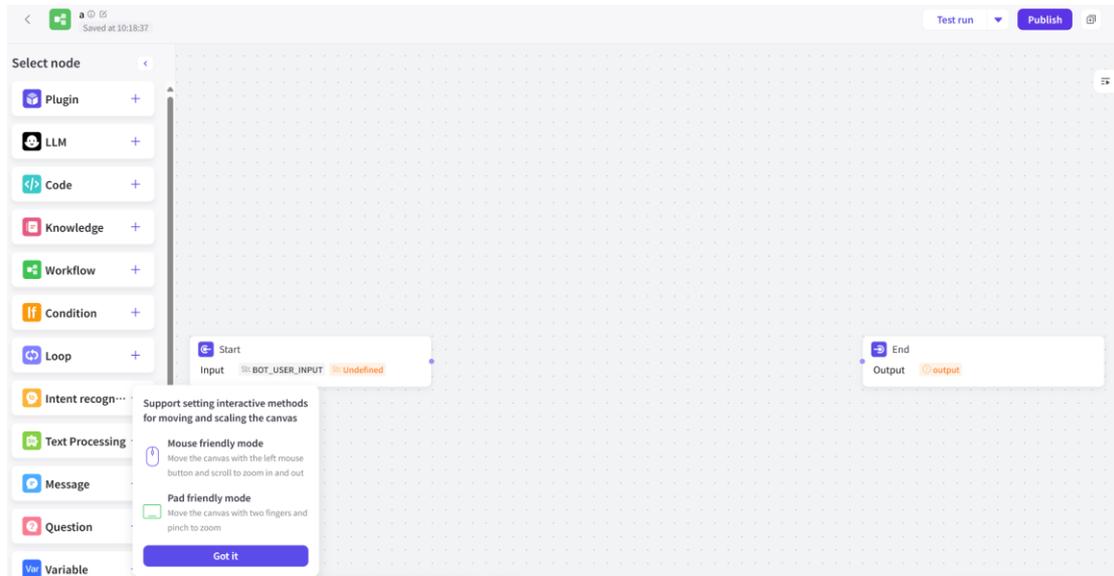


图 13-4

Agent（智能体）是一种能自主感知、学习、决策并执行任务的智能程序或系统。Agent 是基于人工智能（AI）技术的应用，它具备感知能力、推理能力和行动能力，能够根据环境的变化做出适应性决策。智能体可以在没有人工干预的情况下，根据程序预定的目标和规则，自动执行复杂的任务。智能体的核心特点包括自主性、感知能力、决策能力和学习能力。

智能体在工作流中的作用至关重要，它能够不需要人工干预的情况下，自动执行和管理工作流中的任务。工作流系统中引入 Agent 后，系统能够在复杂条件下自动决策，完成任务的分配、审批、执行等多个环节。这使得工作流变得更加智能化和自动化。具体来说，智能体在工作流中的应用包括以下几个方面：

自动化任务执行：Agent 可以根据工作流的定义自动执行任务，比如根据预设的条件审核文件、填写表格、检查数据等。

智能决策与判断：Agent 能够根据系统中的信息和数据进行智能判断，并做出决策。例如，在多重审批流程中，Agent 可以根据规则判断是否跳过某些步骤或请求额外的批准。

动态调整 workflow：智能体能够根据实时数据或外部环境变化调整 workflow 的执行路径。比如，当某个任务因外部条件发生延误时，智能体可以自动调整其他任务的顺序，确保 workflow 的整体效率。

异常处理与优化：智能体能够实时监控 workflow 执行中的问题，发现异常并及时响应。例如，在 workflow 的某个环节出现错误时，Agent 可以自动启动修复或替代操作，确保流程的连续性。

结合 workflow 和智能体 (Agent) 技术，企业和组织可以实现高度自动化的业务处理。智能体可以作为 workflow 系统中的“执行者”自动执行任务，而 workflow 则充当着任务组织、协调和调度的角色。这种自动化和智能化的结合，不仅大幅提高了任务的执行效率，还能够减少人工干预的错误，提高决策的准确性和一致性。尤其在大规模、复杂的业务流程中，workflow 与 Agent 的结合为企业提供了极大的灵活性和处理能力，适应了快速变化的商业环境。

四、使用 Coze 创建智能体

(一) 创建智能体步骤参考

平台中内置了操作步骤，请阅读完毕后，在扣子平台中自行注册账号，并参考文档内容完成实训并截图上传至后一任务中。

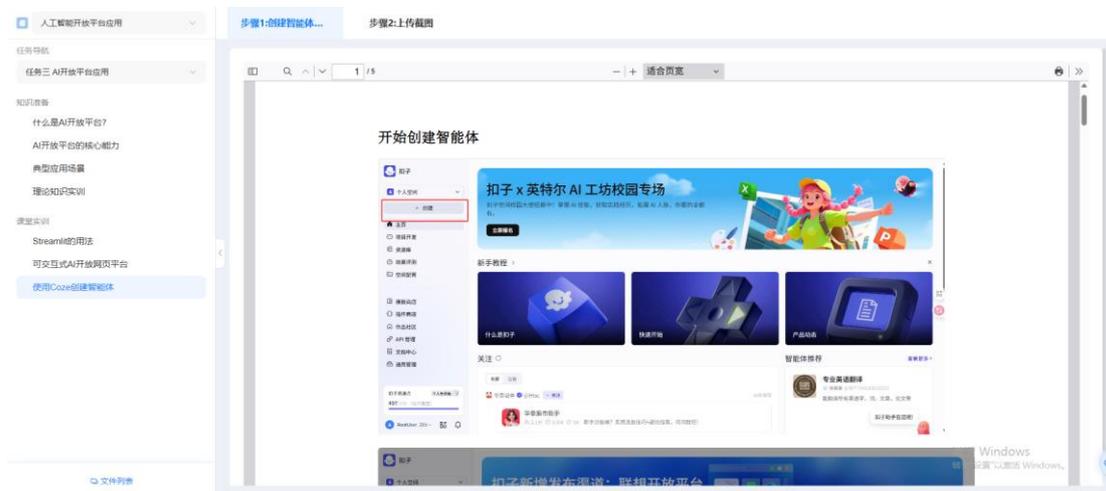


图 13-5

(二) 上传截图

请在下方上传自行创建的智能体问答截图。

下方示例仅供参考，无标准答案



财税常识问答助手

嘿，朋友，很高兴遇见你！我是专业的财税常识问答助手，能清晰准确地为你解答各类基础财税常识问题。要是你有相关疑问，尽管问我。



RootUser_2111950627

介绍一下一般纳税人增值税的概念



财税常识问答助手

运行完毕



一般纳税人是年应税销售额超规定标准的企业等，能准确核算销项税额并按规定报送税务资料。

在增值税制度里，一般纳税人应纳税额为销项税额减进项税额。销项税额是销售货物等时收取的增值税额，进项税额是购进货物等时支付的增值税额。

按《增值税会计处理规定》，一般纳税人在“应交税费”科目下设“应交增值税”明细科目，还会按不同用途设“进项税额”“销项税额”等专栏，来记录管理增值税变动。企业要按税法规定时间和



继续对话...



内容由AI生成，无法确保真实准确，仅供参考。

图 13-6

项目十四：AI 智能体创建

任务一 大模型微调详解

1. 体验 AI 智能客服

1.1 低学习率模型体验

步骤一 模型选择

(1) 根据任务描述下提供的编写电商客服的 AI 模型表格，在【模型选择与训练】下方，选择对应的内容，具体为基础模型处选择“qwen2.5-7B-instruct”，微调方法为“lora”，数据集选择“accounting_0722_1_ds.json”，训练轮次修改为 1，学习率修改为 0.00005，批处理大小为 2，LORA 秩修改为 8，LORA 缩放系数为 16，LORA 随机丢弃为 0，其余表格未呈现的内容无需修改，点击【开始】，如图 3-8-1-1。

模型选择与训练

基础模型: qwen2.5-7B-instruct 微调方法: lora

数据集: accounting_0722_1_ds.json 预览

学习率: 0.00005 训练轮数: 1 最大梯度范数: 1 最大样本数: 100000 计算类型: bf16

截断长度: 2048 批处理大小: 2 梯度累积: 8 验证集比例: 0 学习率调节器: cosine

LORA 参数设置

LORA 秩: 8 LORA 缩放系数: 16 LORA 随机丢弃: 0

启用lora 启用DoRA 启用PSSA

开始

图 3-8-1-1 模型内容填写

(2) 待微调完成后，输出模型训练结果，如图 3-8-1-2。

微调模型名称: ECqwen7.22-LowLearn

模型训练结果:

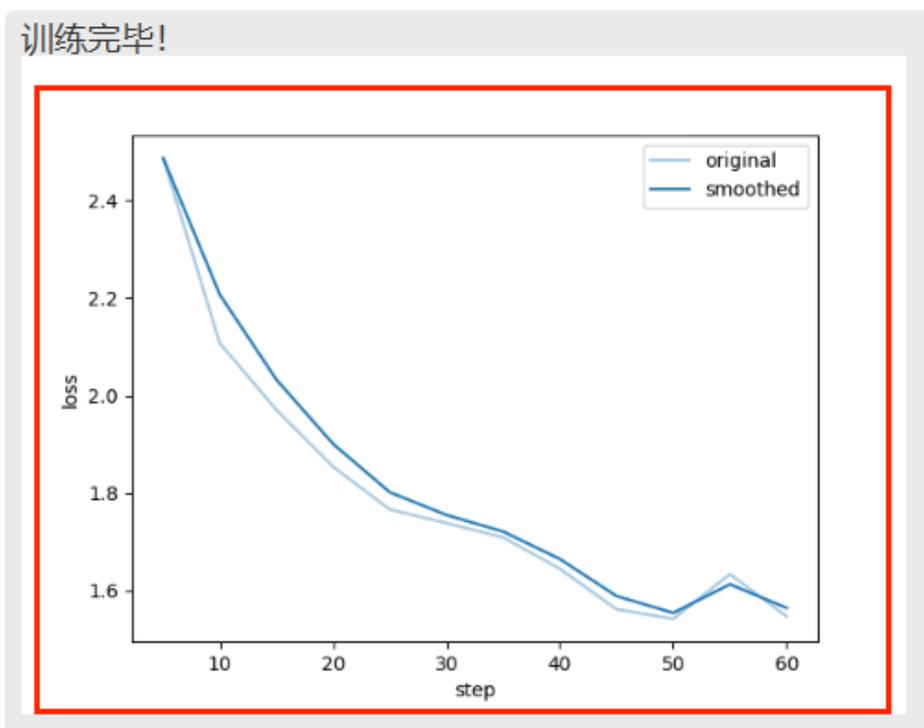


图 3-8-1-2 模型微调结果

步骤二 进行体验测试

(1) 根据任务描述下的要求，调整最大生成长度为 1024，Top-p 采样值为 0.7，温度系数为 0.95，加载模型选择“ECqwen7.22-LowLearn”(如图 3-8-1-3)。



图 3-8-1-3 模型加载与数值调整

(2) 配置完成后，根据任务描述下的示例，在提问框中输入问题，点击【提交】后可查看 AI 输出的结果（如图 3-8-1-4）。在第二次输入问题时，添加系统提示词，再次点击【提交】查看 AI 生成的客服回复（如图 3-8-1-5），可以看出低学习率下，模型对问题的回答效果一般，不能正确回复内容。



图 3-8-1-4 输入问题并查看结果 1



图 3-8-1-5 输入问题并查看结果 2

1.2 高学习率模型体验

步骤一 模型选择

(1) 根据任务描述下提供的编写电商客服的 AI 模型表格，在【模型选择与训练】下方，选择对应的内容，具体为基础模型处选择“qwen2.5-7B-instruct”，微调方法为“lora”，数据集选择“accounting_0722_1_ds.json”，训练轮次修改为 1，学习率修改为 0.0005，批处理大小为 2，LORA 秩修改为 8，LORA 缩放系数为 16，LORA 随机丢弃为 0，其余表格未呈现的内容无需修改，点击【开始】，如图 3-8-1-6。

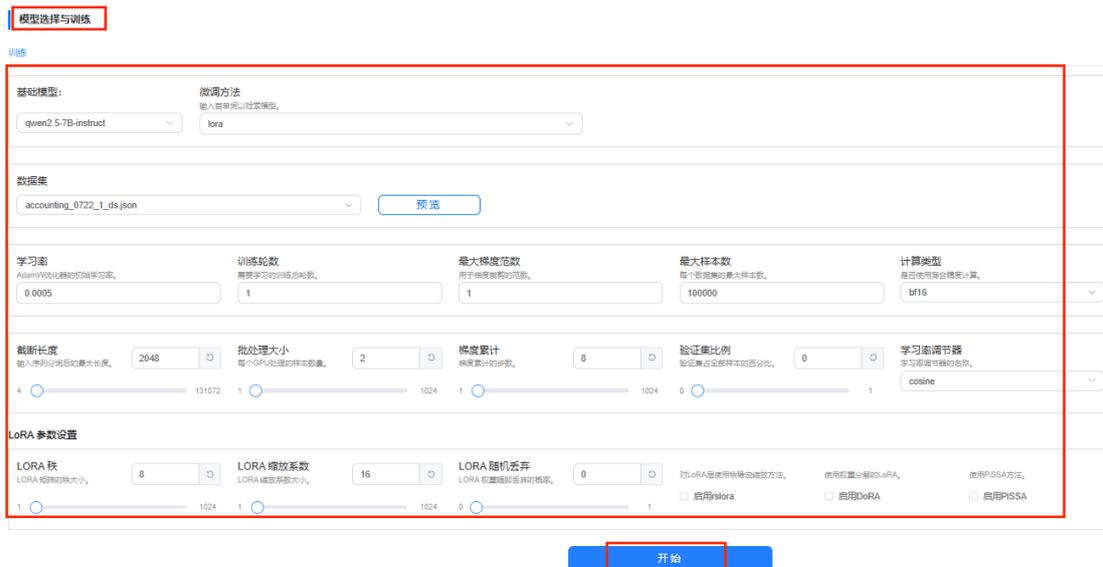


图 3-8-1-6 模型内容填写

(2) 待微调完成后，输出模型训练结果，如图 3-8-1-7。

微调模型名称: ECqwen7.22-HighLearn

模型训练结果:

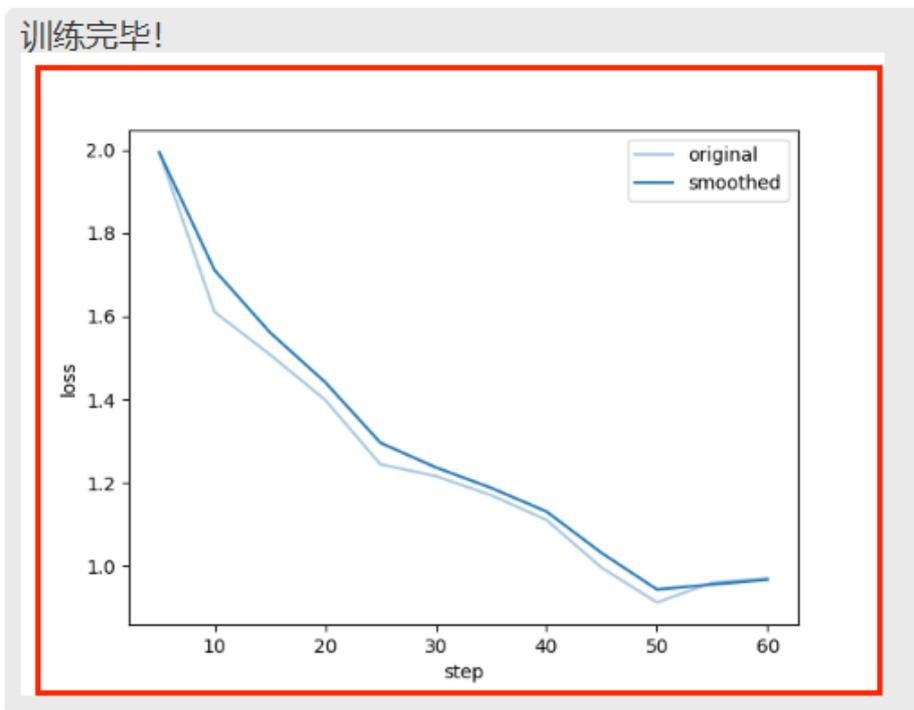


图 3-8-1-7 模型微调结果

步骤二 进行体验测试

(1) 根据任务描述下的要求，调整最大生成长度为 1024，Top-p 采样值为 0.7，温度系数为 0.95，加载模型选择“ECqwen7.22-HighLearn”(如图 3-8-1-8)。

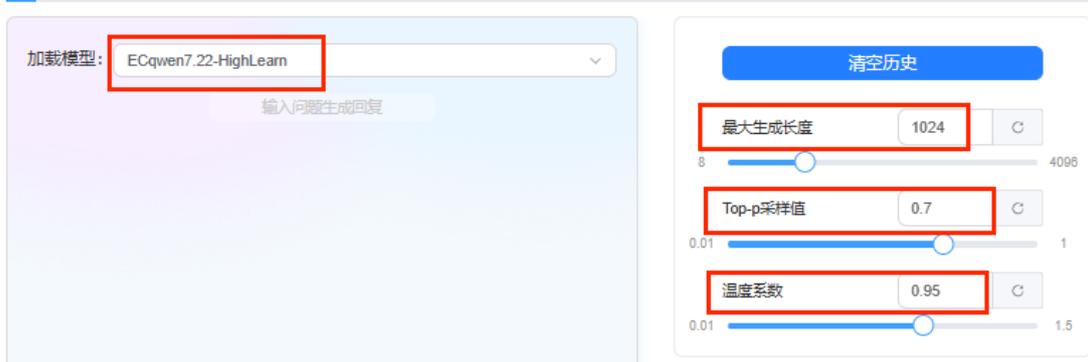


图 3-8-1-8 模型加载与数值调整

(2) 配置完成后，根据任务描述下的示例，在提问框中层层递进的提问以查看模型能力，首先输入场景描述，点击【提交】查看 AI 的回答内容（如图 3-8-1-9），在第一次的基础上，添加系统提示词后点击【提交】查看 AI 的回答内容（如图 3-8-1-10），与正确回复对比，可以看出，在高学习率下，模型回复的准确性提高，说明提高学习率能够有效提高模型的回复正确性。



图 3-8-1-9 AI 回复内容 1



图 3-8-1-10 AI 回复内容 2

(3) 在得到回答后，不点击【清空历史】，连续三次提问“你们发什么快递呀？”，查看 AI 回复内容可知，三次回复无任何差异（如图 3-8-1-11）。点击【清空历史】后，仍三次连续提问“你们发什么快递呀？”，查看 AI 回复内容可知，三次回复依旧无任何差异（如图 3-8-1-12）。



图 3-8-1-11 不清空记录三次提问



图 3-8-1-12 清空记录三次提问

(4) 在提问框中分别使用同义描述、丰富描述和丰富场景描述进行提问，查看三次提问中的 AI 回答内容，如图 3-8-1-13。可见，三种不同描述下，AI 回答的内容差异不大，均围绕发货快递以及指定快递的相关内容回答。



图 3-8-1-13 AI 回答内容查看

1.3 高训练轮次模型体验

步骤一 模型选择

(1) 根据任务描述下提供的编写电商客服的 AI 模型表格，在【模型选择与训练】下方，选择对应的内容，具体为基础模型处选择“qwen2.5-7B-instruct”，微调方法为“lora”，数据集选择“accounting_0722_1_ds.json”，训练轮次修改为 3，学习率修改为 0.0005，批处理大小为 2，LORA 秩修改为 8，LORA 缩放系数为 16，LORA 随机丢弃为 0，其余表格未呈现的内容无需修改，点击【开始】，如图 3-8-1-14。



图 3-8-1-14 模型内容填写

(2) 待微调完成后，输出模型训练结果，如图 3-8-1-15。

微调模型名称: ECqwen7.22-MoreRounds

模型训练结果:

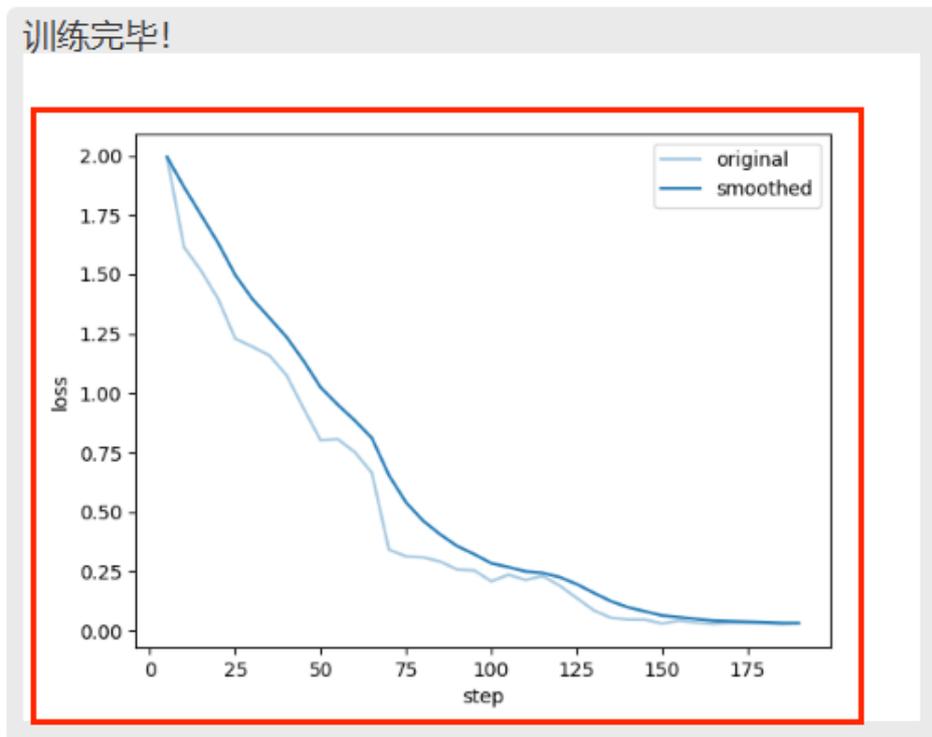


图 3-8-1-15 模型微调结果

步骤二 进行体验测试

(1) 根据任务描述下的要求，调整最大生成长度为 1024，Top-p 采样值为 0.7，温度系数为 0.95，加载模型选择“ECqwen7.22-MoreRounds”(如图 3-8-1-16)。

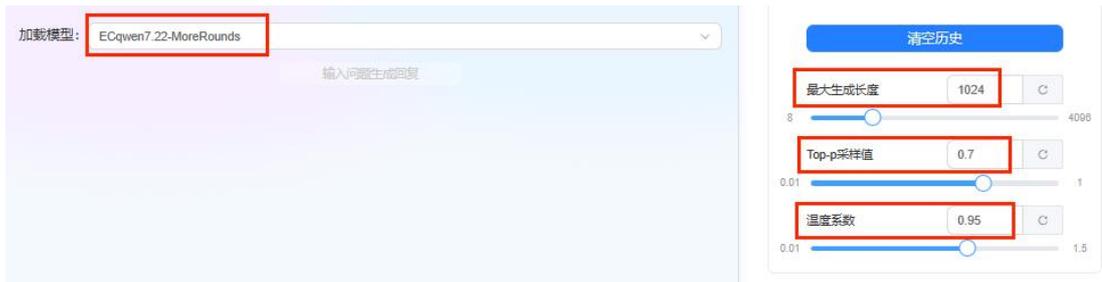


图 3-8-1-16 模型加载和数值填写

(2) 配置完成后，根据任务描述下的示例，在提问框中层层递进的提问以查看模型能力，首先输入场景描述，点击【提交】查看 AI 的回答内容（如图 3-8-1-17），在第一次的基础上，添加系统提示词后点击【提交】查看 AI 的回答内容（如图 3-8-1-18），与正确回复对比，可以看出，在高训练轮次下，模型回复的准确性提高，说明提高训练轮次是提高模型回答准确性和稳定性的有效手段。

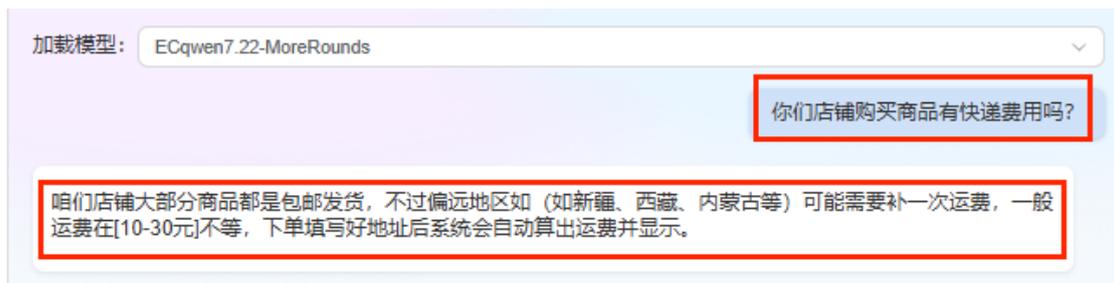


图 3-8-1-17 AI 回复内容 1



图 3-8-1-18 AI 回复内容 2

(3) 在得到回答后，不点击【清空历史】，连续三次提问“你们店铺购买商品有快递费用吗？”，查看 AI 回复内容可知，三次回复无任何差异（如图 3-8-1-19）。点击【清空历史】后，仍三次连续提问“你们店铺购买商品有快递费用吗？”，查看 AI 回复内容可知，清空记录后的回复没有未清空回复的详细

(如图 3-8-1-20)。



图 3-8-1-19 不清空记录三次提问



图 3-8-1-20 清空记录三次提问

(4) 在提问框中分别使用同义描述、丰富描述和丰富场景描述进行提问, 查看三次提问中的 AI 回答内容, 如图 3-8-1-21。可见, 三种不同描述下, AI 回答的内容差异不大, 均围绕发货快递以及指定快递的相关内容回答。



图 3-8-1-21 AI 回答内容查看

2. 训练 AI 智能客服

2.1 同义替换学习训练

步骤一 模型需求获取

(1) 点击【任务描述】下的《客服语料库（同义替换缺陷识别）》以获取语料库（如图 3-8-1-22）。根据任务描述下的步骤进行，首先是审查语料库的内容，其次确定测试目标，即搜索语料库中可能因同义替换而导致模型出错的场景，再基于术语构建测试数据集，最后通过同义替换描述的输出结果与对原始描述的输出结果对比，以分析模型缺陷。

1. 审查语料库内容

请点击下载 [《客服语料库（同义替换缺陷识别）》](#)

图 3-8-1-22 下载语料库

(2) 此处以任务描述的示范操作为例进行操作，先选择加载模型为“ECqwen7.22-MoreRounds”，修改最大生成长度为 1024，Top-p 采样值为 0.7，温度系数为 0.95(如图 3-8-1-23)。选定具体消费咨询场景为“墨兰是什么时候开花的？”，明确其中需测试的核心术语为“墨兰”，确定“墨兰”的同义替换文本有“日向”、“企黑”、“珍珠龙”、“白墨”、“黑珍珠”、“兰阳之松”、“宇宙殿”和“吴字翠”等品种。确定完后，需依次替换“墨兰”为同义替换的文本向 AI 机器人发送，以日向、企黑为例（如图 3-8-1-24），其中与墨兰描述下的输出结果对比，可见描述为日向时的输出结果与原始描述存在差异，企黑的输出结果与原始描述的输出结果差异不大。此外，经验证描述为珍珠龙、兰阳之

松、宇宙殿的输出结果都与原始描述存在差异。



图 3-8-1-23 模型加载与数值修改



图 3-8-1-24 同义替换发送

步骤二 准备训练数据

(1) 点击【任务描述】下的“初始训练数据文件”可下载步骤一中的语料库文件，“同义替换补充语料”可下载对初始语料库补充的场景描述及相应的客服回复内容，如图 3-8-1-25。在《客服语料库（同义替换缺陷识别）》中补充《同义替换补充语料》的内容（如图 3-8-1-26），点击保存并命名为《客服语料库（同义替换补充）》。

②根据步骤1中识别的模型缺陷，在原有的**初始训练数据文件**中（步骤1的文件），及相应的客服回复内容，如**同义替换补充语料**所示。

图 3-8-1-25 语料下载

1014	解答关于缺货少件的解决方案，需要给客户道歉，再提供解决方案，比如：非常抱歉，请您提供照 我收到的兰花叶子数量和商品介绍中提到的不一致，	
1013	解答关于缺货少件的解决方案，需要给客户道歉，再提供解决方案，比如：非常抱歉，请您提供照 我收到的兰花叶子数量和商品介绍中提到的不一致，	
1014	解答关于缺货少件的解决方案，需要给客户道歉，再提供解决方案，比如：非常抱歉，请您提供照 我收到的兰花叶子数量和商品介绍中提到的不一致，	
1015	解答关于缺货少件的解决方案，需要给客户道歉，再提供解决方案，比如：非常抱歉，请您提供照 我收到的兰花没有购买时提到的防虫网，能补发给我	
1014	解答关于缺货少件的解决方案，需要给客户道歉，再提供解决方案，比如：非常抱歉，请您提供照 我收到的兰花没有购买时提到的防虫网，能补发给我	
101	解答关于墨兰的相关产品信息，包括墨兰品种、墨兰开花季节、墨兰养护等内容	你们店铺销售的墨兰主要有哪些？
1013	解答关于墨兰的相关产品信息，包括墨兰品种、墨兰开花季节、墨兰养护等内容	你们店铺有销售什么品种的墨兰？
1013	解答关于墨兰的相关产品信息，包括墨兰品种、墨兰开花季节、墨兰养护等内容	我想要购买一篇墨兰，你可以给我介绍一下吗？
102	解答关于墨兰的相关产品信息，包括墨兰品种、墨兰开花季节、墨兰养护等内容	我朋友在你这购买了一盆墨兰，说很好看，我也想买
102	解答关于墨兰的相关产品信息，包括墨兰品种、墨兰开花季节、墨兰养护等内容	春节要到了，我想买一盆墨兰回家装饰一下，你们
102	解答关于墨兰的相关产品信息，包括墨兰品种、墨兰开花季节、墨兰养护等内容	企黑是墨兰吗？我看他好像很好看，你们店铺有其它
102	解答关于墨兰的相关产品信息，包括墨兰品种、墨兰开花季节、墨兰养护等内容	墨兰是只有一种吗？还是有多种的，可以给我介绍一
102	解答关于墨兰的相关产品信息，包括墨兰品种、墨兰开花季节、墨兰养护等内容	企黑是墨兰吗？
102	解答关于墨兰的相关产品信息，包括墨兰品种、墨兰开花季节、墨兰养护等内容	日向是墨兰吗？
102	解答关于墨兰的相关产品信息，包括墨兰品种、墨兰开花季节、墨兰养护等内容	企黑什么时候开花？
102	解答关于墨兰的相关产品信息，包括墨兰品种、墨兰开花季节、墨兰养护等内容	企黑开花季节是什么时候？
102	解答关于墨兰的相关产品信息，包括墨兰品种、墨兰开花季节、墨兰养护等内容	企黑是冬季开花的吗？
102	解答关于墨兰的相关产品信息，包括墨兰品种、墨兰开花季节、墨兰养护等内容	企黑可以存春节前开花吗？

图 3-8-1-26 语料补充

(2) 在利用微调数据集训练大模型时，需将微调数据集转换为 json 文件，在任务描述下有两种方法，首先介绍方法一的具体操作：点击“Excel 转换成 json 文件.zip”以下载相关 VBA 宏，解压后可获得 VBA 宏的文本。以使用 wps 为例，在《客服语料库（同义替换补充）》中，进入【VB 编辑器】，点击【插入】下的【模块】（如图 3-8-1-27）。将解压后的 VBA 宏文本内容复制，粘贴至模块的代码编辑处，点击【运行】（如图 3-8-1-28）。运行后，在保存路径处找到 json 格式的文件，右键文件后打开方式选择【记事本】，在【文件】处点击【另存为】，修改文件名为“WordFlex”，【编码】处选择“UTF-8”后点击【保存】（如图 3-8-1-29）。

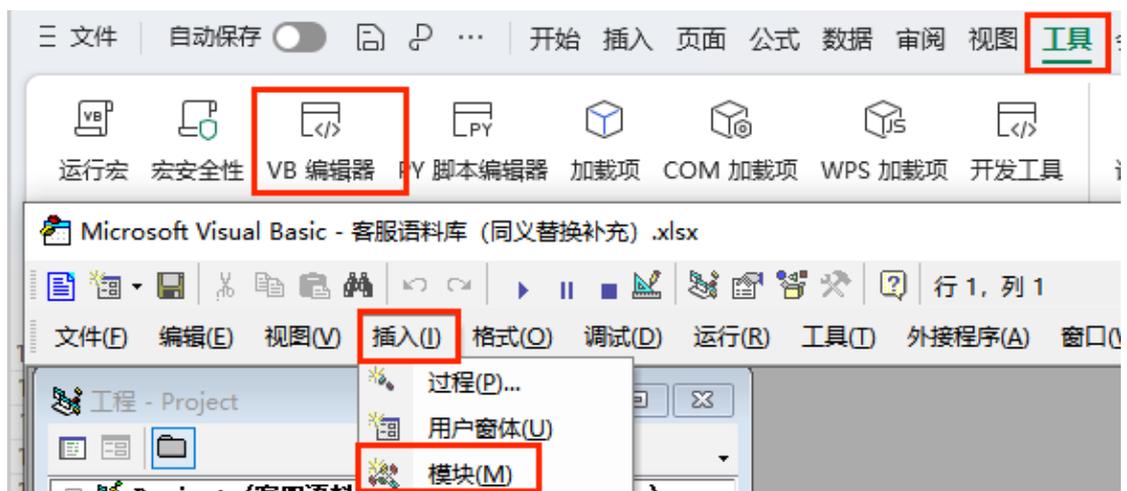


图 3-8-1-27 插入模块

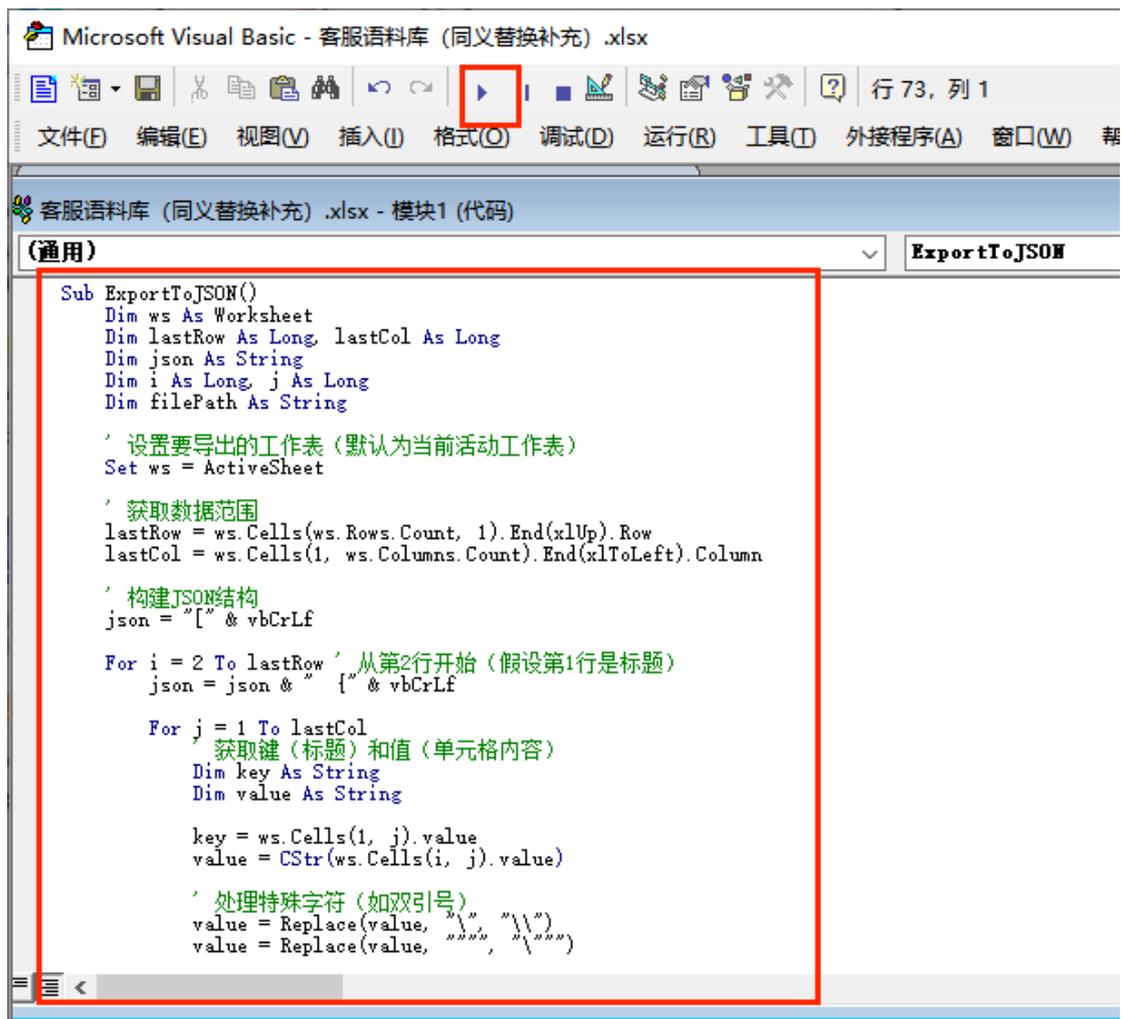


图 3-8-1-28 运行代码

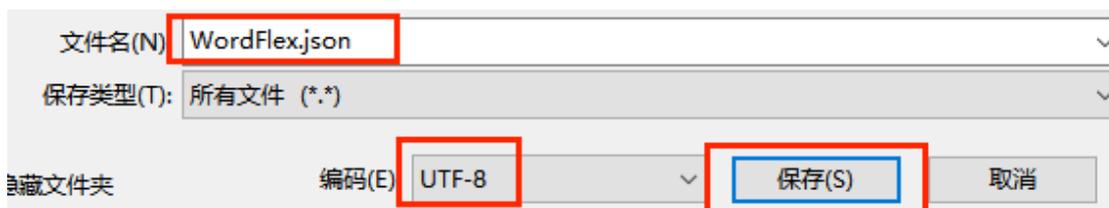


图 3-8-1-29 修改文件名称和格式

(3) 方法二是利用 Python 转换数据，先在【文件列表】处上传《客服语料库（同义替换补充）》（如图 3-8-1-30）。根据任务描述下的要求“转换为 json 文件、编码格式为 utf-8”，在【技术需求转化】处对应关键词填写参数，其中上传的数据文件为“客服语料库（同义替换缺陷识别）.xlsx”，输出的 json 文件名为“WordFlex”，json 文件编码格式为“utf-8”（如图 3-8-1-31）。参数填写后点击【执行并显示结果】，传入参数数据，执行数据代码，执行完成后可查看输出结果，点击文件右侧的【下载】可下载相应压缩包（如图 3-8-1-32），

解压后即可获取转换后的数据文件

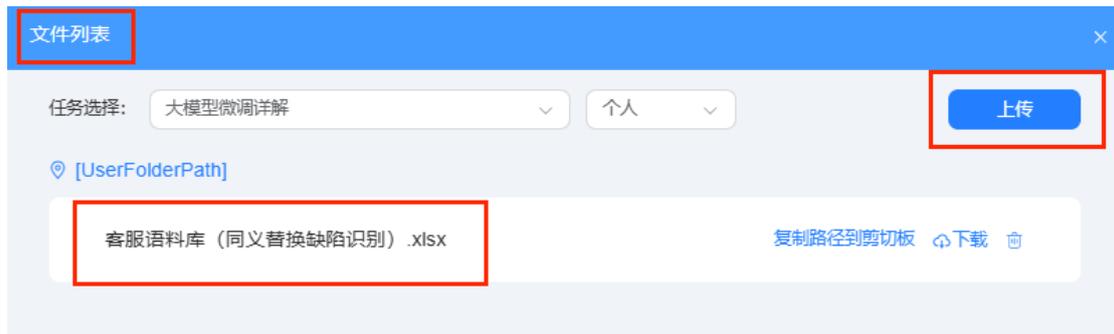


图 3-8-1-30 文件上传

json格式转换 (2.00分)



图 3-8-1-31 参数填写

json格式转换 (2.00分)



图 3-8-1-32 数据下载

步骤三 进行模型训练

(1) 根据任务描述下提供的编写电商客服的 AI 模型表格，在【模型训练】下方，选择对应的内容，具体为基础模型处选择“qwen2.5-7B-instruct”，微

调方法为“lora”，数据集选择步骤二中转换的 json 文件“WordFlex.json”，数据集训练轮次修改为 3，学习率修改为 0.0005，批处理大小为 2，LORA 秩修改为 16，LORA 缩放系数为 32，LORA 随机丢弃为 0，其余表格未呈现的内容无需修改，点击【开始】，如图 3-8-1-33。

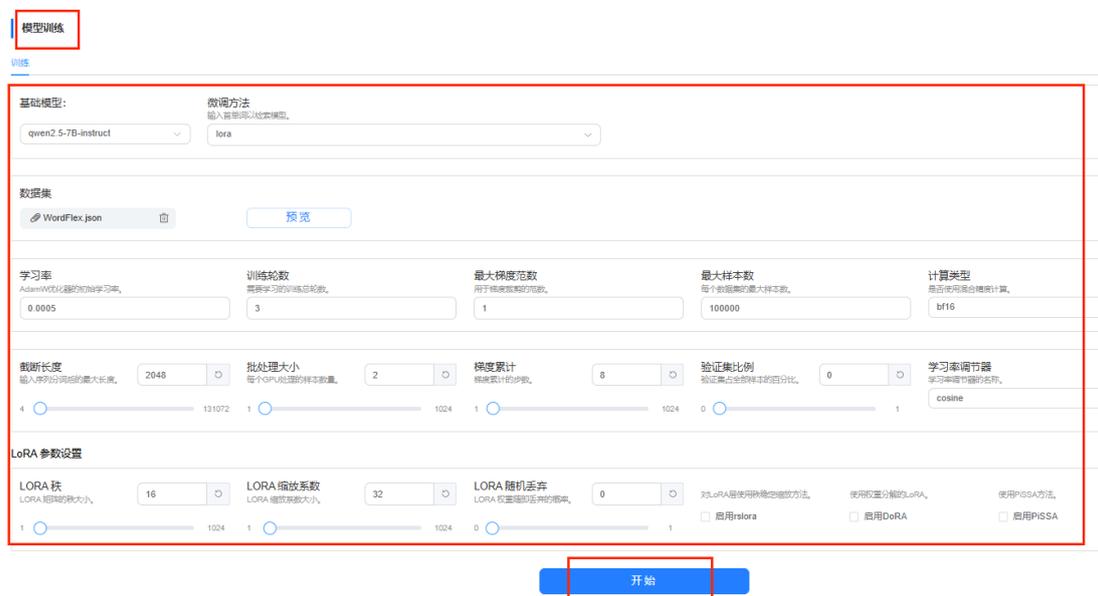


图 3-8-1-33 模型内容填写

(2) 待微调完成后，输出模型训练结果，如图 3-8-1-34。

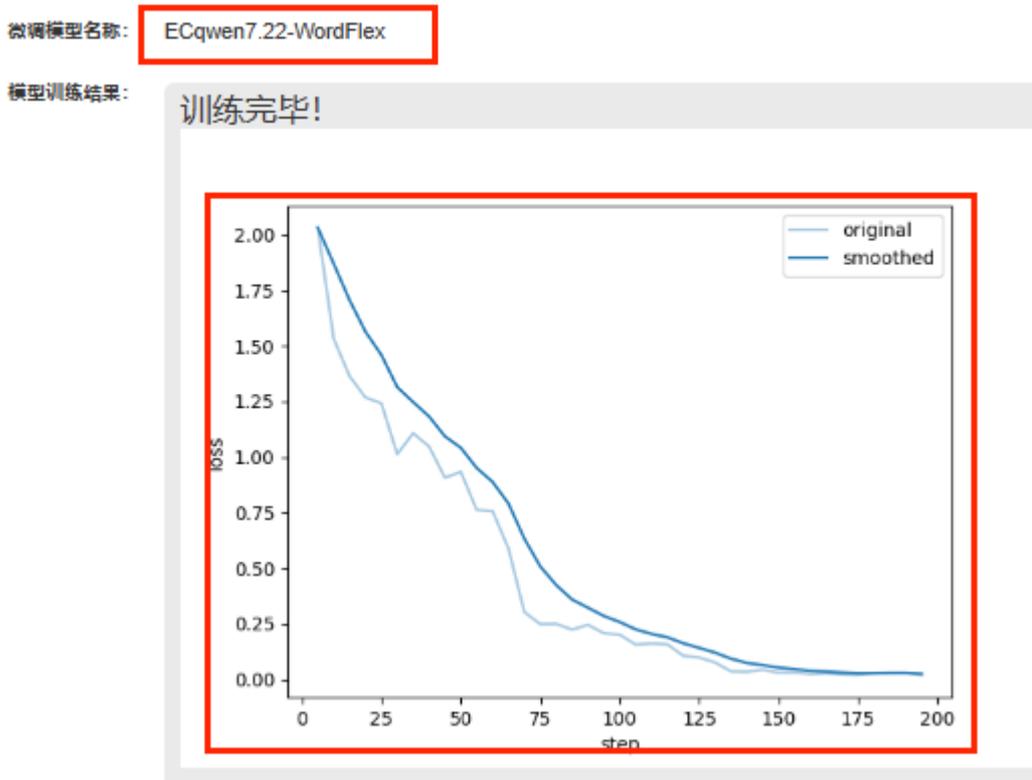


图 3-8-1-34 模型训练结果输出

步骤四 验证训练模型

(1) 根据任务描述下的要求，调整最大生成长度为 1024，Top-p 采样值为 0.7，温度系数为 0.95，加载模型选择“ECqwen7.22-WordFlex”(如图 3-8-1-35)。

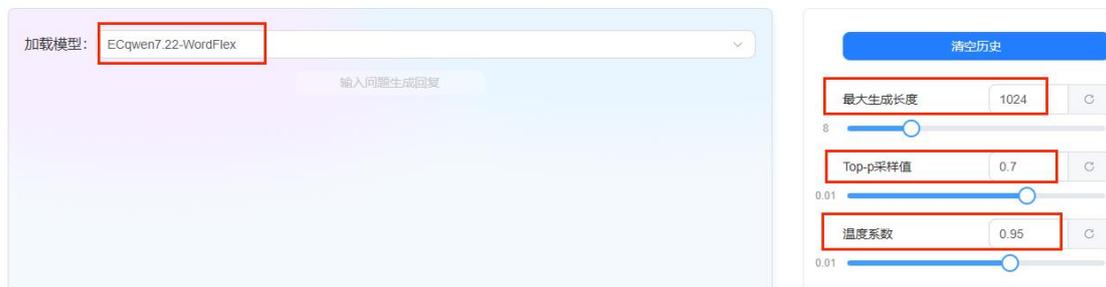


图 3-8-1-35 模型加载与数值修改

(2) 根据任务描述下的步骤进行训练模型验证，首先针对核心术语进行同义替换以评估模型对同义替换词语的处理能力，其次从“记忆力、理解力和推理力”三个关键维度设计验证测试，再对模型验证打分，最后邀请外部验证与综合评估。此处以任务描述下的示例进行具体操作，核心术语为“墨兰”，同义替换墨兰的一些子品种“日向”、“企黑”、“珍珠龙”、“白墨”、“黑珍珠”、“兰阳之松”、“宇宙殿”和“吴字翠”等。

(3) 原始的具体咨询场景仍为“墨兰是什么时候开花？”，点击【清空历史】(如图 3-8-1-36)之后依次替换“墨兰”为上述确定的一些子品种，以查看同义替换下模型的回复情况。依旧以“日向”和“企黑”为例，如图 3-8-1-37。其中描述为日向时其回答仍与原始描述的答案存在一定差异，可见模型在同义替换下仍存在部分不稳定。此外，经验证除珍珠龙和白墨的描述下答案会有所差异外，其余均与原始描述所输出的答案无太大差异，说明模型对同义替换词语的处理能力有所提高。

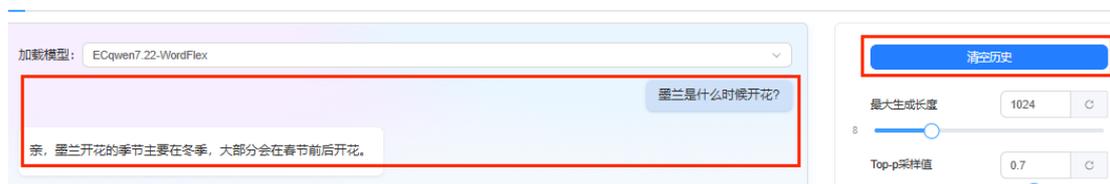


图 3-8-1-36 原始描述



图 3-8-1-37 同义替换描述

步骤五 模型训练报告

(1) 点击【答题卡下载】下的“训练报告.xlsx”以下载分析报告模板（如图 3-8-1-38）。



图 3-8-1-38 模板下载

(2) 根据步骤三中模型参数填写，完成《训练报告》中的“AI 模型参数信息”表格内容填写（如图 3-8-1-39），并根据自己使用情况与调查他人的使用情况，结合表格中验证内容及评分标准，在“模型验证得分”表格内打分，如图 3-8-1-40。在子表“统计和可视化分析”中可直观呈现模型各维度得分情况。完成后点击保存将报告上传至系统。

AI模型参数信息	
使用语料	WordFlex.xlsx
训练轮次	3
学习率	0.0005
其他参数	批处理大小: 2; lora秩: 16; lora缩放系数: 32; lora随机丢弃: 0
训练得分	90.5

图 3-8-1-39 AI 模型参数信息填写

模型验证得分				
验证维度	验证内容	分数设置	自评得分	他评得分
			完全满足验证：满分； 大部分满足验证：扣1分； 部分满足验证：扣2分； 少量满足验证：扣3分； 完全不满足验证：0	完全满足验证：满分； 大部分满足验证：扣1分； 部分满足验证：扣2分； 少量满足验证：扣3分； 完全不满足验证：0
记忆力维度	1. 模型能够根据消费者咨询输出客服回复	5	4	4
	2. 模型输出的客服回复准确	5	4	4
	3. 模型在多次询问相同的消费者咨询后，不存在记忆衰退或者输出不稳定的问题	10	9	8
理解力维度	1. 模型可以根据不同描述但本质相同的业务场景输出正确的客服回复。	10	9	8
	2. 模型可以处理包含错误或不完整信息的消费者咨询	12	11	12
	3. 模型在面临特点领域的专有名称替换为其他内容的业务场景时，能输出正确的客服回复。	12	10	11
推理力维度	1. 模型在高重合度场景测试中，能准确地应用已学习到的知识，输出正确的客服回复。	12	12	11
	2. 模型在低重合度场景测试中，能准确地应用已学习到的知识，输出正确的客服回复。	16	15	14

图 3-8-1-40 模型验证得分填写

2.2 情感理解强化学习训练

步骤一 模型需求获取

(1) 点击【任务描述】下的《客服语料库（情感理解缺陷识别）》以获取语料库（如图 3-8-1-41）。根据任务描述下的步骤进行，首先是审查语料库的内容，其次确定测试目标，即标记语料库中涉及消费者情感的咨询场景，再基于标记的场景，用同义表示构建测试数据集，最后通过对比大模型对情感情绪的识别情况，以分析模型缺陷。

任务描述: 

请点击下载 [《客服语料库（情感理解缺陷识别）》](#) 仔细

图 3-8-1-41 下载语料库

(2) 此处以任务描述的示范操作为例进行操作，先选择加载模型为“ECqwen7.22-WordFlex”，修改最大生成长度为 1024，Top-p 采样值为 0.7，温度系数为 0.95(如图 3-8-1-42)。选定具体消费咨询场景为“我很急，你们什么时候能发货送到？”，明确其中需测试的咨询场景为“加急紧迫性”，发送原始场景描述后，用同义表示描述为“我大后天就要用，你们快递能送达吗？”再次输入给模型，两次回答如图 3-8-1-43 所示，可以发现两次回答均不能对客户有所安抚，帮提供解决方案，因此需要进行语料库的调整。

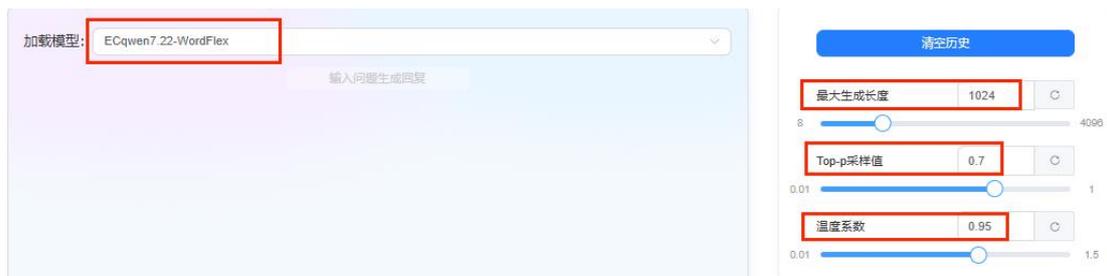


图 3-8-1-42 模型加载与数值修改



图 3-8-1-43 同义表示发送

步骤二 准备训练数据

(1) 点击【任务描述】下的“初始训练数据文件”可下载步骤一中的语料库文件，“情感理解补充语料”可下载对初始语料库补充的场景描述及相应的客服回复内容，如图 3-8-1-44。在《客服语料库（情感理解缺陷识别）》内补充《情感理解补充语料》的内容（如图 3-8-1-45），点击保存并命名为《客服语料库（情感理解补充）》。

任务描述

1.准备语料数据:

- ①详细审阅提供的用于训练智能客服AI机器人的语料库，确保充分理解其输入与输出
- ②根据步骤1中识别的模型缺陷，在原有的**初始训练数据文件**中（步骤1的文件），域的场景描述，以及相应的客服回复内容，如**情感理解补充语料**所示。

图 3-8-1-44 语料下载

1019	解答关于墨兰的相关产品信息，包括墨兰品种、墨兰开花季节、墨兰养护等内容	日向什么时候开花
1020	解答关于墨兰的相关产品信息，包括墨兰品种、墨兰开花季节、墨兰养护等内容	日向什么时候开花
1021	识别用户物流的紧迫性和指定快递需求，针对不同的紧迫性与用户需求，推荐发货的快递。	我很急，你们什么时候能发货送到？
1022	识别用户物流的紧迫性和指定快递需求，针对不同的紧迫性与用户需求，推荐发货的快递。	我过两天就要送人了，很急，能不能
1023	识别用户物流的紧迫性和指定快递需求，针对不同的紧迫性与用户需求，推荐发货的快递。	可以帮我发顺丰快递吗？我很急
1024	识别用户物流的紧迫性和指定快递需求，针对不同的紧迫性与用户需求，推荐发货的快递。	我大后天就要用，你们快递能送达吗
1025	识别用户物流的紧迫性和指定快递需求，针对不同的紧迫性与用户需求，推荐发货的快递。	你们发货了吗？我很急，能不能给我
1026	识别用户物流的紧迫性和指定快递需求，针对不同的紧迫性与用户需求，推荐发货的快递。	急！求优先发货
1027	识别用户物流的紧迫性和指定快递需求，针对不同的紧迫性与用户需求，推荐发货的快递。	领导说这批兰花这两天就要送到，你

图 3-8-1-45 语料补充

(2) 在利用微调数据集训练大模型时，需将微调数据集转换为 json 文件，在任务描述下有两种方法，首先介绍方法一的具体操作：点击“Excel 转换成 json 文件.zip”以下载相关 VBA 宏，解压后可获得 VBA 宏的文本。以使用 wps 为例，在《客服语料库（情感理解补充）》中，进入【VB 编辑器】，点击【插入】下的【模块】（如图 3-8-1-46）。将解压后的 VBA 宏文本内容复制，粘贴至模块的代码编辑处，点击【运行】（如图 3-8-1-47）。运行后，在保存路径处找到 json 格式的文件，右键文件后打开方式选择【记事本】，在【文件】处点击【另存为】，修改文件名为“FeelSense”，【编码】处选择“UTF-8”后点击【保存】（如图 3-8-1-48）。

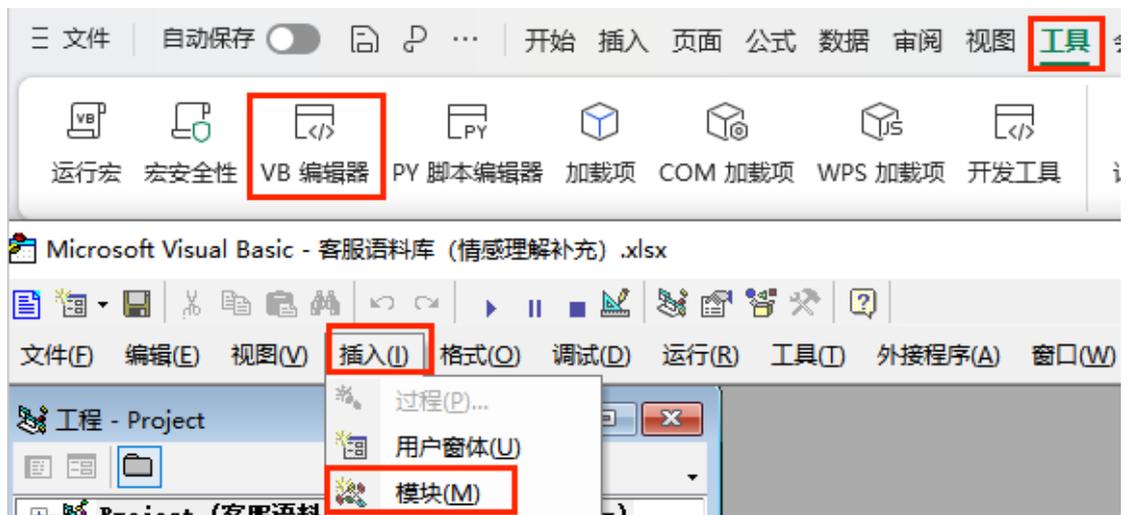


图 3-8-1-46 插入模块

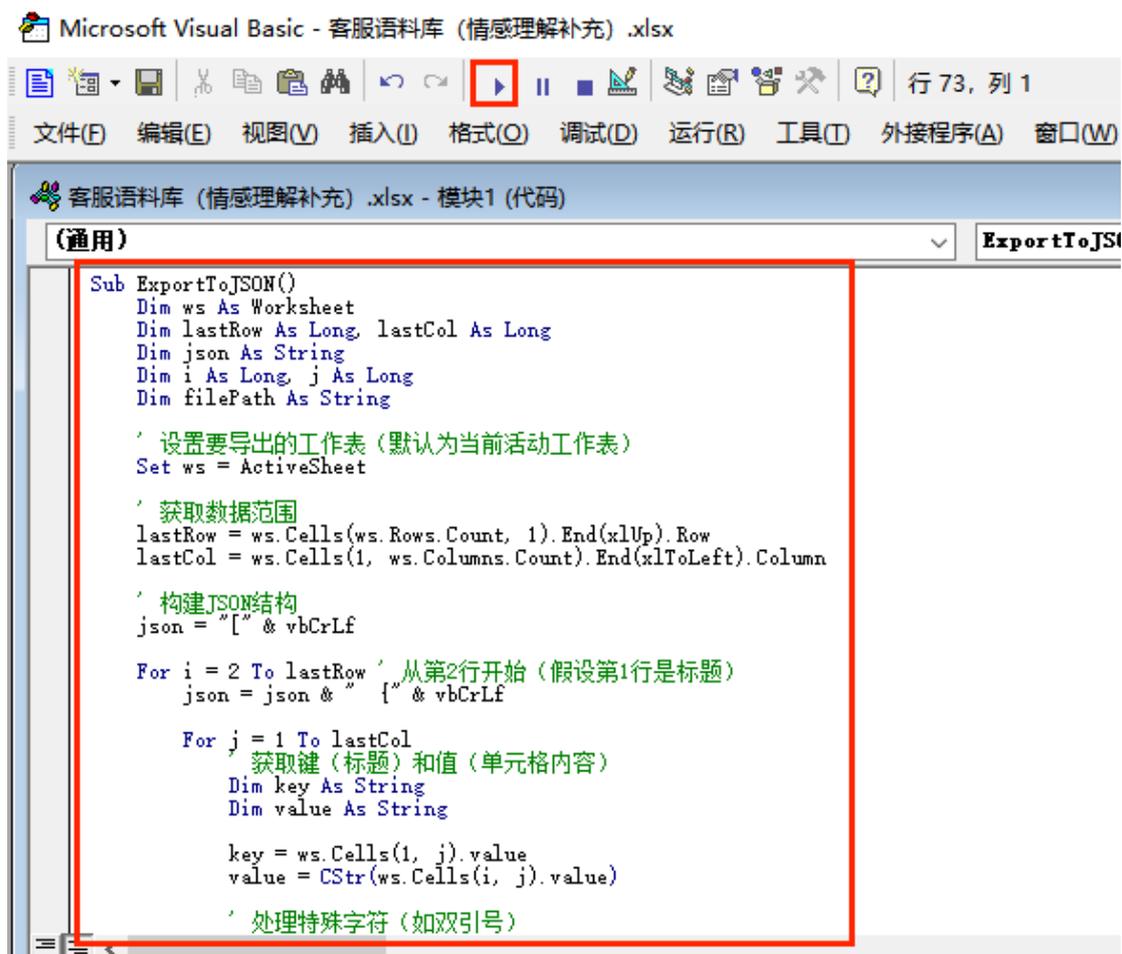


图 3-8-1-47 运行代码

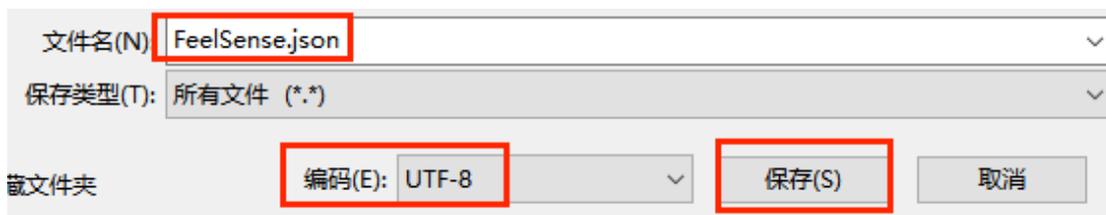


图 3-8-1-48 修改文件名称和格式

(3) 方法二是利用 Python 转换数据，先在【文件列表】处上传《客服语料库（情感理解补充）》（如图 3-8-1-49）。根据任务描述下的要求“转换为 json 文件、编码格式为 utf-8”，在【技术需求转化】处对应关键词填写参数，其中上传的数据文件为“客服语料库（情感理解补充）.xlsx”，输出的 json 文件名为“FeelSense”，json 文件编码格式为“utf-8”（如图 3-8-1-50）。参数填写后点击【执行并显示结果】，传入参数数据，执行数据代码，执行完成后可查看输出结果，点击文件右侧的【下载】可下载相应压缩包（如图 3-8-1-51），解压后即可获取转换后的数据文件。

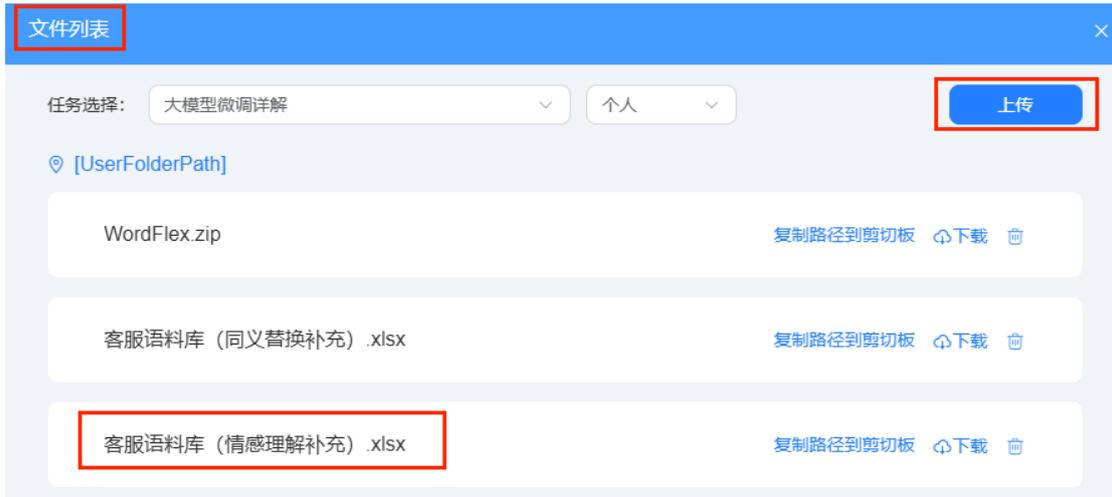


图 3-8-1-49 文件上传



图 3-8-1-50 参数填写



执行并显示结果



图 3-8-1-51 数据下载

步骤三 进行模型训练

(1) 根据任务描述下提供的编写电商客服的 AI 模型表格，在【进行模型训练】下方，选择对应的内容，具体为基础模型处选择“qwen2.5-7B-instruct”，微调方法为“lora”，数据集选择步骤二中转换的 json 文件“FeelSense.json”，数据集训练轮次修改为 3，学习率修改为 0.0005，批处理大小为 2，LORA 秩修改

为 16，LORA 缩放系数为 32，LORA 随机丢弃为 0，其余表格未呈现的内容无需修改，点击【开始】，如图 3-8-1-52。

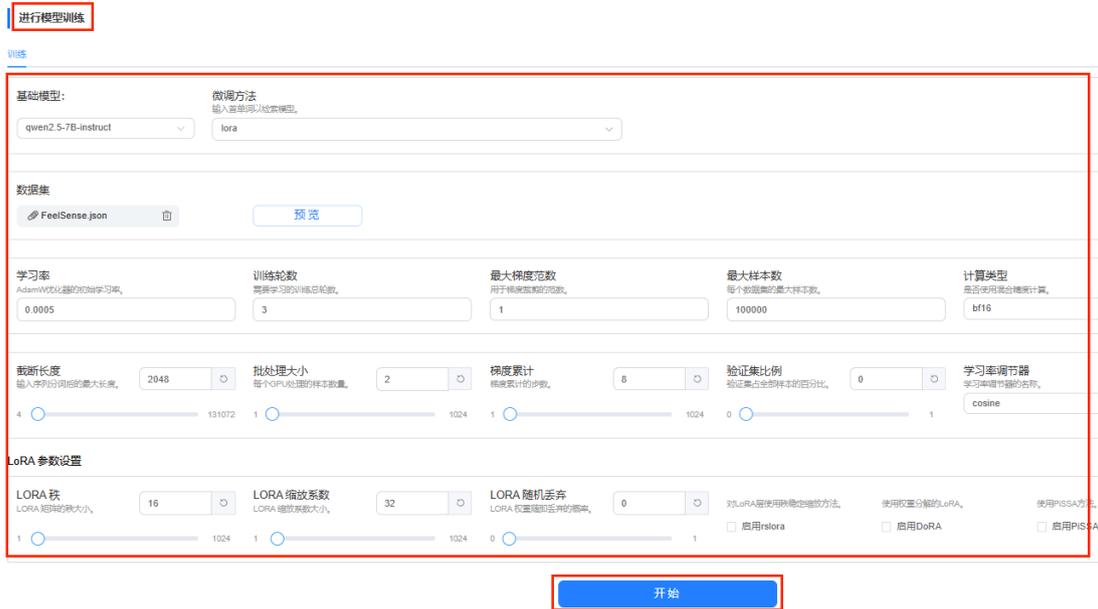


图 3-8-1-52 模型内容填写

(2) 待微调完成后，输出模型训练结果，如图 3-8-1-53。

微调模型名称: ECqwen7.22-FeelSense

模型训练结果:

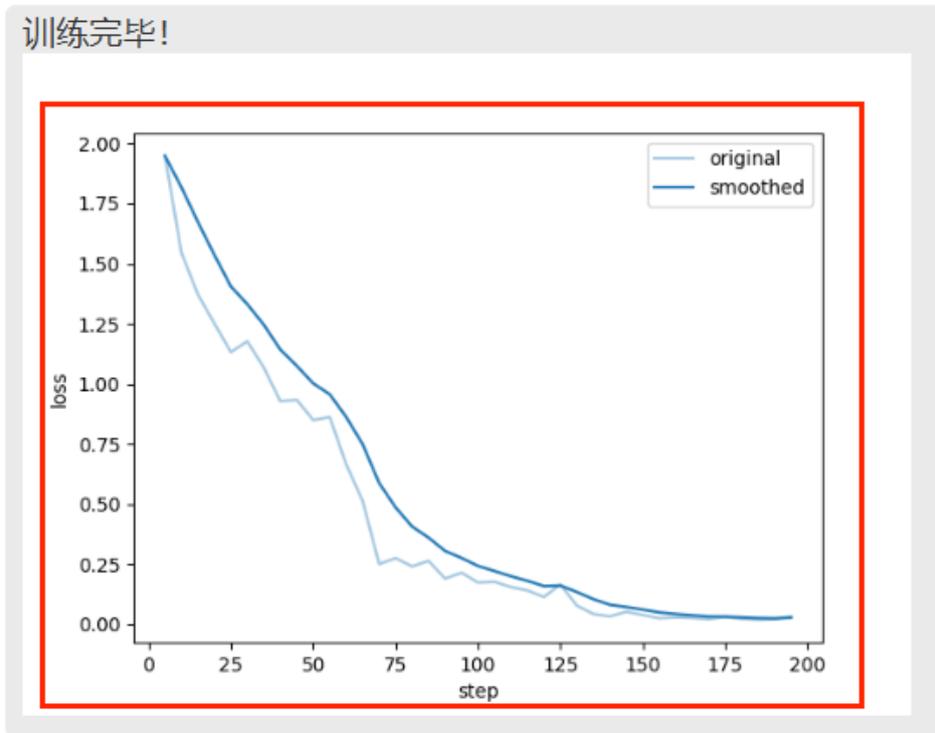


图 3-8-1-53 模型训练结果输出

步骤四 验证训练模型

(1) 根据任务描述下的要求，调整最大生成长度为 1024，Top-p 采样值为

0.7, 温度系数为 0.95, 加载模型选择“ECqwen7.22-FeelSense”(如图 3-8-1-54)。

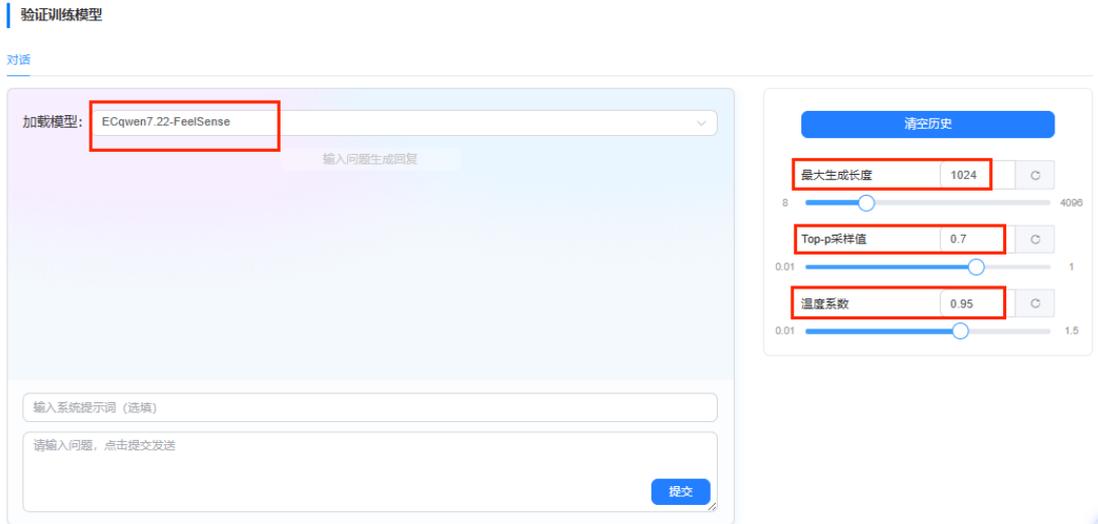


图 3-8-1-54 模型加载与数值修改

(2) 根据任务描述下的步骤进行训练模型验证, 首先针对场景替换同义表述以评估模型对同义表示场景的处理能力, 其次从“记忆力、理解力和推理力”三个关键维度设计验证测试, 再对模型验证打分, 最后邀请外部验证与综合评估。此处以任务描述下的示例进行具体操作, 场景为“加急紧迫性”, 同义表述为“我在你们这购买了盆墨兰, 由于我要送人, 比较急, 能不能给我快点送到?”。

(3) 原始的具体咨询场景仍为“你们发货了吗? 我很急, 能不能给我优先发货。”, 点击【清空历史】(如图 3-8-1-55)之后输入同义表述, 以查看同义表述下模型的回复情况(如图 3-8-1-56)。可以看到, 无论是原始的咨询内容还是同义的咨询内容, 经过语料库的调整, 模型已经能够给出解决方案, 模型的实用性大大加强。



图 3-8-1-55 原始描述

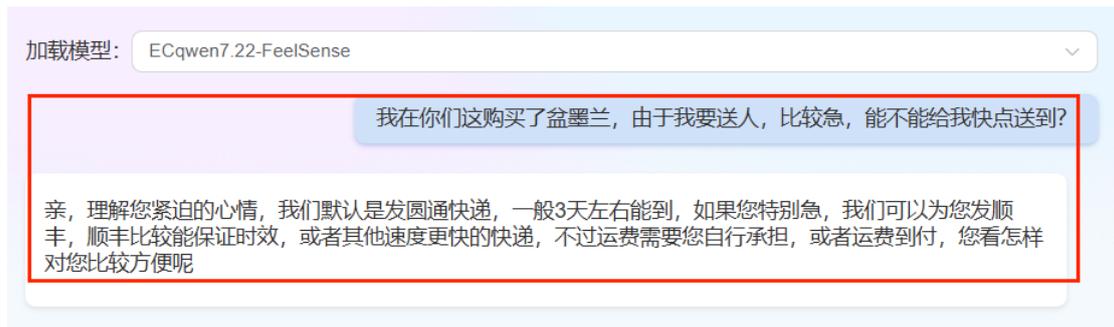


图 3-8-1-56 同义替换描述

步骤五 模型训练报告

(1) 点击【答题卡下载】下的“训练报告.xlsx”以下载分析报告模板（如图 3-8-1-57）。



图 3-8-1-57 模板下载

(2) 根据步骤三中模型参数填写，完成《训练报告》中的“AI 模型参数信息”表格内容填写（如图 3-8-1-58），并根据自己使用情况与调查他人的使用情况，结合表格中验证内容及评分标准，在“模型验证得分”表格内打分，如图 3-8-1-59。在子表“统计和可视化分析”中可直观呈现模型各维度得分情况。完成后点击保存将报告上传至系统。

AI模型参数信息	
使用语料	FeelSense.json
训练轮次	3
学习率	0.0005
其他参数	lora秩: 16; lora缩放系数: 32; lora随机丢弃: 0
训练得分	97

图 3-8-1-58 AI 模型参数信息填写

模型验证得分				
验证维度	验证内容	分数设置	自评得分	他评得分
			完全满足验证：满分； 大部分满足验证：扣1分； 部分满足验证：扣2分； 少量满足验证：扣3分； 完全不满足验证：0	完全满足验证：满分； 大部分满足验证：扣1分； 部分满足验证：扣2分； 少量满足验证：扣3分； 完全不满足验证：0
记忆力维度	1. 模型能够根据消费者咨询输出客服回复	5	5	
	2. 模型输出的客服回复准确	5	5	
	3. 模型在多次询问相同的消费者咨询后，不存在记忆衰退或者输出不稳定的问题	10	10	1
理解力维度	1. 模型可以根据不同描述但本质相同的业务场景输出正确的客服回复。	10	10	1
	2. 模型可以处理包含错误或不完整信息的消费者咨询	12	12	1
	3. 模型在面临特点领域的专有名称替换为其他内容的业务场景时，能输出正确的客服回复。	12	12	1
	1. 模型在高重合度场景测试中，能准确地应用已学习到的知识，输出正确的客服回复。	12	12	1
	2. 模型在低重合度场景测试中，能准			

图 3-8-1-59 模型验证得分填写

2.3 计算能力强化学学习训练

步骤一 模型需求获取

(1) 点击【任务描述】下的《客服语料库（计算缺陷识别）》以获取语料库（如图 3-8-1-60）。根据任务描述下的步骤进行，首先是审查语料库的内容，其次确定测试目标，即针对由于涉及计算而导致模型输出错误的电商场景以明确测试的核心术语，再基于确定的术语，用与场景一致，同样涉及计算的消费者咨询场景构建测试数据集，最后分析模型缺陷。

任务描述： 

1. 审查语料库内容

请点击下载 [《客服语料库（计算缺陷识别）》](#)。

图 3-8-1-60 下载语料库

(2) 此处以任务描述的示范操作为例进行操作，先选择加载模型为“ECqwen7.22-FeelSense”，修改最大生成长度为 1024，Top-p 采样值为 0.7，温度系数为 0.95(如图 3-8-1-61)。选定具体消费咨询场景为“你们店铺的价保周期是多少天？”输入给模型（如图 3-8-1-62），其中的“价保周期”为核心测试术语，以此确定核心术语的一系列涉及计算的文本，在点击【清空历史】后再输入给模型（如图 3-8-1-63）。查看涉及计算的文本输入后的回答可知，

模型对价保期的计算有误，2月11日到2月28日超过价保周期，但模型判断2月28日在价保周期内。

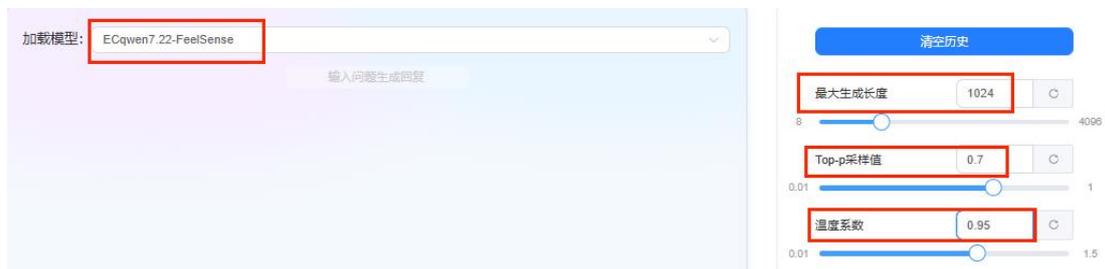


图 3-8-1-61 模型加载与数值修改



图 3-8-1-62 具体消费咨询场景

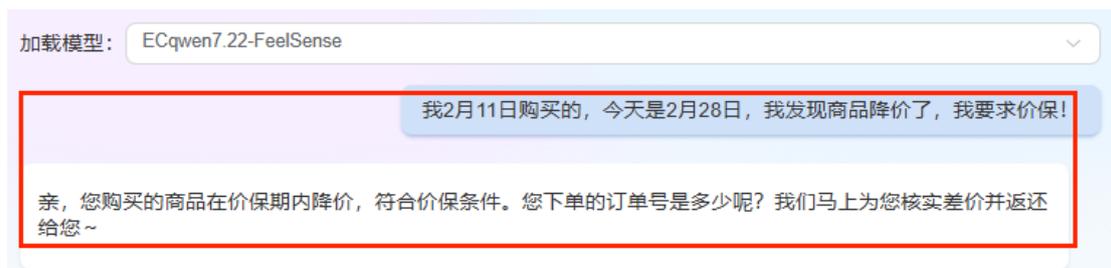


图 3-8-1-63 涉及计算场景

步骤二 准备训练数据

(1) 点击【任务描述】下的“初始训练数据文件”可下载步骤一中的语料库文件，“计算能力补充语料”可下载对初始语料库补充的场景描述及相应的客服回复内容，如图 3-8-1-64。在《客服语料库（计算缺陷识别）》内补充《计算能力补充语料》的内容（如图 3-8-1-65），点击保存并命名为《客服语料库（计算能力补充）》。

任务描述

1.准备语料数据:

- ①详细审阅提供的用于训练智能客服AI机器人的语料库,确保充分理解其输入与输出
- ②根据步骤1中识别的模型缺陷,在原有的**初始训练数据文件**中(步骤1的文件),
容,如**计算能力补充语料**所示。

图 3-8-1-64 语料下载

1026	识别用户物流的紧迫性和指定快递需求,针对不同的紧迫性与用户需求,推荐发货的快递。	急!求优先发货
1027	识别用户物流的紧迫性和指定快递需求,针对不同的紧迫性与用户需求,推荐发货的快递。	领导说这批兰花这两天就要送到,你们可以做
1028	对价保周期进行计算,判断用户的价保是否超过价保期限(周期),超过则不允许价保	我2月11日购买的,今天是2月28日,我发现商
1029	对价保周期进行计算,判断用户的价保是否超过价保期限(周期),超过则不允许价保	我3月11号购买的,今天是3月28号,我发现商
1030	对价保周期进行计算,判断用户的价保是否超过价保期限(周期),超过则不允许价保	我3月28号购买的,今天是4月11号,我发现商
1031	对价保周期进行计算,判断用户的价保是否超过价保期限(周期),超过则不允许价保	我5月29日购买的,今天是6月9日,我发现商
1032	对价保周期进行计算,判断用户的价保是否超过价保期限(周期),超过则不允许价保	我发现商品降价了,这才几天,把差价补给我
1033	对价保周期进行计算,判断用户的价保是否超过价保期限(周期),超过则不允许价保	我发现商品降价了,这才几天,把差价补给我
1034	对价保周期进行计算,判断用户的价保是否超过价保期限(周期),超过则不允许价保	你们店铺有价保服务对吧,我7月1日购买的商
1035	对价保周期进行计算,判断用户的价保是否超过价保期限(周期),超过则不允许价保	你们价保周期是多久,我9月29日购买的兰花,
1036	对价保周期进行计算,判断用户的价保是否超过价保期限(周期),超过则不允许价保	我前天买的兰花是35.9,今天怎么去看就25.9
1037	对价保周期进行计算,判断用户的价保是否超过价保期限(周期),超过则不允许价保	我今天看这个兰花就降价了,我3号购买的三
1038	对价保周期进行计算,判断用户的价保是否超过价保期限(周期),超过则不允许价保	我周一购买的兰花,今天周五就降价了,我要
1039	对价保周期进行计算,判断用户的价保是否超过价保期限(周期),超过则不允许价保	我上周五购买的兰花,今天才周四,就降价了
1040	对价保周期进行计算,判断用户的价保是否超过价保期限(周期),超过则不允许价保	周一我才下单买了这盆兰花,今天才周日,就
1041	对价保周期进行计算,判断用户的价保是否超过价保期限(周期),超过则不允许价保	我购买兰花还不到5天,它就降价了,你们要
1042	对价保周期进行计算,判断用户的价保是否超过价保期限(周期),超过则不允许价保	我5月11号购买的,今天是5月14号,我发现商

图 3-8-1-65 语料补充

(2) 在利用微调数据集训练大模型时,需将微调数据集转换为 json 文件,在任务描述下有两种方法,首先介绍方法一的具体操作:点击“Excel 转换成 json 文件.zip”以下载相关 VBA 宏,解压后可获得 VBA 宏的文本。以使用 wps 为例,在《客服语料库(计算能力补充)》中,进入【VB 编辑器】,点击【插入】下的【模块】(如图 3-8-1-66)。将解压后的 VBA 宏文本内容复制,粘贴至模块的代码编辑处,点击【运行】(如图 3-8-1-67)。运行后,在保存路径处找到 json 格式的文件,右键文件后打开方式选择【记事本】,在【文件】处点击【另存为】,修改文件名为“QuickMath”,【编码】处选择“UTF-8”后点击【保存】(如图 3-8-1-68)。

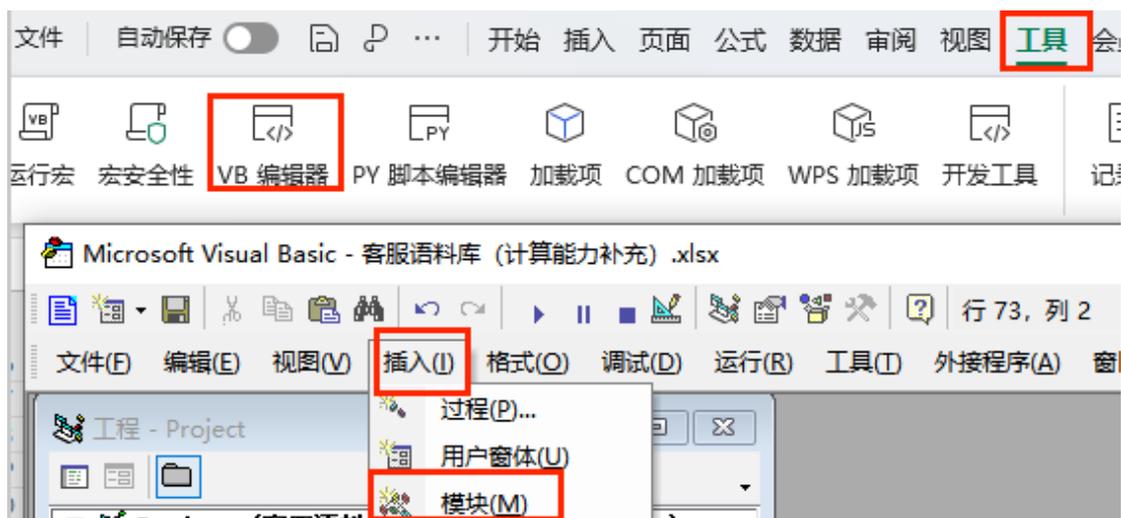


图 3-8-1-66 插入模块

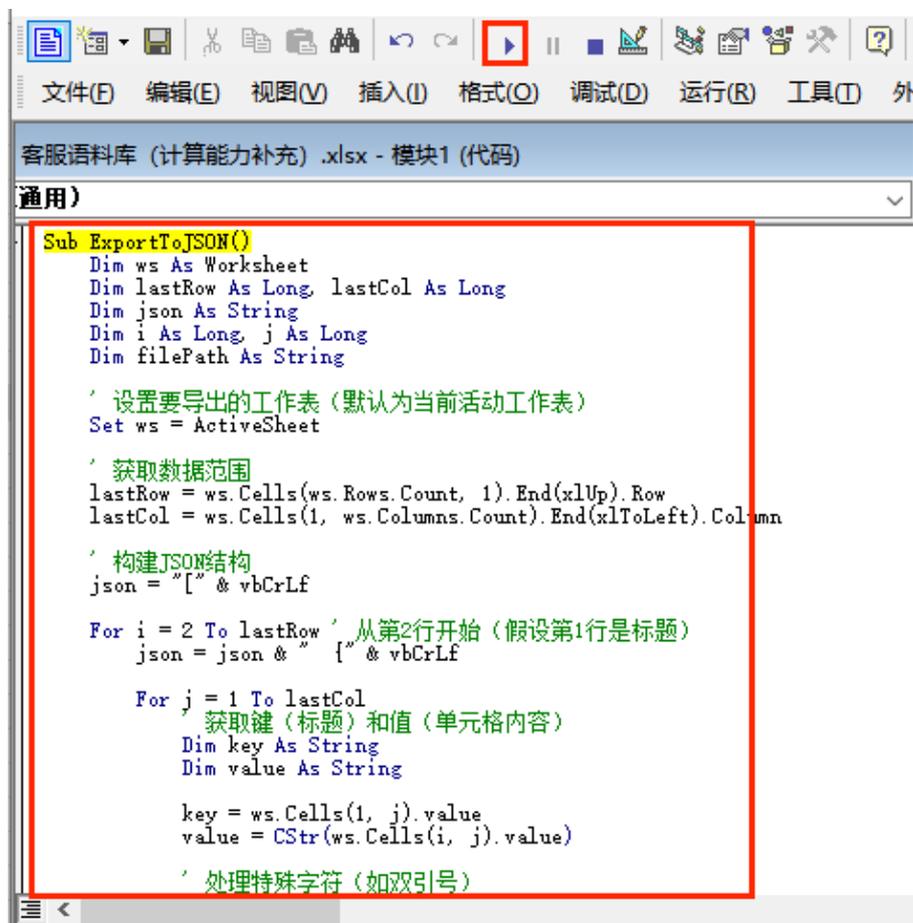


图 3-8-1-67 运行代码

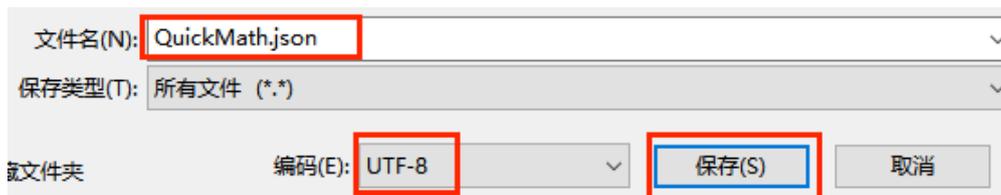


图 3-8-1-68 修改文件名称和格式

(3) 方法二是利用 Python 转换数据，先在【文件列表】处上传《客服语料库（计算能力补充）》（如图 3-8-1-69）。根据任务描述下的要求“转换为 json 文件、编码格式为 utf-8”，在【技术需求转化】处对应关键词填写参数，其中上传的数据文件为“客服语料库（计算能力补充）.xlsx”，输出的 json 文件名为“QucikMath”，json 文件编码格式为“utf-8”（如图 3-8-1-70）。参数填写后点击【执行并显示结果】，传入参数数据，执行数据代码，执行完成后可查看输出结果，点击文件右侧的【下载】可下载相应压缩包（如图 3-8-1-71），解压后即可获取转换后的数据文件。



图 3-8-1-69 文件上传

json格式转换 (2.00分)



技术需求转化

显示答案

关键词	参数
上传的数据文件	客服语料库 (计算能力补充) .xlsx
输出的json文件名	QuickMath
json文件编码格式	utf-8

图 3-8-1-70 参数填写



执行并显示结果

执行时间: 2025-10-28 14:44:25.056

执行状态: ● 运行结束

执行结果:

输出:
转换完成

运行结束

其他:

(1) 文件: 客服语料库 (计算能力补充) .xlsx 下载

(2) 文件: QuickMath.zip 下载

图 3-8-1-71 数据下载

步骤三 进行模型训练

(1) 根据任务描述下提供的编写电商客服的 AI 模型表格，在【进行模型训练】下方，选择对应的内容，具体为基础模型处选择“qwen2.5-7B-instruct”，微调方法为“lora”，数据集选择步骤二中转换的 json 文件“QuickMath.json”，数据集训练轮次修改为 3，学习率修改为 0.0005，批处理大小为 2，LORA 秩修改为 16，LORA 缩放系数为 32，LORA 随机丢弃为 0，其余表格未呈现的内容无需修改，点击【开始】，如图 3-8-1-72。

进行模型训练

基础模型: qwen2.5-7B-instruct | **微调方法:** lora

数据集: QuickMath.json 预览

学习率: 0.0005 | **训练轮数:** 3 | **最大梯度范数:** 1 | **最大样本数:** 100000 | **计算类型:** bf16

截断长度: 2048 | **批处理大小:** 2 | **梯度累积:** 8 | **验证集比例:** 0 | **学习率调节器:** cosine

LORA 参数设置:

LORA 秩: 16 | LORA 缩放系数: 32 | LORA 随机丢弃: 0

启用 lora | 启用 DoRA | 启用 PISA

开始

图 3-8-1-72 模型内容填写

(2) 待微调完成后，输出模型训练结果，如图 3-8-1-73。

微调模型名称: ECqwen7.22-QuickMath

模型训练结果:

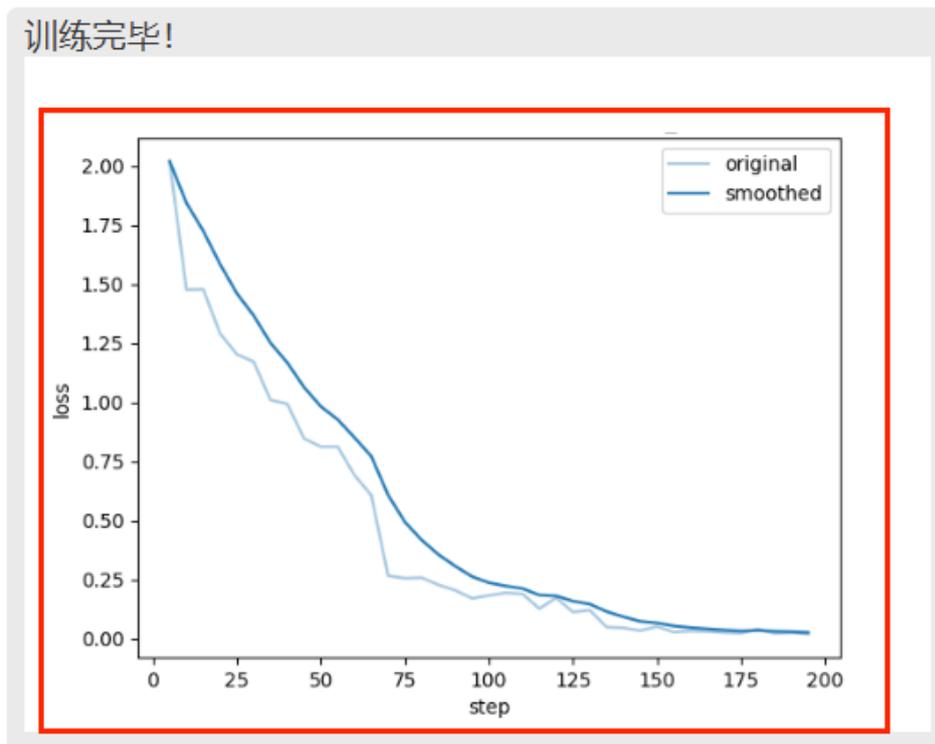


图 3-8-1-73 模型训练结果输出

步骤四 验证训练模型

(1) 根据任务描述下的要求, 调整最大生成长度为 1024, Top-p 采样值为 0.7, 温度系数为 0.95, 加载模型选择“ECqwen7.22-QuickMath”(如图 3-8-1-74)。

验证训练模型

对话

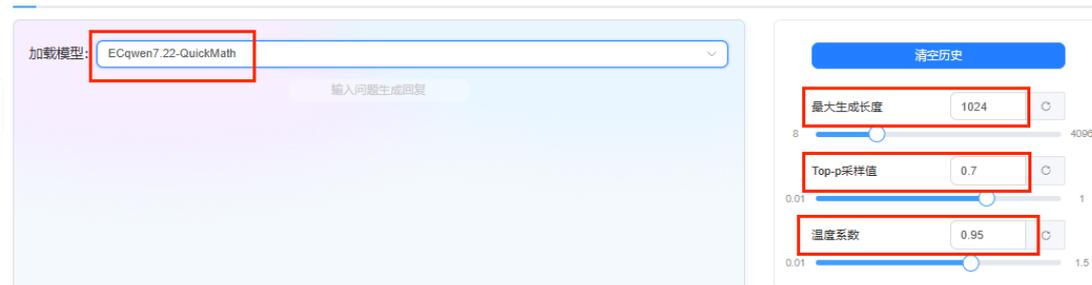


图 3-8-1-74 模型加载与数值修改

(2) 根据任务描述下的步骤进行训练模型验证, 首先针对核心测试术语设置计算相关问题, 以评估模型的计算能力, 其次从“记忆力、理解力和推理力”三个关键维度设计验证测试, 再对模型验证打分, 最后邀请外部验证与综合评估。此处以任务描述下的示例进行具体操作, 核心测试术语为“价保周期”。

(3) 先提问新增的微调语料库内容“我 2 月 11 日购买的, 今天是 2 月 28

日，我发现商品降价了，我要求价保！”，点击【清空历史】（如图 3-8-1-75）之后更换数字，从能价保和不能价保两个结果分别问出一个问题，以查看模型的计算能力（如图 3-8-1-76），可以看到，经过语料库的调整和模型的重新训练，模型已经可以对日期类型的价保问题进行回答。

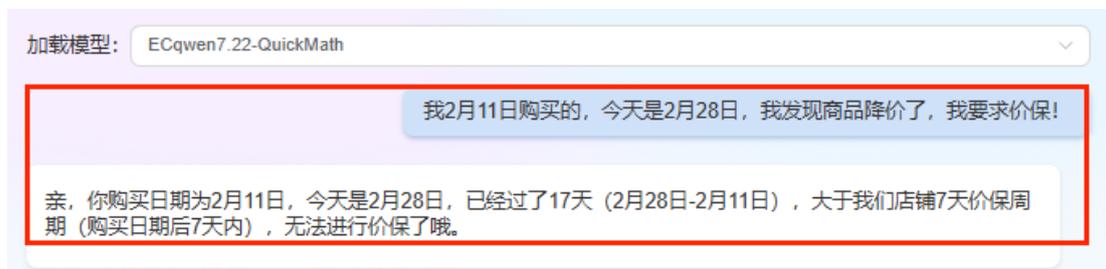


图 3-8-1-75 原始描述

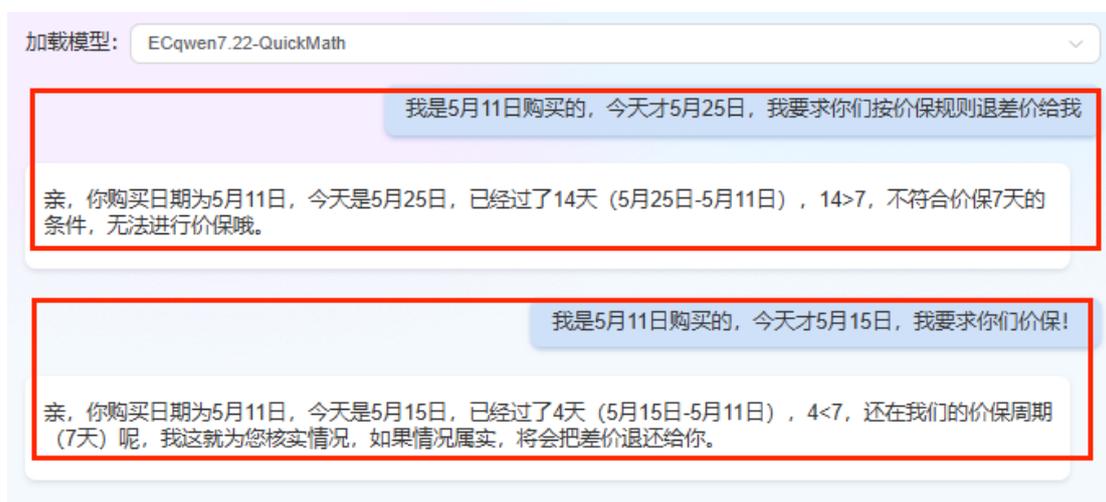


图 3-8-1-76 新增计算

步骤五 模型训练报告

（1）点击【答题卡下载】下的“训练报告.xlsx”以下载分析报告模板（如图 3-8-1-77）。



图 3-8-1-77 模板下载

（2）根据步骤三中模型参数填写，完成《训练报告》中的“AI 模型参数信息”表格内容填写（如图 3-8-1-78），并根据自己使用情况与调查他人的使用

情况，结合表格中验证内容及评分标准，在“模型验证得分”表格内打分，如图 3-8-1-79。在子表“统计和可视化分析”中可直观呈现模型各维度得分情况。完成后点击保存上传至系统。

AI模型参数信息	
使用语料	QuickMath.json
训练轮次	3
学习率	0.0005
其他参数	lora秩: 16; lora缩放系数: 32; lora随机丢弃: 0
训练得分	87.5

图 3-8-1-78 AI 模型参数信息填写

模型验证得分				
验证维度	验证内容	分数设置	自评得分	他评得分
			完全满足验证: 满分; 大部分满足验证: 扣1分; 部分满足验证: 扣2分; 少量满足验证: 扣3分; 完全不满足验证: 0	完全满足验证: 满分; 大部分满足验证: 扣1分; 部分满足验证: 扣2分; 少量满足验证: 扣3分; 完全不满足验证: 0
记忆力维度	1. 模型能够根据消费者咨询输出客服回复	5	4	4
	2. 模型输出的客服回复准确	5	4	3
	3. 模型在多次询问相同的消费者咨询后, 不存在记忆衰退或者输出不稳定的问题	10	9	8
理解力维度	1. 模型可以根据不同描述但本质相同的业务场景输出正确的客服回复。	10	10	9
	2. 模型可以处理包含错误或不完整信息的消费者咨询	12	10	9
	3. 模型在面临特点领域的专有名称替换为其他内容的业务场景时, 能输出正确的客服回复。	12	10	10
推理力维度	1. 模型在高重合度场景测试中, 能准确地应用已学习到的知识, 输出正确的客服回复。	12	12	10
	2. 模型在低重合度场景测试中, 能准确地应用已学习到的知识, 输出正确的客服回复。	16	16	15

图 3-8-1-79 模型验证得分填写

实践案例 民生审计与社会保障

任务一 收集资料及数据

任务二 执行询问程序

任务三 高值耗材收费分析

任务四 低值耗材收费分析

任务五 诊疗收费项目异常

任务六 撰写审计报告

实践案例 自然资源与生态环境审计

任务一 环境资源审计介绍

任务二 审计前准备工作

任务三 审计程序—内部数据库核查

任务四 林地识别

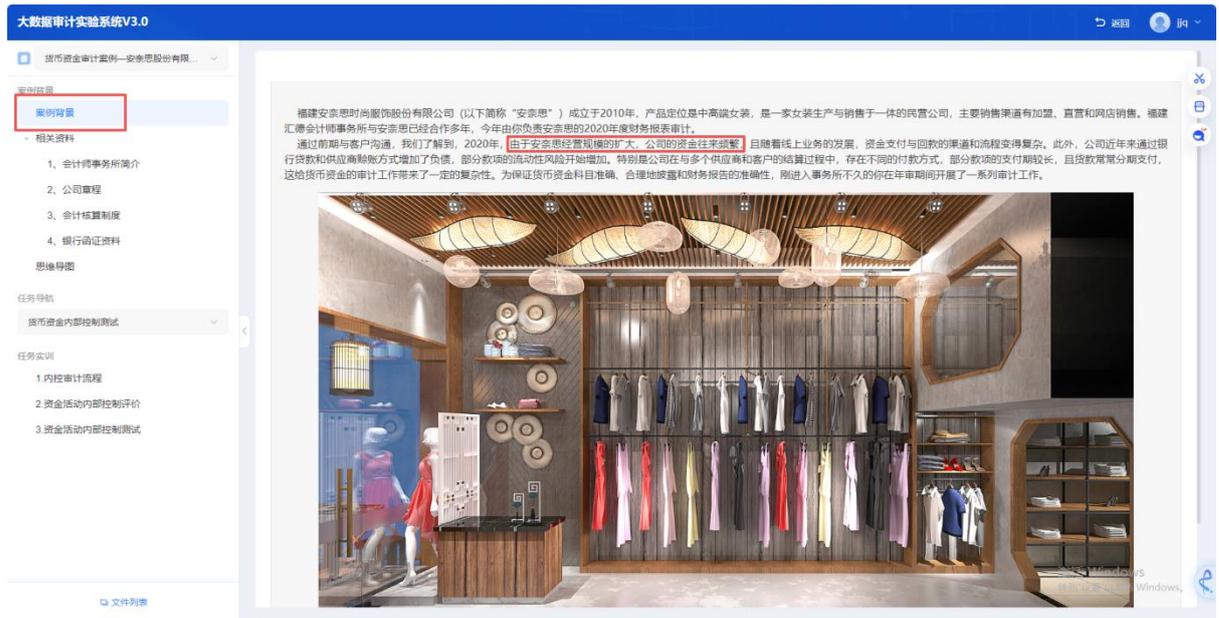
任务五 审计程序—空间数据收集与核查

实践案例 数智审计与数字治理

第 1 课时 步骤 1: 了解被审计单位及其环境

1.1 案例背景

点击行业状况，根据场景对话和相关资料回答问题，如图 2-1-1 所示。



1.2 事务所简介

点击相关资料查阅会计师事务所简介，查阅有关审计事项的确定。



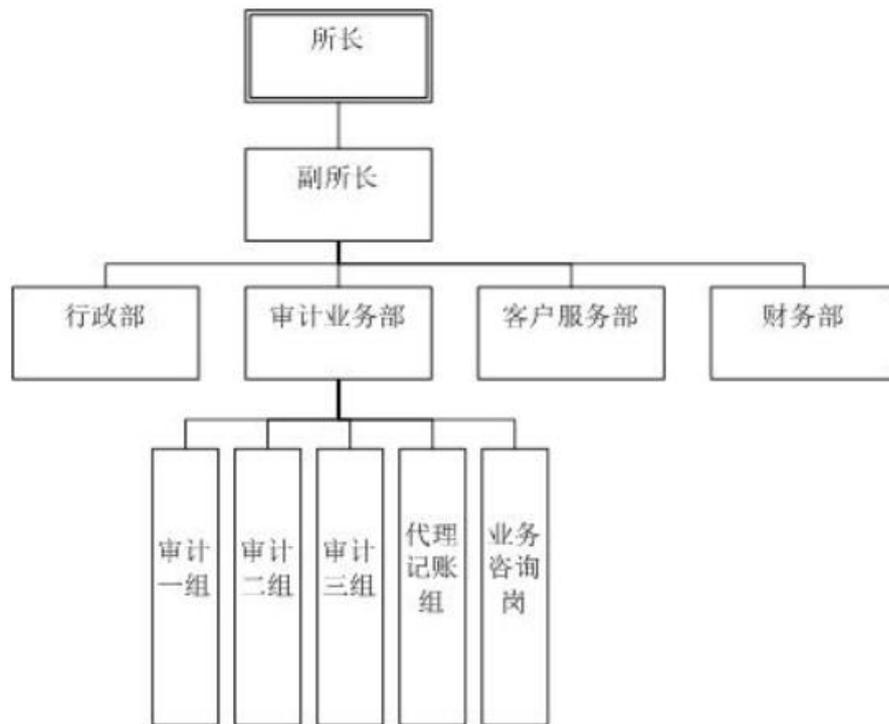
福建汇德会计师事务所

一、基本情况

福建汇德会计师事务所是由财政部门批准,经福州市工商行政管理局登记注册的会计师事务所,成立于 2004 年 6 月 5 日,专门从事审计、验资、办理企业合并、分立、清算事项中的审计事务、财务会计顾问、税务咨询及代理服务等业务汇德会计师事务所现有注册会计师 16 名,专业咨询顾问等财税专业人员 29 名,具备财务、会计、审计、税务、管理咨询等方面知识和经验。现有长期合作客户约 200 余家。拥有可以进行基本建设施工预决(结)算审计验证;法律、行政法规规定的其它审计业务等长期的合作伙伴和专业人才资源。

公司电话:0591-85510532, 公司传真:0591-82763846, 公司邮编:350001 地址:福州市鼓楼区华大街道华林路 166 号

二、公司架构



三、竞争优势

·专业强

拥有一支专业、经验丰富的队伍，其中注册会计师 16 人，其他专业人员 29 名。拥有强大的数据资源，积累了数千家企业和项目的资料和案例。

·信誉好

以科学严谨的工作作风、高效优质的服务，赢得了广大客户和相关政府机构的支持与信任，拥有稳定、广泛的客户群，并与各大企业、机构保持良好的合作关系。

·领域广

在积极开展审计业务的同时，提供一系列的业务相关服务，覆盖面较广。

四、业务范围

（一）审计业务范围：

- 1.财务报表审计：是指对企业资产负债表、损益表、现金流量表、会计报表附注及相关附表所进行的审计，是最常规的审计业务。通常提供给公司的股东、上级单位，以及政府主管部门。
- 2.清算审计；
- 3.离任审计；
- 4.验资；
- 5.代理记账；
- 6.财务顾问
- 7.特殊目的审计。

（二）客户服务业务：

- 1.专业培训；
- 2.代理办理相关业务。

五、收费标准

根据福州市物价局文件通知精神，本会计师事务所收费标准如下：

	序号	服务项目名称	计价单位	50万元以下	50-100万元	100-500万元	500-1000万元	1000-5000万元	
				收费标准					
计 件 收 费	1	财务报表审计	年	3000元	0.20%	0.09%	0.07%	0.05%	
	2	资本验证	次	货币出资收费标准	1500元	0.15%	0.04%	0.03%	2.50%
				其他出资收费标准	按货币出资方式收费标准的120%计收				
	3	合并、分立、清算审计		按年度财务报表审计标准150%计收，再讲分年度收费额累加计算；超过三年的年度可以按照年度财务报表审计标准120%计收。					
	4	经济责任审计		按年度财务报表审计标准150%计收，再讲分年度收费额累加计算；超过三年的年度可以按照年度财务报表审计标准120%计收。					
	5	清产核资		按财务报表审计收费标准的2倍计收。					
	6	外汇收支审核		按计时收费，不低于1000元。					
7	特殊目的审计		参照4、5收费标准计收或按计时收费。						
计 时 收 费	初级助理：300元/小时；助理：600元/小时；注册会计师：1000元/小时；项目经理：1500元/小时；部门经理：2000元/小时；合伙人（主任）会计师：3000元/小时								

备注：

- 1.财务报表审计按被审计单位资产总额与销售收入孰高分档累进计费；
- 2.资本验证按实收资本分档累进计费。如同时存在货币出资及其他出资情况，可分别计算后再相加计费；
- 3.表内的几件收费标准可下浮不超过30%。
- 4.表内的计时收费标准可上下浮40%。证券期货相关业务及大型金融企业鉴证业务可上浮60%。

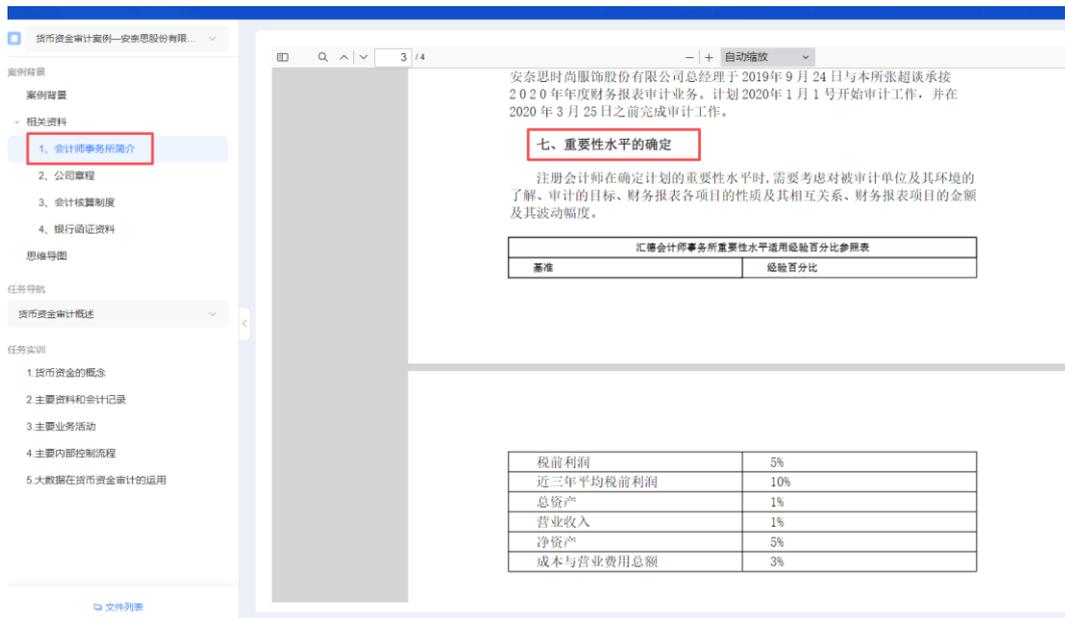
六、业务介绍

福建安奈思时尚服饰股份有限公司主营业务为生产并销售中高端女装。福建安奈思时尚服饰股份有限公司总经理于2019年9月24日与本所张超谈承接2020年年度财务报表审计业务。计划2020年1月1日开始审计工作，并在2020年3月25日之前完成审计工作。

七、重要性水平的确定

注册会计师在确定计划的重要性水平时，需要考虑对被审计单位及其环境的了解、审计的目标、财务报表各项目的性质及其相互关系、财务报表项目的金额及其波动幅度。

基准	经验百分比
税前利润	5%
近三年平均税前利润	10%
总资产	1%
营业收入	1%
净资产	5%
成本与营业费用总额	3%



1.3 安奈思公司章程 查阅公司章程如下：



- (一) 以其认缴的出资额为限对公司承担责任；
- (二) 应当按期足额缴纳本章程载明的各自所认缴的出资额；以货币出资的，应当将货币出资足额存入公司在银行开设的账户；以非货币财产出资的，应当依法办理其财产权转移到公司名下的手续；
- (三) 遵守公司章程，保守公司秘密；
- (四) 支持公司的经营管理，促进公司业务发展；
- (五) 不得抽逃出资；
- (六) 不得滥用股东权利损害公司或者其他股东的利益；
- (七) 不得滥用公司法人独立地位和股东有限责任损害公司债权人的利益。

第四章 股东的出资额、出资时间和出资方式

第十五条 股东的出资额、出资时间和出资方式：

1、股东姓名（名称）：朱家文，认缴出资 500 万元，在 2015 年 01 月 10 日前缴足，其中，以货币出资全额 500 万元。

2、股东姓名（名称）：刘希希，认缴出资 300 万元，在 2015 年 01 月 10 日前缴足，其中，以货币出资全额 300 万元。

3、股东姓名（名称）：王军，认缴出资 200 万元，在 2015 年 01 月 10 日前缴足，其中，以货币出资全额 200 万元。

第十六条 股东以非货币财产出资的，对出资的非货币财产须评估作价，核实财产，不得高估或者低估作价。法律、行政法规对评估作价有具体规定的，从其规定。

第十七条 股东应当以自己的名义出资。

第十八条 股东的出资期限不得超过本章程规定的公司营业期限。

第十九条 股东不按照本章程规定缴纳出资的，除应当向公司足额缴纳外，还应当向已按期足额缴纳出资的股东承担违约责任。

第二十条 公司成立后，发现作为设立公司出资的非货币财产的实际价额显著低于公司章程所定价额的，应当由交付该出资的股东补足其差额；公司设立时的其他股东承担连带责任。

第二十一条 公司发生债务纠纷或者依法解散清算时，如资不抵债，未缴足出资的股东应先缴足出资。

第五章 公司的股权转让

第二十二条 股东之间可以相互转让其全部或者部分股权。

股东向股东以外的人转让股权，应当经其他股东过半数同意。股东应就其股权转让事项书面通知其他股东征求同意，其他股东自接到书面通知之日起满三十日未答复的，视为同意转让。其他股东半数以上不同意转让的，不同意的股东应当购买该转让的股权；不购买的，视为同意转让。经股东同意转让的股权，在同等条件下，其他股东有优先购买权。两个以上股东主张行使优先购买权的，协商确定各自的购买比例；协商不成的，按照转让时各自的实缴出资比例行使优先购买权。本条第二款规定的“视为同意转让”，可由公司出具书面证明。

第二十三条 人民法院依照法律规定的强制执行程序转让股东的股权时，其他股东在同等条件下有优先购买权，其他股东自人民法院通知之日起满二十日不行使优先购买权的，视为放弃优先购买权。

第二十四条 转让股权后，公司应当注销原股东的出资证明书，向新股东签发出资证明书，并相应修改本章程和股东名册中有关股东及其出资的记载。

第二十五条 股东可以依照《公司法》的规定，请求公司按照合理的价格收购其股权。股东与公司不能达成股权收购协议的，股东可以依法向人民法院提起诉讼。公司收购本公司股权后，应当办理减资登记。

第二十六条 股东未履行或者未全面履行出资义务即转让股权的，受让人应当承继转让人的出资义务。

第二十七条 自然人股东死亡后，其合法继承人可以继承股东资格。股东未履行或者未全面履行出资义务的，继承人应当承继股东的出资义务。

第六章 公司的法定代表人

第二十八条 公司法定代表人由执行董事担任。

第二十九条 法定代表人的职权：

（一）法定代表人是法定代表公司行使职权的签字人。

（二）法定代表人在法律、行政法规以及本章程规定的职权范围内行使职权，代表公司参加民事活动，对企业的生产经营和管理全面负责。

（三）公司法定代表人可以委托他人代行职权，委托他人代行职权时，应当出具《授权委托书》。法律、行政法规规定必须由法定代表人行使的职权，不得委托他人代行。

第三十条 法定代表人应当遵守法律、行政法规以及本章程的规定，不得滥用职权，不得作出违背公司股东会、董事会决议的行为，不得违反对公司的忠实义务和勤勉义务。法定代表人违反上述规定，损害公司或者股东利益的，应当承担相应的责任。

第三十一条 法定代表人出现下列情形的，应当解除其职务，重新产生符合法律、行政法规和本章程规定的任职资格的法定代表人：

- （一）法定代表人有法律、行政法规或者国务院决定规定不得担任法定代表人的情形的；
- （二）法定代表人由执行董事或者经理担任，但其丧失执行董事或者经理资格的；
- （三）正在被执行刑罚或者正在被执行刑事强制措施，无法履行法定代表人职责的；
- （四）正在被公安机关或者国家安全机关通缉的；
- （五）其他导致法定代表人无法履行职责的法定情形。

第七章 公司的组织机构及其产生办法、职权、议事规则

第三十二条 公司股东会由全体股东组成，股东会是公司的最高权力机构。

第三十三条 股东会行使下列职权：

- （一）决定公司的经营方针和投资计划；
- （二）选举和更换非由职工代表担任的执行董事、监事，决定有关执行董事、监事的报酬事项；
- （三）审议批准执行董事的报告；
- （四）审议批准监事的报告；
- （五）审议批准公司的年度财务预算方案、决算方案；
- （六）审议批准公司的利润分配方案和弥补亏损方案；
- （七）对公司增加或者减少注册资本作出决议；
- （八）对发行公司债券作出决议；
- （九）对公司合并、分立、解散、清算或者变更公司形式作出决议；
- （十）修改公司章程。

上述事项股东以书面形式一致表示同意的，可以不召开股东会会议，直接作出决定，并由全体股东在决定文件上签名、盖章。

第三十四条 股东会会议分为定期会议和临时会议。定期会议在每会计年度期末召开一次。代表十分之一以上表决权的股东，执行董事，监事提议召开有限责任公司股东会临时会议的，应当召开临时会议。

第三十五条 股东会会议由执行董事召集并主持；执行董事不能履行或者不履行召集股东会会议职责的，由监事召集并主持；监事不召集的，代表十分之一以上表决权的股东可以自行召集并主持。

第三十六条 召开股东会会议，应当于会议召开 1-5 日前通知全体股东；但是，全体股东另有约定的除外。会议通知的内容应当包括：股东会召开的时间、地点、议题等。

第三十七条 股东会应当对股东会会议通知情况、股东出席情况、表决情况以及所议事项的决定作成会议记录，出席会议的股东应当在会议记录上签名。

第三十八条 股东按照认缴出资比例行使表决权。

第三十九条 股东会会议作出修改公司章程、增加或者减少注册资本的决议，以及公司合并、分立、解散或者变更公司形式的决议，必须经代表三分之二以上表决权的股东通过。股东会会议作出的其他决议，应当经代表过半数表决权的股东通过。股东会会议作出公司合并、分立以及减少注册资本决议的，公司应当自作出决议之日起十日内通知债权人，并于三十日内在报纸上公告。

第四十条 公司设执行董事一人，对公司股东会负责，由股东会选举产生。

第四十一条 执行董事每届任期 3 年。执行董事任期届满，连选可以连任。
执行董事任期届满未及时改选，或者执行董事在任期内辞职的，在改选出的执行董事就任前，原执行董事仍应当依照法律、行政法规和公司章程的规定，履行执行董事职务。

第四十二条 执行董事对股东会负责，行使下列职权：

- (一) 负责召集和主持股东会，并向股东会报告工作；
- (二) 执行股东会的决议；
- (三) 决定公司的经营计划和投资方案；
- (四) 制订公司的年度财务预算方案、决算方案；
- (五) 制订公司的利润分配方案和弥补亏损方案；
- (六) 制订公司的增加或减少注册资本的方案；
- (七) 制订公司合并、分立、变更公司形式、解散的方案；
- (八) 决定公司内部管理机构的设置；
- (九) 聘任或者解聘公司经理、财务负责人，决定其报酬事项；
- (十) 制定公司的基本管理制度。

第四十三条 公司设经理，由执行董事聘任或者解聘。经理对执行董事负责，行使下列职权：

- (一) 主持公司的生产经营管理工作，组织实施执行董事的决定；
- (二) 组织实施公司年度经营计划和投资方案；
- (三) 拟订公司内部管理机构设置方案；
- (四) 拟订公司的基本管理制度；
- (五) 制定公司的具体规章；
- (六) 提请聘任或者解聘公司副经理、财务负责人；
- (七) 决定聘任或者解聘除应由执行董事决定聘任或者解聘以外的负责管理人员；
- (八) 执行董事授予的其他职权。

第四十四条 公司设监事 1 人，监事由股东会选举产生。

第四十五条 监事每届任期三年。监事任期届满，连选可以连任。

第四十六条 执行董事、高级管理人员以及财务负责人不得兼任监事。

第四十七条 监事行使下列职权：

- (一) 检查公司财务；
- (二) 对执行董事、高级管理人员执行公司职务的行为进行监督，对违反法律、行政法规、公司章程或者股东会决议的执行董事、高级管理人员提出罢免的建议；
- (三) 当执行董事、高级管理人员的行为损害公司的利益时，要求执行董事、高级管理人员予以纠正；
- (四) 提议召开临时股东会会议，在执行董事不履行本章程规定的召集和主持股东会会议职责时召集和主持股东会会议；
- (五) 向股东会会议提出提案；
- (六) 依照《公司法》第一百五十一条的规定，对执行董事、高级管理人员提起诉讼；
- (七) 对执行董事决定的事项提出质询或者建议。

第四十八条 有下列情形之一的，不得担任公司的执行董事、监事、高级管理人员：

- (一) 无民事行为能力或者限制民事行为能力；
- (二) 因贪污、贿赂、侵占财产、挪用财产或者破坏社会主义市场经济秩序，被判处刑罚，执行期满未逾五年，或者因犯罪被剥夺政治权利，执行期满未逾五年；
- (三) 担任破产清算的公司、企业的董事或者厂长、经理，对该公司、企业的破产负有个人责任的，自该公司、企业破产清算完结之日起未逾三年；

(四)担任因违法被吊销营业执照、责令关闭的公司、企业的法定代表人，并负有个人责任的，自该公司、企业被吊销营业执照之日起未逾三年；

(五)个人所负数额较大的债务到期未清偿。公司违反前款规定选举执行董事、监事或者聘任高级管理人员的，该选举或者聘任无效。执行董事、监事、高级管理人员在任职期间出现本条第一款所列情形的，公司应当解除其职务。

第四十九条 执行董事、监事、高级管理人员应当遵守法律、行政法规和公司章程的规定，对公司负有忠实义务和勤勉义务：

(一)谨慎、认真、勤勉地行使股东、公司赋予的权利，以保证公司的商业行为符合国家法律、行政法规以及国家各项政策的要求，商业活动符合公司章程规定的业务范围；

(二)公平对待所有股东；

(三)及时了解公司业务经营管理状况；

(四)对公司定期报告签署书面确认意见；

(五)如实向监事提供有关情况 and 资料，不得妨碍监事行使职权；

(六)保证公司所披露的信息真实、准确、完整；

(七)法律、行政法规和公司章程规定的其他义务。

第五十条 执行董事、高级管理人员不得有下列行为：

(一)挪用公司资金；

(二)将公司资金以其个人名义或者以其他个人名义开立账户存储；

(三)违反公司章程的规定，未经股东会同意，将公司资金借贷给他人或者以公司财产为他人提供担保；

(四)违反公司章程的规定或者未经股东会同意，与本公司订立合同或者进行交易；

(五)未经股东会同意，利用职务便利为自己或者他人谋取属于公司的商业机会，自营或者为他人经营与所任职公司同类的业务；

(六)接受他人与公司交易的佣金归为己有；

(七)擅自披露公司秘密；

(八)违反对公司忠实义务的其他行为。

执行董事、高级管理人员违反前款规定所得的收入应当归公司所有。

第五十一条 执行董事、监事、高级管理人员执行公司职务时违反法律、行政法规或者公司章程的规定，给公司造成损失的，应当承担赔偿责任。

第五十二条 公司依照法律、行政法规和国家财政主管部门的规定建立财务、会计制度。公司应当在每一个会计年度终了时制作财务会计报告，并依法经会计师事务所审计。公司聘用、解聘承办公司审计业务的会计师事务所，应当由股东会决定。公司依法律规定在分配当年税后利润时，提取利润的百分之十列入公司法定公积金，法定公积金累计额为公司注册资本的百分之五十以上的，可不再提取。公司从税后利润中提取法定公积金后，经股东会决议，可以从税后利润中提取任意公积金。公司弥补亏损和提取公积金后所余税后利润，按照股东的实缴出资比例分配。不按实缴的出资比例分取红利的，应经全体股东同意。公司的公积金用于弥补亏损，扩大公司生产经营或者转为增加公司资本。公司除法定的会计账册外，不得另立会计账册。对公司的资产，不得以任何个人名义开立账户存储。任何个人不得挪用公司资金或者将公司资金借贷给他人；不得侵占公司的财产。

第五十三条 公司应当在下一会计年度开始之后 2 个月前将公司财务会计报告送交各股东。

第五十四条 公司的行政部门负责保管公司的公章、营业执照。

第八章 公司的解散、清算

第五十五条 公司因下列原因解散：

- (一) 公司章程规定的营业期限届满；
- (二) 股东会决议解散；
- (三) 因公司合并或者分立需要解散；
- (四) 依法被吊销营业执照、责令关闭或者被撤销；
- (五) 人民法院依公司法第一百八十二条的规定予以解散。

第五十六条 公司出现除上一条第（三）项以外的解散事由时，应当在解散事由出现之日起十五日内成立清算组，开始清算。清算组由股东组成，其中非自然人股东应当指定人员行使相应权利。

第五十七条 清算组在清算期间行使下列职权：

- (一) 清理公司财产，分别编制资产负债表和财产清单；
- (二) 通知、公告债权人；
- (三) 处理与清算有关的公司未了结的业务；
- (四) 清缴所欠税款以及清算过程中产生的税款；
- (五) 清理债权、债务；
- (六) 处理公司清偿债务后的剩余财产；
- (七) 代表公司参与民事诉讼活动。

第五十八条 清算组应当自成立之日起十日内通知债权人，并于六十日内在报纸上进行公告。

第五十九条 清算组在清算公司财产、编制资产负债表和财产清单后，应当制定清算方案，并报股东会或者人民法院确认。公司财产在分别支付清算费用、职工工资、社会保险费用和法定补偿金，缴纳所欠税款，清偿公司债务后的剩余财产，按照股东的实缴出资比例分配。清算期间，公司存续，但不得开展与清算无关的经营活动。公司财产在未依照前款规定清偿前，不得分配给股东。

第六十条 公司清算结束后，清算组应当制作清算报告，报股东会或者人民法院确认，并向公司登记机关申请注销公司登记，公告公司终止。

第九章 公司的其他规定

第六十一条 股东、执行董事、监事应当把联系方式（包括通信地址、电话、电子邮箱等）报公司置备，发生变动的，应及时报公司予以更新。

第六十二条 本章程涉及的股东会会议，可以采取口头、电子邮件、书面等方式通知。

第六十三条 公司可以向其他企业投资或者为他人提供担保，并由股东会决议。

公司向其他企业投资的，除法律另有规定外，不得成为对所投资企业的债务承担连带责任的出资人。公司为公司股东或者实际控制人提供担保的，必须经股东会决议。前款规定的股东或者受前款规定的实际控制人支配的股东，不得参加前款规定事项的表决。该项表决由出席会议的其他股东所持表决权的过半数通过。

第六十四条 公司应当通过企业信用信息公示系统向社会公示章程、年度报告、股东缴纳出资情况等信息，具体公示内容按国家相关规定执行。

第六十五条 本章程于 2015 年 01 月 08 日订立。

1.4 公司会计核算制度



查阅公司会计核算制度如下：

福州歌莱斯服饰股份有限公司 财务会计基本核算制度 第一章 总则

第一章 总则

第一条 规范公司财务会计行为，依据《中华人民共和国会计法》《企业财务通则》和《企业会计准则》等有关法律法规，结合本公司的实际情况，制定本制度。

第二条 公司设立财务部，负责财务会计方面的日常工作，财务经理负责公司财务收支计划的执行，主管公司的会计核算和资金管理工作，主持制定公司的财务规章制度，规范财务管理实施细则，统一公司会计核算办法，配合总经理协调金融、财政、税务等方面的外部关系。

第三条 公司会计机构和会计人员应当依照会计法律法规的规定，按照真实、完整的要求，进行会计核算，实行会计监督，并向公司董事会、总经理定期汇报会计工作情况，报送财务会计报告和其他必要资料。

第四条 公司会计年度采用公历年度，会计核算以人民币为记账本位币，会计记账采用借贷记账法。

第五条 公司审核原始凭证、填制会计凭证、登记会计账簿、编报财务会计报告和管理会计档案，应当遵循有关法律法规和本制度的要求。

第二章 会计管理体制

第六条 公司董事长、财务经理根据岗位职责分工负责本公司的财务会计工作，并对会计工作和会计资料的真实性和完整性负责，财务部负责具体的日常会计工作。

第七条 公司财务负责人为财务经理，依法行使如下会计管理职责：

- （一）依法设立会计机构、配备会计人员，并支持他们的工作，保障会计人员依法履行职责；
- （二）正确执行相关法律法规和规章制度，纠正违反财务会计制度的行为；
- （三）最终审批公司所有财务支出事项；
- （四）对会计工作情况及会计资料的真实性、完整性负责，签署对外公布的财务会计报告；
- （五）法律法规规定或上级主管部门授予的其他职责。

第八条 会计机构应当按照有关法律法规的规定、上级主管部门及公司主要负责人的要求，强化会计岗位责任制度、内部控制制度和会计稽核制度。

第九条 公司领导、部门经理和其他员工应当按照本制度的规定，在各自的权限内真实、完整地审核和提供各种经济业务的原始资料。公司有关人员在对所管业务的财务收支进行具体的预算管理和目标控制时，对有关的原始资料负有记录、保管和及时进行传递的责任。

第三章 会计机构和会计人员

第十条 公司财务部为依法设立的会计机构，并根据工作需要配备相应的会计人员。财务部设财务经理、会计主管、出纳、会计四个岗位。财务部直接对总经理负责。

第四章 主要会计政策

第十一条 以每年元月一日至十二月三十一日为一个会计年度。

第十二条 记账原则和方法：以权责发生制为原则，实行借贷记账法。

第十三条 企业的会计核算以人民币为记账本位币。

第十四条 公司的固定资产按直线法计提折旧。

第十五条 按照有关规定计提坏账准备，存货跌价准备和其他资产减值准备。

第五章 会计科目的设置

第十六条 公司执行财政部颁发的《企业会计制度》《企业会计准则》和《公司经营企业会计制度》，并根据具体经营特点选用总账会计科目，其他明细科目根据编制报表要求，便于理解，方便管理和符合公司经济活动分析的需要合理设置。

第十七条 财务部在具体账务处理中必须正确使用确定的会计科目，在实际工作中可根据实际情况适当增设相关会计科目。

第六章 会计原始凭证与记账凭证

第十八条 原始凭证是指涉及确定的经济业务并需要进行会计处理的文件、合同、发票收据等各种书面记录，包括文件、合同、签单、审批件、发票账单、结算单、验收单、工资表、收款收据、支票存根、汇款回单等。各项财务收支业务需要的原始记录，应当具备规定的内容，充分证明经济业务发生的性质、原因、时间、数量和金额。所有原始记录必须做到真实、完整。

第十九条 各种原始凭证必须规范，符合会计核算的要求，其应具备如下一些内容：

- (一) 名称：原始凭证必须有特定的名称，文件、合同、签呈、呈批件等必须有标题；
- (二) 日期：一般是指经济业务发生的时间或文件、合同、签呈、呈批件的做成时间；
- (三) 业务内容：业务内容应当清楚，并表明该记录设计的经济性质。
- (四) 数量；
- (五) 单价；
- (六) 金额：各原始凭证上记载的金额应当填写清楚，计算准确，大小写必须一致，格式化的原始凭证上的金额还应当封顶；
- (七) 接受原始凭证的单位名称；
- (八) 出具原始凭证单位的名称：以行政公章、财务专用章标识时，印章必须清晰可辨；
- (九) 有关经办人员签字认可。

第二十条 原始凭证出现错误，必须按规定进行更正。

- (一) 文字出现错误时，必须由开具单位予以更正，并在更正处加盖公章；
- (二) 数字和金额出现错误时，不得在错误凭证上更正，必须由原单位重新开具；
- (三) 如果原始凭证由两个单位共同完成，任何错误均应当由双方共同更正、盖章。

第二十一条 同类经济业务的原始凭证数量较多时，可填制原始凭证汇总表。

第二十二条 原始凭证必须由出具原始凭证单位盖印财务专用章，同时必须由有关经办人员，验收人员和主管人员的签名或盖章，并注明用途。属于公司自制的原始凭证必须按照财务部制定的统一、规范的格式，并由经办人和具有审批权限的相关人员签名或盖章。

第二十三条 公司依照凭证按照记账凭证的支持对原始凭证划分为两种类型：主要凭证和辅助凭证。主要凭证是对记账凭证起主要支持作用的原始凭证。

主要凭证有：

（一）证明款项和有价证券收付业务的凭证，包括证明款项是否支付、有价证券是否交割的资料，以及证明款项（或有价证券）种类、金额价值、性质和权属的资料；

（二）证明财物收发、增减和使用的凭证，包括证明收发、性质和使用原因、来源取向和审批过程的资料，以及证明资产价值、种类、性质和权属的资料；

（三）证明债权债务发生和结算的凭证，包括证明发生原因、金额计算过程和依据资料，以及证明债权债务相对人、清偿责任人的资料；

（四）证明资本、基金增减业务的凭证，包括确定其增减原因、来源去向及审批过程的资料，以及确定所增减资本（基金）种类、金额价值、性质和权属的资料；

（五）证明收入、支出、费用、成本计算业务的凭证，包括记录其计算过程、计算依据和审批过程的资料，以及证明收入（支出、费用、成本）种类、经办人、金额和性质的资料；

（六）证明财务成果计算和处理业务的凭证，包括说明计算（处理）依据和过程的资料，以及说明财务成果性质、金额的资料。

第二十四条 主要凭证必须符合国家法律法规和会计制度的规定，并附在记账凭证的后面。对两张或两张以上记账凭证起主要支持作用的原始凭证，可附在其中一张记账凭证的后面，其他记账凭证后附原始凭证的附件或复印件，但其他记账凭证上必须签注原始凭证原件所在记账凭证的编号。

第二十五条 辅助凭证是对主要凭证起辅助证明作用的原始凭证，包括不直接支持某一张记账凭证或者对多张记账凭证起支持作用的合同、文件、呈批件等。

第二十六条 辅助凭证可不附在记账凭证的后面，但应当单独按照档案管理办法进行分类整理，及时归档保存。未附辅助凭证的记账凭证应当注明该辅助凭证的名称、编号。

第二十七条 对于涉及多笔经济业务或需要多次备查使用的主要凭证和辅助凭证，经办部门和需要查询的部门，应当同时保管相关原始凭证的原件。

第二十八条 会计人员应当认真审核原始记录，并在通过审核记录上签字。

审核的内容包括：

（一）经纪业务是否本单位发生的，内容是否真实、合法、合规；

（二）原始凭证是否具有法律效力，公章是否清晰；

（三）经办人是否经过授权或在授权范围内填制内容；

（四）数量、单价、金额的计算是否正确；

（五）必备内容是否完备齐全。

第二十九条 会计人员在受理原始记录时，对于不真实的，有权拒绝受理，并可将情况报告单位负责人；对不完整的，有权退回，经办人补正后再受理。

第三十条 会计人员应当依据真实、完整的原始记录进行会计处理，并按照本制度的相关规定将原始凭证附在记账凭证后面。

第三十一条 需要进行会计处理的原始凭证按照如下的程序进行传递。

（一）经办人取得原始凭证后，需要提交证明、检查或验收人签字的，交有关人员签字；

（二）有关人员签字后，交会计人员审核；

（三）按有关规定经部门经理、公司负责人签字批准；

（四）会计人员受理原始记录，进行会计处理。

第三十二条 会计人员制作记账凭证，应当以真实、完整的原始凭证为依据。无法取得原始凭证的，应当按照要求自制或者要求相关人员制作原始凭证。

第三十三条 记账凭证应当使用财务会计软件按规定进行填制，不得手工填制。填制记账凭

证必须符合如下要求：

- （一）以本人的身份进入财务会计软件填制凭证的界面；
- （二）必备内容，包括日期、摘要、应借应贷科目及金额等；
- （三）填写要齐全、清楚，并与所附原始凭证及其记载的内容相符；附件注明原始凭证张数，时间按编制当日起填写；填写要正确使用会计科目，保证科目使用与相关业务在定义上的一致；
- （四）借贷的金额要相等；
- （五）按所附原始凭证准确填写与借记或贷记的不同科目相符的摘要。摘要必须清楚，简明扼要；
- （六）购买多种物品应注明主要物品的名称；往来业务凭证，借款和报销业务应具备对方单位名称或个人所属部门及姓名；
- （七）提取现金应注明原因；收支的结转，费用的预提摊销应注明期间。

第三十四条 第四十五条审核人员、出纳人员、总账会计通过财务会计软件审核记账凭证和记账后，应当在输出的纸质记账凭证上盖章或签名。

第三十五条 第四十六条通过财务会计软件输出纸质记账凭证，应当符合《会计基础工作规范》和会计电算化有关工作的规范的要求，并与所附原始凭证一起装订成册。会计凭证的装订应当符合会计档案管理的要求，当月第一本凭证的最前面应当附有会计科目发生额及余额表，最后一本凭证应当附有银行存款余额调节表和银行对账单。

第三十六条 第四十七条会计凭证的归档，保管等事项按照《会计档案管理办法》的规定办理。

第七章 会计账簿的设置

第三十七条 财务部按《会计法》的规定，设置会计账簿，包括总账、明细账、日记账和其他辅助性账簿。

第三十八条 根据电算化的要求，当期所有记账凭证数据和明细账数据都储存于计算机。

第三十九条 明细账根据会计核算系统，结账后数据自动生成资产、负债、权益、损益四大类明细账。账本格式按照会计科目的设置、管理查账的需要，合理编排为借贷余三栏式、成本费用类多栏式、多借一贷式、多贷一借式和多借多贷式。

第四十条 日记账分为银行日记账和现金日记账，不得跨年度使用，账簿不得用银行对账单或其他方式代替。开设若干个银行存款账户时，应按开设的账户分别登记银行日记账，日记账由出纳根据银行收（付）款、现金收（付）款凭证登记，按凭证发生日期、凭证号依次登记，做到日清月结，登记时不得出现银行存款或现金的红字。

第四十一条 各种租借设备、物资、有价证券、应收应付票据、项目核算资料等应设置辅助账。各种账簿必须有统一固定的格式，要做到内容真实完整、数字准确、摘要清楚、登记及时，不漏记、错记或重记。如果发生错误或者隔页、跳行的，应按规定的办法更正，并由会计人员和会计机构负责人在更正处盖章。第四十二条 财务部应按规定定期对会计账簿记录金额与库存金额、货币资

金、有价证券、往来单位或个人等进行核实，做到账证、账账、账表、账实相符，对账工作每年至少一次，如遇人员调动或发生非常事件，也应及时对账。

第四十三条 会计账簿按银行存款日记账、现金日记账、总账、明细账、辅助账分别整理立卷，归档保管。

第四十四条 本公司使用财务会计软件进行会计核算，软件输出的会计账簿应当符合国家统一会计制度的要求，并及时装订。除会计软件输出的正式账簿外，不得设立其他正式账簿。

第四十五条 通过会计软件输出的正式账簿，应当按照封面、扉页、科目发生额及余额表、总账和明细账的顺序进行打印和装订，各科目的总账和明细账按科目编号的顺序打印和装订。

第四十六条 会计账簿应当保持整洁、清晰，除手工登记的备查登记簿外，不得涂改、挖补，手工登记的备查登记簿需要更正的，由登记人员进行更正并盖章，登记人员以外的人员不得进行改正。

第四十七条 通过会计软件输出的年度账簿，应当在扉页上加盖公章，并贴划印花税票。

第四十八条 财务部经理应当在装订的会计账簿封面签章。在装订年度正式账簿时，科目发生及余额表后应当附有当年的年度会计报表。通过会计软件输出的正式账簿，由会计机构按照会计档案管理的规定进行保管。

第四十九条 会计账簿的电子数据应当每月备份一次，每次备份时应当在至少两台不同的计算机硬盘上各保存一份。电子数据在年末结账后，应当备份到单独的电子媒介，按照档案管理的规定妥善保管。

第八章 所有者权益和负债的核算

第五十条 本公司筹集资本必须是货币资金（现金），未经股东大会同意，不得以债权、实物资产、有价证券等投入资金。

第五十一条 经营期内，投资者对其投入的资本，除依法转让外，不得以任何方式抽走。投资者应按照投资比例或者合同、章程规定，分享本公司的收益或承担本公司风险。

第五十二条 投资者以现金投入的资本，应当以实际收到或者存入公司开户银行的金额作为实收资本入账。实际收到或者存入企业开户银行的金额超过公司注册资本中所占份额的部分，计入资本公积。

第五十三条 投资者以非现金资产投入的资本，应在办理实物产权转移手续时，按投资各方确认的价值或合同。协议约定的价值作为实收资本入账。对于投资各方确认的价值超过其在被投资方注册资本中所占份额的部分，应计入资本公积。

第五十四条 实收资本除下列情况外，不得随意变动：

（1）符合增资条件，并经有关部门批准增资的，在实际取得投资者的出资时，登记入账。

（2）企业按法定程序报经批准减少注册资本的，在实际发还投资时登记入账。

第五十五条 资本公积按其形成的类别设置四个明细科目，包括“资本溢价”“接受捐赠非现金资产准备”“外币资本折算差额”“其他资本公积”。资本公积准备项目不能转增资本。

第五十六条 公司将盈余公积分为法定盈余公积，任意盈余公积两部分。法定盈余公积按税后利润的 10%提取，法定盈余公积累计达到注册资本 50%时可不再提取；任意盈余公积由执行董事与总经办决议确定提取比例。

第五十七条 资本公积和盈余公积都可按照投资者的投资比例转为实收资本，转增后留存的盈余公积的数额不得少于注册资本的 25%。

第九章 货币资金的核算

第五十八条 公司的货币资产包括现金、银行存款和其他货币资金。

第五十九条 财务部门必须定期或不定期对货币资金进行盘点、对账，并确保账款相符。

第六十条 货币资金管理必须遵守《货币资金管理制度》的有关规定。

第十章 存货的核算

第六十一条 购入材料物资采用历史成本计量，按实际成本对入库原材料进行计价，实际成本包括存货采购价、运杂费、包装费、仓储费、保险费、运输途中合理损耗等。

第六十二条 月末在途材料按实际成本计价，记入“在途物质”。对于原材料已运到并验入库，但月末发票账单等结算凭证仍未收到，未办理付款的业务，原材料应按合同价格暂估入账，下月初用红字予以冲销。

第六十三条 材料领用采用一次加权平均法进行计价。

第六十四条 产成品按实际成本计价核算，月末结转销售成本时按加权平均法核算。

第六十五条 低值易耗品于领用时一次摊销计入成本、费用。

第六十六条 存货盘点采用永续盘存制，仓库的存货每年盘点两次；生产部门的在产品每月盘点一次。盘存数如果与账面记录不符，应及时查明原因，进行会计处理。

第六十七条 经相关权力人员审批后，对于盘盈的存货，冲减相关费用；盘亏或毁损的存货，在扣除过失人或保险公司的赔款和残料价值后，计入相关费用；由于非常原因造成的存货净损失，计入营业外支出。

第六十八条 存货跌价准备的确认标准和计提方法：当存货遭受毁损、陈旧过时或市价低于其账面成本时，按其可变现净值与账面成本的差额、按存货类别，提取存货跌价准备。

第十一章 固定资产的核算

第六十九条 使用年限在一年（不含一年）以上，单位价值在 2000 元（不含 2000 元）以上，并在使用过程中保持原有物质形态的资产，包括房屋及建筑物、电子设备、运输设备、机器设备、工具、器具。当一项固定资产的某组成部分在使用效能上与流动资产相对独立，并且具有不同使用年限时，应将该组成部分单独确认为固定资产。

第七十条 不属于生产、经营的主要设备的物品，单位价值在 5000 元以下，一次性计入当期损益，只做固定资产数量列管。

第七十一条 不符合固定资产条件的物品，为低值易耗品。低值易耗品应比固定资产管理办法，设立分类登记簿和低值易耗品卡片账，建立登记、保管和定期盘点制度。

第七十二条 财务部门要定期对固定资产、低值易耗品进行全面盘点、清查，每年年终前为定期盘点清查的时间。

第七十三条 所有固定资产和低值易耗品都必须按国家有关规定进行折旧或摊销。

第七十四条 为购建固定资产、无形资产等借款所发生的借款费用，满足资本化条件的应予以资本化，其他借款费用均应于发生当期确认为费用，直接计入当期财务费用。

第七十五条 为购置某项固定资产而借入的专门借款所发生的利息、折价或溢价的摊销、汇兑差额，在同时满足以下三个条件时，计入所购建固定资产的成本：

（1）资产支出（只包括为购建固定资产而以支付现金、转移非现金资产或者承担带息债务形式发生的支出）已经发生；

（2）借款费用已经发生；

（3）为使资产达到预定可使用状态所必要的购建活动已经开始。

第七十六条 对固定资产盘盈、盘亏，由财务部、行政事务部调查核实，弄清情况，提出意见，逐笔报领导审批，财务部门根据规定做账务处理。

第十二章 在建工程的核算

第七十七条 在建工程包括施工前期准备、正在施工中的建筑工程、安装工程、技术改造工程等。在建工程按实际发生的支出入账。所建造的固定资产自达到预定可使用状态之日起，根据工程预算、造价或工程实际成本等，按估计的价值转入固定资产并计提折旧，待办理了竣工决算手续后再调整固定资产入账价值，但已提折旧额不再调整。

第七十八条 在建工程借款所发生的利息支出，在工程达到预定可使用状态前计入工程成本，工程交付使用后计入当期损益。

第七十九条 自建工程按直接材料、直接工资、直接机械施工费以及所分摊的工程管理费和有关的利息支出等计价。

第八十条 出包工程按照应当支付的工程价款以及所分摊的工程管理费和有关的利息支出等计价。

第八十一条 设备安装工程按照所安装设备的原价、工程安装费用、工程试运转支出以及所分摊的工程管理费和有关借款的利息支出等计价。

第十三章 长期和短期投资的核算

第八十二条 公司的证券投资必须符合信用，稳健、效益的原则，确保资产的保值增值。

第八十三条 购入有价证券按有价证券的面值和规定的利率计算应收利息，分期计入损益。

第八十四条 出售经营性证券可以采用先进先出法、加权平均法、移动加权平均法等确定其实际成本。计价方法一经确定，不得随意变更。出售经营性有价证券实际收到的价款与账面成本的差额，计入当期损益。

第八十五条 购入折价或溢价发行的长期债券，实际支付的款项与票面价值的差额应在债券到期前，分期增加或冲减投资收益。

第八十六条 对外证券投资分析得的股利或利润，计入投资收益，并按规定进行税务处理。

第十四章 其他类资产的核算

第八十七条 公司其他类资产包括无形资产、递延资产及其他资产。

第八十八条 无形资产按取得时的实际成本计价。

第八十九条 无形资产自开始使用之日起在有效使用期限内平均摊入成本

第九十条 递延资产是指不能全部计入当年损益，应当在以后年度内分期摊销的费用，包括开办费、以经营租赁方式租入的固定资产技改支出、摊销期超过 1 年的修理费以及摊销期超过 1 年的其他待摊费用等。

第九十一条 开办费是指筹建期间发生的费用，自营业之日起分 5 年分期摊入营业费用。

第十五章 债权债务的核算

第九十二条 对债权债务账户应按单位或个人设立明细账，及时、详细、完整登记每笔往来款项，每年进行一次账龄分析。

第九十三条 坏账核算方法：按应收账款及其他应收款期末余额的 5%进行计提，采用备抵法对坏账准备进行核算。坏账的确认标准如下：

- (1) 因债务人破产，依照法定程序清偿后，确定无法收回的应收款项；
- (2) 因债务人死亡，既无遗产可供清偿，又无义务承担人，确定无法收回的应收款项；
- (3) 因债务人逾期三年未履行偿债义务，经公司股东大会批准，可以全额列作坏账的应收款项。

第十六章 生产成本的核算

第九十四条 生产过程中实际消耗的直接材料、直接工资、其他直接支出计入生产成本。

第九十五条 各车间为组织和管理生产所发生的各项间接费用计入制造费用。

第九十六条 公司实行品种法核算产品成本，按产品品种归集成本费用。

第九十七条 公司设置直接材料、直接人工、制造费用三个成本项目，三个成本项目之和构成产品制造成本。

第十七章 经营成本的核算

第九十八条 在公司业务经营中发生的与经营有关的各项支出等按规定计入成本。

第九十九条 营业及管理费用主要包括：职工工资、职工福利费、广告及业务宣传费、业务招待费、差旅费、水电费、租赁费、员工培训经费、办公费、会议费、其他费用等。

第一百条 产品销售过程中所发生的各种费用计入销售费用。

第一百〇一条 公司为组织和管理生产经营活动而发生的各项费用计入管理费用。

第一百〇二条 公司为筹集生产经营所需资金而发生的费用计入财务费用。

第一百〇三条 下列各项开支不得列入成本费用：

- (1) 为购置和建造固定资产的支出；
- (2) 无形资产和其他资产的支出；
- (3) 对外投资的支出；
- (4) 被没收财物、支付滞纳金、罚款、违约金、赔偿金的支出；
- (5) 捐赠、赞助性支出；
- (6) 国家规定不得列入成本、费用的其他支出。

第一百〇四条 公司的下列费用按照下列规定进行提取和管理:

- (1) 职工养老保险按公司职工岗位工资总额的 14%提取;
- (2) 职工失业保险按公司职工岗位工资总额的 0.48%提取;
- (3) 住房公积金按公司职工岗位工资总额的 12%提取;
- (4) 职工教育经费按公司职工岗位工资总额的 1.5%提取;
- (5) 职工工伤保险按公司职工岗位工资总额的 0.2%提取;
- (6) 职工生育保险按公司职工岗位工资总额的 0.85%提取;
- (7) 职工医疗保险按公司职工岗位工资总额的 8%提取;

第一百〇五条 费用控制及管理严格按有关制度执行。

第十八章 经营收入的核算

第一百〇六条 公司已将产品所有权上的重要风险和报酬转移给买方, 并不再对产品实施继续管理权和实际控制权, 相关的收入已经收到或取得收款证据, 并且与销售产品有关的成本能够可靠地计量时, 确认为经营收入的实现。

第一百〇七条 公司在生产经营活动中按照合同、协议收取的佣金, 计入营业外收入或者冲减有关经营成本; 按照合同、协议支付的佣金, 计入有关成本费用。

第一百〇八条 销售过程中发生的现金折扣在实际发生时计入当期费用。销售折让在实际发生时冲减当期收入。

第一百〇九条 公司已确认收入的售出商品发生销售退回的, 应当冲减退回当期的收入: 年度资产负债表日及以前售出的商品, 在资产负债表日至财务会计报告批准报出日之间发生的退回的, 应当作为资产负债表日后事项调整事项处理, 调整资产负债表日编制的会计报表有关收入、费用、资产、负债、所有者权益等项目的数字。

第十九章 投资收益与营业外收支的核算

第一百一十条 投资收益是本公司以各种方式对外直接投资取得的利润、利息等收入。

第一百一十一条 营业外收支是指与本公司经营无直接关系的各项收入, 包括固定资产经营收入、租赁收入、固定资产盘盈、固定资产清理净收益、教育费附加返还款、罚款收入、出纳款收入、证券交易差错收入、因债权人的特殊原因却是无法支付的应付款项等。

第二十章

第一百一十二条 本公司利润总额按下列公式计算: 利润总额=营业利润纳税、利润及其分配+投资收益+营业外收支净额

第一百一十三条 净利润=利润总额-应纳所得税

第一百一十四条 营业利润=主营业务收入-主营业务成本+其他业务利润-营业费用-管理费用-财务费用
国债投资收益根据规定不缴企业所得税。

第一百一十五条 所得税后利润根据公司章程的规定, 按如下顺序分配:

- (一) 弥补上一年度的亏损;
- (二) 提取法定公积金 10%;
- (三) 提取任意公积金;
- (四) 支付股东股利; 公司法定公积金累计为公司注册资本的 50%以上的, 可以不再提取。提取法定公积金后, 是否提取任意公积金由股东大会决定。公司不在弥补亏损和提取法定公积金或公益金之前向股东分配利润。

第一百一十六条 本公司须依法申报和缴纳增值税、城市维护建设税、教育费附加、企业所得税及其他应纳税种。

第二十一章 会计报表的种类与编制

第一百一十七条 会计人员应当根据会计账簿的记录编制会计报表。本公司应分年, 半年、季度和月度填制下列报表。

第一百一十八条 企业向外提供的会计报表包括：资产负债表；利润表；现金流量表；资产减值准备明细表；利润分配表；股东权益增减变动表；分部报表；其他有关附表。内部管理报表主要由银行存款收支月报表、成本月报表、债权债务明细表等，企业应根据管理的实际需要设计编制内部管理报表，并根据企业发展需要增加或减少报表种类，调整报表内容、报表格式。内部管理财务报表的设计、编制、报送须经财务经理批准，其他任何部门和人员无权直接要求财务部门

提供财务报表、财务数据。

第一百一十九条 各种报表、项目之间的勾稽关系必须对应、准确；本期报表与上期报表之间的有关数据必须相互衔接。

第一百二十条 向外公布的报表如发现有错误，应及时办理订正手续，除本单位留存报表外，还应同时通知接收报表的单位；错误较多的应重新编报。对外报送的财务报表在报送之前必须由分管财务的财务经理和其他高管审核、签字。

第一百二十一条 上报会计报表时，应当制作会计报表封面，并加盖公司公章。公司负责人、财务部经理/副经理和报表编制人应当在报表封面上签章。

附第二十二章 则

第一百二十二条 本制度由财务部负责解释和修订。

第一百二十三条 本制度经公司自颁布之日起实施。

2015 年 01 月 01 日

1.5 银行函证资料

查阅工商银行函证要求如下：

The screenshot displays the 'Big Data Audit Training Platform' (大数据审计实训平台) interface. The main content area is titled '任务描述:' (Task Description) and contains the text '工商银行函证要求' (Bank Confirmation Requirements), which is highlighted with a red box. The left sidebar shows a navigation menu with '案例背景' (Case Background) and '相关资料' (Related Materials). Under '相关资料', there is a list of items: '1、会计师事务所简介' (Introduction of Accounting Firm), '2、公司章程' (Articles of Association), '3、会计核算制度' (Accounting System), and '4、银行函证资料' (Bank Confirmation Materials), with the last item also highlighted by a red box. The bottom of the interface shows a search bar and a page indicator '1 / 1'.

关于银行函证业务受理的公告

尊敬的客户：

根据财政部、国家金融监督管理总局关于银行函证业务的相关规定，现对我行函证业务受理有关事项公告如下：

一、电子询证函办理方式

我行受理的电子询证函业务包括函证e信和中国银行业协会银行函证区块链服务平台两种办理渠道。

（一）函证e信

1、签署协议、客户对接

函证e信支持会计师事务所从API或我行企业网银发起业务申请，会计师事务所与我行签署《电子询证函业务协议》并完成技术对接后，即可使用函证e信开展业务。

2、函证单位发起函证申请

选择API接入模式的会计师事务所，需与我行API开放平台系统对接，通过自身审计系统发起函证申请。

选择企业网银接入模式的会计师事务所，需开通我行企业网银，通过我行企业网银发起业务申请。

3、被函证单位授权

被函证单位通过我行企业网银、工银聚融等方式授权确认。

4、回函

函证e信自动处理函证数据，通过API接口或企业网银生成回函，反馈给函证单位。

5、业务咨询

会计师事务所可向被函证单位开户网点、各分行询证函业务牵头部门咨询函证e信业务。

（二）通过中国银行业协会银行函证区块链服务平台向我行发函

1、签署协议

会计师事务所与中国银行业协会签署函证区块链平台服务协议，签署风险知情书。

2、业务咨询

对银行函证区块链服务平台接入流程、收发函流程的相关咨询，请联系中国银行业协会；其他银行询证函业务咨询请联系95588；关于被函证单位授权的相关咨询，请联系被函证单位开户行。

二、纸质询证函集中办理方式

我行纸质询证函业务由一级分行函证服务团队集中处理。

（一）集中受理要求

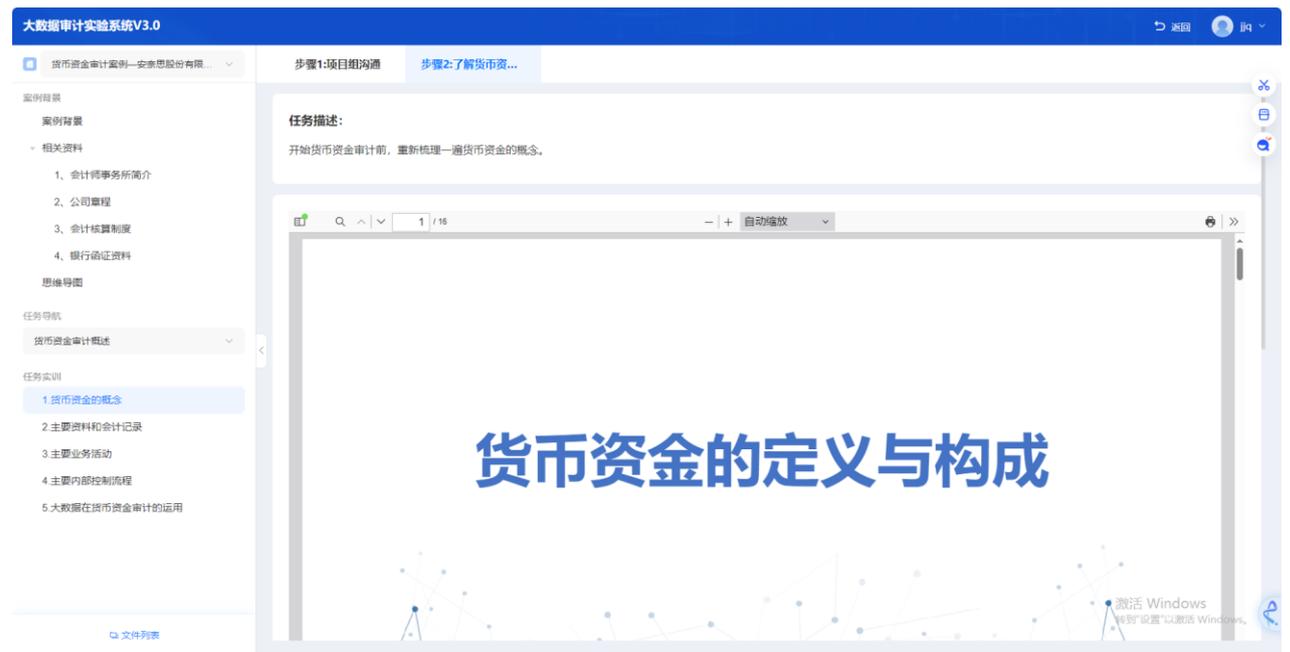
1、函证服务团队接受境内会计师事务所邮寄函证，暂不接受其他第三方（非会计师事务所及其与会计师事务所联合）询证函、跟函方式询证函。

2、函证服务团队接受符合《财政部办公厅 金融监管总局办公厅关于印发〈银行函证工作操作指引〉的通知》（财办会〔2024〕2号）（以下简称“《通知》”）要求的银行询证函标准格式（格式一或格式二，推荐格式一）

第2课时 步骤2：货币资金审计概述

2.1 货币资金的概述

在进行货币资金审计之前先让学生重新认识货币资金的内容。



货币资金指企业拥有的以货币形态存在的流动资产，像企业保险柜里的现金，银行账户中的存款等，是流动性最强的资产项目。

会计要素属性：它属于资产负债表中的流动资产，能反映企业即时支付能力。例如企业有 100 万银行存款，可随时用于支付货款。

基本特征：货币资金具备普遍可接受性、即时支付能力、价值稳定性三大核心特征。

货币资金的功能价值

经营活动基础：货币资金是企业经营活动的基础，支撑日常采购、薪资发放等运营支出，维持企业持续运转，如每月用资金发工资 50 万。

偿债能力体现：它是衡量短期偿债能力的关键指标，影响企业信用评级。若企业货币资金充足，能按时偿还短期债务，信用评级会更高。

战略灵活性保障：货币资金为投资机会提供即时资金支持，增强企业应对风险能力。如遇优质投资项目，可立即用资金投入。

货币资金的内部控制体系

岗位分离原则：出纳与会计岗位严格分离，严禁一人兼任不相容岗位，如出纳不得兼任稽核、会计档案保管等，保障资金管理独立性。

授权审批制度：建立分级授权体系，大额资金支付需多层审批。如超过 50 万的资金支出，需部门经理、财务总监、总经理三级审批。

定期盘点机制：现金每日盘点，银行存款每月对账，确保资金安全。每月末编制银行存款余额调节表，及时发现未达账项。

常见问题与风险防范

主要风险点：货币资金主要风险点包括资金挪用、 欺诈风险、流动性不足风险等。如曾有企业员工挪用资金炒股致公司损失。

防范措施：可实施资金监控系统、定期内部审计、强化员工职业道德培训。像企业安装监控系统实时掌握资金流向。

监管要求：企业货币资金管理需符合《企业内部控制规范》及行业监管机构的专项要求，确

保资金管理合规。

2.2 主要资料和会计记录

点击进行主要资料和会计记录的查看。了解审计货币资金所需要的相关材料。

任务描述:
完整且规范的审计资料和审计证据,是保证审计质量的关键。在进行货币资金审计前,认真思考需要哪些审计资料。

货币资金审计所需相关资料:

会计科目	资料名称	格式要求	企业数据(样例)
库存现金	12月企业自盘现金盘点表	包含盘点人签字和盖章版本	
	已开立银行账户清单	记载户名、开户行、账号	
	各银行账户的对账单	系统导出含银行电子章版本	

明确每个资料所对应的审计作用。

明确货币资金相关审计资料的作用

项目	作用
企业自盘表	<p>评估内部控制有效性: 审计人员可通过企业自盘表了解企业盘点计划的制定情况,评估其是否充分考虑了存货的特点、盘点的流程和方法等,从而判断企业内部控制的设计是否合理。</p> <p>检验盘点程序执行情况: 自盘表可以反映企业盘点程序的实际执行情况。审计人员可以检查盘点过程中是否遵循了既定的程序,如盘点人员的操作是否规范。</p>
企业开户清单	<p>验证账户完整性: 审计人员可将企业开户清单与账面账户情况进行逐一核对,以确认企业是否存在未入账的银行账户,即是否存在体外循环和账外账的情况。比如,若企业账面记录只有3个银行账户,但开户清单显示有5个,那就可能存在账外账户,这有助于发现潜在的财务舞弊行为。</p> <p>确认账户权利和义务: 审计时要确认银行存款的开户人是否为被审计单位。若存款人非被审计单位,需获取该账户户主和被审计单位的书面声明,并确认资产负债表日是否需要调整,以此明确企业对账户的权利和义务关系。</p> <p>关注账户限制情况: 关注开户清单中是否存在质押、冻结等对发现有限制或存在境外的款项。这些受限资金会影响企业的资金流动性和偿债能力,审计人员需要根据清单信息,要求企业作出必要的调整和披露。</p> <p>为其他审计程序提供依据: 开户清单为银行函证等的编制提供了基础信息,审计人员可以根据清单上的账户信息向银行发出询证函,进一步验证账户余额、发生额以及是否存在质押、冻结等情况,提高审计证据的可靠性。</p>

项目	作用
企业自盘表	<p>评估内部控制有效性: 审计人员可通过企业自盘表了解企业盘点计划的制定情况,评估其是否充分考虑了存货的特点、盘点的流程和方法等,从而判断企业内部控制的设计是否合理。</p> <p>检验盘点程序执行情况: 自盘表可以反映企业盘点程序的实际执行情况。审计人员可以检查盘点过程中是否遵循了既定的程序,如盘点人员的操作是否规范。</p>

项目	作用
	<p>范、盘点记录是否完整等，进而评估企业内部控制的有效性。若盘点表中记录的盘点程序与企业规定的程序不一致，可能意味着内部控制存在缺陷。</p> <p>减少重复工作:企业自盘表为企业内部的盘点结果，审计人员可以利用这些已有的盘点信息，减少重复盘点的工作量，节省审计时间和成本。同时，自盘表中的相关记录和数据也可以为审计人员提供参考，有助于审计人员更全面地了解企业的存货情况</p>
企业开户清单	<p>验证账户完整性:审计人员可将企业开户清单与账面账户情况进行逐一核对，以确认企业是否存在未入账的银行账户，即是否存在体外循环和账外账的情况。比如，若企业账面记录只有3个银行账户，但开户清单显示有5个，那就可能存在账外账户，这有助于发现潜在的财务舞弊行为。</p> <p>确认账户权利和义务:审计时要确认银行存款的开户人是否为被审计单位。若存款人非被审计单位，需获取该账户户主和被审计单位的书面声明，并确认资产负债表日是否需要调整，以此明确企业对账户的权利和义务关系。</p> <p>关注账户限制情况:关注开户清单中是否存在质押、冻结等对变现有限制或存在境外的款项。这些受限资金会影响企业的资金流动性和偿债能力，审计人员需要根据清单信息，要求企业作出必要的调整和披露。</p> <p>为其他审计程序提供依据:开户清单为银行询证函的编制提供了基础信息，审计人员可以根据清单上的账户信息向银行发出询证函，进一步验证账户余额、发生额以及是否存在质押、冻结等情况，提高审计证据的可靠性。</p>
银行对账单	<p>①核对企业银行存款日记账与对账单余额，验证账务准确性，发现记账错误或未达账项(如企付银未付、银收企未收等)；</p> <p>②追踪资金流向，确认交易真实性，防范重复付款、盗刷等操作风险及外部欺诈；</p> <p>③评估企业资金管理内控有效性，检查审批流程、不相容职务分离等制度执行情况；</p> <p>④发现异常交易(如大额频繁资金流动、与经营不符款项)，预警财务风险；</p> <p>⑤为审计提供关键线索，支撑银行存款真实性、权利义务认定等审计结论。</p>
银行账户的余额调节表	<p>主要用于核对企业银行存款账面余额与银行对账单余额的差异：</p> <p>①通过调节表，可确认未达账项(如企业已记账但银行未处理的收支、银行已处理但企业未记账的收支)，验证账面记录的准确性，发现记账错误或遗漏；</p> <p>②检查调节事项的合理性，识别潜在未披露交易或资金异常(如长期未达账项可能隐含挪用风险)，辅助评估内控有效性，确保银行存款余额真实完整，为审计结论提供关键证据支撑。</p>
征信报告	<p>①验证企业信用状况真实性(如贷款、担保记录是否与账面负债匹配)，识别潜在或有负债(如未入账的对外担保、违约记录)；</p> <p>②评估企业融资能力及偿债风险(如贷款逾期、授信额度使用情况)</p> <p>③发现异常信用事件(如频繁借贷、高额负债)，辅助判断财务稳定性；</p> <p>④为关联交易、资金链风险等审计重点提供佐证，支撑审计结论的全面性与准确性</p>

2.3 有关货币资金审计作用的自我检测

按照指示，进行自我检测。

1. 审计中发现某银行账户对账单存在多笔“相同金额一收一付”交易，最可能表明：（单选题 1分）

- A. 正常资金调度
- B. 虚构交易或资金循环舞弊
- C. 银行系统自动冲正
- D. 汇率波动调整

正确答案:B

试题解析:短期内相同金额的一收一付属于典型异常交易，可能用于虚增流水、掩盖资金占用或配合收入造假(如监管指引提示的舞弊迹象)。选项 A、C、D 虽可能发生，但需结合业务合理性判断，而该模式更符合舞弊特征。

2. 为识别通过个人账户转移资金的舞弊，注册会计师应优先执行：（单选题 1分） ?

- A. 检查银行存款余额调节表
- B. 核对开户清单与交易单据中的收付款账户，关注非清单内账户（如个人账户）
- C. 函证银行存款余额
- D. 重新计算银行手续费

正确答案:B

试题解析:开户清单包含所有对公账户，若交易单据中收款方为个人或未在清单内的账户，可能表明资金被转移至账外。选项 A、C、D 主要验证已记录账户无法直接发现通过个人账户的挪用行为。

3. 为识别通过隐藏账户转移资金，注册会计师应如何结合征信报告与开户清单？（单选题 1分） ?

- A. 仅核对两者账户数量是否一致
- B. 交叉验证开户清单的账户与征信报告信贷记录，排查无信贷记录却有大量流水的账户
- C. 依赖企业提供的账户说明
- D. 重点关注征信报告中逾期账户

正确答案:B

试题解析:隐藏账户可能无信贷记录但存在大额异常流水(如挪用资金通道)，需交叉比对开户清单与征信信贷流水。选项 A 忽略交易细节;选项 C 易被误导;选项 D 与账户隐藏无关。

2.4 查阅了解货币资金的主要业务活动



流程步骤:收款时严格遵循审核单据、收款、开具收据、登记日记账的步骤，付款则是审批申请、复核、付款、登记，保证每笔收支清晰可查。

审批权限:金额分级审批明确，如：5000元以下由部门经理审批，超过5000元需财务总监审批，保障资金支付安全。

凭证管理:收付凭证需双人复核，且附件要完整，如发票、验收单等原始单据，以确保凭证真实有效。

限额管理:依据企业日常现金用量设定库存限额，超额部分当日送存银行，降低现金滞留风险，提高资金安全性。

日清日结:每日下班前对现金进行盘点，仔细核对账实，若有差异及时查明原因并处理，保证现金账实相符。

安全措施:配备专业保险柜，由双人分别保管钥匙，同时监控全面覆盖，还需定期检查安防设备，确保现金存放安全。

根据所学内容完成如下测试：

1. 现金管理中，出纳员每日对库存现金自行盘点后，应编制 **(单选题 1分)** ?

A. 银行存款余额调节表

B. 现金报表

C. 财务报表

D. 库存现金盘点表

正确答案:B

试题解析:出纳员每日对库存现金自行盘点，需编制现金报表，计算当日现金收入、支出及结余，并将结余与实际库存额进行核对。银行存款余额调节表是在月末由会计主管指定人员编制;财务报表是企业整体财务状况的反映，并非出纳盘点现金后编制;库存现金盘点表是在月末会计主管指定人员盘点时编制。

2. 下列各项中,符合货币资金业务内部控制要求的是 (单选题 1分) ?

- A. 出纳员负责登记现金日记账和总账
- B. 出纳员负责编制银行存款余额调节表
- C. 内部审计人员对货币资金业务进行审查
- D. 财务主管统一保管各类印章

正确答案:C

试题解析:良好的货币资金内部控制要求岗位分离、相互制约和监督。选项 A, 出纳员负责登记现金日记账, 但总账应由其他会计人员登记, 出纳不能同时登记总账, 否则无法保证账目的准确性和监督性;选项 B, 银行存款余额调节表应由出纳员以外的人员编制, 以确保调节过程的独立性和客观性, 出纳员编制可能导致调节不彻底或掩盖问题;选项 D, 财务主管统一保管各类印章, 不符合印章管理要求, 财务专用章和个人名章应分别由专人保管, 严禁一人保管支付款项所需的全部印章, 以防止印章被滥用。而选项 C, 内部审计人员对货币资金业务进行审查, 有助于发现内部控制中的问题, 符合内部控制的要求。

3. 在货币资金审计中,获取被审计单位银行账户开户清单的主要目的是 (单选题 1分)

- A. 核对银行存款日记账与总账的一致性
- B. 确认被审计单位是否存在未入账的银行账户 (如账外账户)
- C. 检查银行存款余额调节表的编制是否正确
- D. 验证银行对账单的真实性

正确答案:B

试题解析:开户清单能帮助审计师全面掌握被审计单位所有银行账户信息,与账面记录比对,识别是否存在未披露的账户(如账外账、小金库等),是发现潜在舞弊的重要程序。其他选项(A、C、D)分别对应其他审计程序(如账务核对、调节表检查、银行函证)与开户清单的直接作用无关。

2.5 查阅了解货币资金的内部控制



内涵界定：货币资金内部控制是企业为保证货币资金安全、完整，防范舞弊风险而建立的一系列控制流程与制度体系。例如企业会制定严格的资金审批流程。

控制对象：控制对象涵盖现金、银行存款、其他货币资金及票据、印章等关键载体。像企业对印章的保管会有专人负责。

监管依据：其监管依据基于《企业内部控制基本规范》及财会〔2010〕11号文等法规要求，确保企业内控合法合规。

库存限额管理：根据日常用量设定库存上限，超额现金及时送存银行，如某企业规定每日库存不超

过5万元，确保现金安全。

收支两条线：收款与付款岗位严格分离，严禁坐支现金，确保资金流向清晰可追溯，避免资金混乱。

日清日结制度：每日下班前盘点库存现金，与账面余额核对一致，编制《现金盘点表》，保证账实

相符。

现金使用范围控制：明确现金使用范围，如仅用于小额零星支出，超范围结算需通过银行转账，防止现金滥用。

操作风险：操作风险常见于日常业务处理，如出纳未及时登记日记账，某企业就因这一疏忽产生3万元短款，影响资金核算准确性。

舞弊风险：舞弊风险危害极大，像伪造银行对账单掩盖挪用行为，某公司出纳借此挪用资金200万元，严重损害公司利益。

合规风险：合规风险涉及违法违规操作，如违规对外担保或出借资金，面临监管处罚，最高罚款可达50万元。

不相容岗位分离：不相容岗位分离是重要防控手段，规定出纳不得兼任稽核等工作，可有效防止错误和舞弊发生。

印章分存管控：印章分存管控保障资金安全，财务专用章和法人章分别由会计和出纳保管，

严禁一人保管全部印章。

信息化监控系统：信息化监控系统实时监控资金动态，部署资金管理系统可对大额支付、异常交易预警，响应时间缩至 2 小时。

定期审计检查：内部审计部门每季度开展资金内控专项审计，严格把控资金流向，发现问题整改率需达 100%，确保资金安全。

穿行测试：模拟资金支付全流程测试控制有效性，年度测试覆盖率不低于 80%，全面评估内控流程的可靠性。

考核问责制度：将内控执行情况纳入财务人员 KPI，与绩效薪酬直接挂钩，增强财务人员的责任意识和执行力度。

根据所学内容完成如下测试：

1. 在对甲公司2020年度财务报表进行审计时，A注册会计师负责审计货币资金项目。2020年12月28日对甲公司全部现金进行监盘后，确认有现金数额为4500元。甲公司2020年12月29日至12月31日现金收入总额为9300元、现金支出总额为9500元，则推断2020年12月31日库存现金余额应为（ ）元（单选题 1分）
- A. 4100
 - B. 4200
 - C. 4300
 - D. 4400

正确答案:C

试题解析:选项 C 正确:将资产负债表日前的盘点金额调整至资产负债表日的金额，即:期末实存数=盘点日的金额+盘点日至资产负债表日现金收入总额-盘点日至资产负债表日现金支出总额=4500+9300-9500=4300(元)。

2. 企业在确定职权和岗位分工过程中，应当体现不相容职务相互分离的要求。下列不属于不相容职务的是（ ）。
- A. 采购审批与执行
 - B. 客户信用调查与审批
 - C. 现金日记账与银行存款日记账的编制
 - D. 可行性研究与决策审批

正确答案:C

试题解析:现金日记账与银行存款日记账的编制都是出纳的职责，不属于不相容职务。

3. 下列现金盘点内部控制设计中，存在重大缺陷的是（单选题 1分） ?
- A. 会计主管指定应付账款会计每月末盘点现金，并编制现金盘点表
 - B. 出纳员每日下班前自行清点现金，记录台账备查
 - C. 财务经理复核现金盘点表，对超限额差异调查审批
 - D. 现金盘点需在第三方监督下突击进行

正确答案:A

试题解析:现金盘点应由独立于现金保管和记录职责的人员执行(如会计主管或内部审计)，选项 A 中应付账款会计负责现金盘点，违背职责分离原则。选项 B、C、D 符合突击盘点、复核监督等内控要求。

2.6 查阅了解大数据工具在货币资金审计的应用

了解 Excel、BI 工具、Python 等工具在大数据审计中的应用。

一、Excel 在货币资金审计中的应用

1、基础数据分析

流水核对:通过 **VLOOKUP** 函数匹配银行对账单与账面记录,快速定位未达账项透视表汇总:按账户、日期分类汇总收支金额,识别异常波动(如月末集中大额支出)。

2、内控测试与合规检查

交叉索引:利用条件格式标记重复付款、超限额现金交易等违规行为。

3、局限性

处理超 10 万条数据时性能下降,需结合 **Power Query** 或数据库工具:、**BIT** 具在货币资金审计中的应用

1、可视化监控

动态仪表盘:实时监控现金比率、存贷双高等风险指标,桑基图展示资金流向:直观呈现跨账户、跨主体资金流转路径,发现隐匿关联交易,

2、智能分析与预警

预算偏差分析:自动比对实际支出与预算,标记超支项目。欺诈模式识别:内置机器学习模型(如毕马威审计罗盘)匹配异常交易特征(如高频深夜转账)

3、整合能力

支持多源数据(ERP、网银、税务系统)融合,生成合规性审计报告。

三、Python 在货币资金审计中的应用

1、自动化数据处理

数据清洗与预处理:使用 **Pandas** 库快速清洗银行流水数据,处理缺失值、重复记录及异常值(如识别跨期入账或白条抵库现象)。交易流水分析:通过正则表达式解析非结构化数据(如支付宝/微信交易备注),提取关键信息(对方账户、用途)并与账面记录核对。异常交易识别:应用机器学习算法(如孤立森林、**K-Means** 聚类)检测异常资金流向,例如频繁小额转账至关联方或异常时间交易

2、风险建模与可视化

资金流向图谱:利用 **NetworkX** 构建资金流转网络,识别循环交易或体外资金循环。

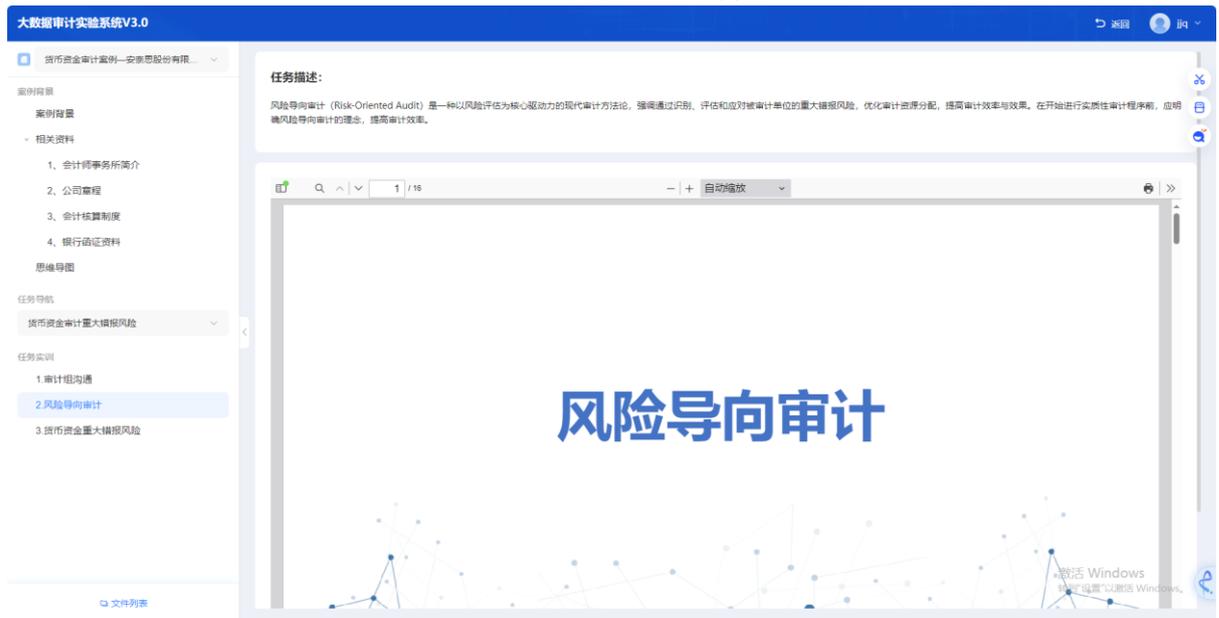
动态监控:结合 **Kafka** 流处理技术实时监控大额资金异动,触发预机制。



第3课时 步骤3: 货币资金重大错报风险

3.1 风险导向审计

点击货币资金重大错报风险——风险导向审计, 观看 ppt。



风险导向审计是一种以风险评估为核心驱动力的现代审计方法论, 强调通过识别、评估和应对被审计单位的重大错报风险, 优化审计资源配置, 提高审计效率与效果, 在开始进行实质性审计程序前, 应明确风险导向审计的理念, 提高审计效率。

3.2 货币资金错大重报风险

货币资金的重大错报风险是指在审计过程中, 与货币资金的交易、账户余额和披露相关的认定层面出现重大差错的可能性。

货币资金审计案例—安泰股份的有限...

案例背景
案例背景
相关资料
思维导图

任务导航
货币资金审计重大错报风险

任务实训
1. 审计组沟通
2. 风险导向审计
3. 货币资金重大错报风险

任务描述:

货币资金的重大错报风险是指在审计过程中，与货币资金的交易、账户余额和披露相关的认定层面出现重大差错的可能性。请学习相关知识点并完成各部分练习：

认定层次重大错报风险	具体情形
虚假余额或交易风险	被审计单位可能存在虚假的货币资金余额或交易，导致银行存款余额的“存在”认定或交易的“发生”认定出现重大错报。例如，企业可能虚构现金收支业务来调节利润或掩盖其他财务问题。
外币交易计价错误风险	若被审计单位存在大额的外币交易和余额，可能因未采用正确的折算汇率，导致外币交易或余额未被准确记录，影响“准确性、计价和分摊”认定。比如，企业有外币现金或外币银行存款，但在核算时使用了错误的汇率。
截止性错误风险	银行存款的期末收支可能存在大额的截止性错误，如被审计单位期末存在金额重大且异常的银付企未付、企收银未收事项，影响“截止”认定。
披露风险	被审计单位可能未能按照企业会计准则的规定对货币资金作出恰当披露，如期末持有使用受限制的大额银行存款，但在编制财务报表时未在财务报表附注中进行披露，存在“列报”认定风险。

认定层次重大错报风险	具体情形
虚假余额或交易风险	被审计单位可能存在虚假的货币资金余额或交易，与导致银行存款余额的“存在”认定或交易的“发生”认定出现重大错报。例如，企业可能虚构现金收支业务来调节利润或掩盖其他财务问题。
外币交易计价错误风险	若被审计单位存在大额的外币交易和余额，可能因未采用正确的折算汇率，导致外币交易或余额未被准确记录，影响“准确性、计价和分摊”认定。比如，企业有外币现金或外币银行存款，但在核算时使用了错误的汇率。
截止性错误风险	银行存款的期末收支可能存在大额的截止性错误，如被审计单位期末存在金额重大且异常的银付企未付、企收银未收事项，影响“截止”认定。
披露风险	被审计单位可能未能按照企业会计准则的规定对货币资金作出恰当披露，如期末持有使用受限制的大额银行存款，但在编制财务报表时未在财务报表附注中进行披露，存在“列报”认定风险。

学习货币资金重大错报风险后，完成学习实训训练内容。

1. 下列情形中，注册会计师最需关注货币资金“存在”认定重大错报风险的是（ ）（单选题 1分）

A. 被审计单位存在大额外币交易，但未按汇率波动调整账面价值
 B. 银行存款日记账余额与银行对账单差异较大，且存在长期未达账项
 C. 企业将资金存放于管理层个人账户进行结算
 D. 货币资金收支与现金流量表经营活动现金流不匹配

正确答案:C

试题解析:选择 C:

A.被审计单位存在大额外币交易但未按汇率波动调整账面价值风险性质:涉及外币折算错误，属于计价和分摊认定(准确性)问题。与“存在”认定的关联:未调整汇率可能导致账面价值不准确，但不直接质疑资金是否存在。

结论:非“存在”认定的核心风险。8.银行存款日记账余额与银行对账单差异较大，且存在长期未达账项风险性质:可能暗示记账错误、截止问题或资金挪用，影响完整性或截止认定。

与“存在”认定的关联:长期未达账项可能掩盖虚构存款才能直接指向“存在”风险。

结论:需警惕, 但非最直接证据。C.企业将资金存放于管理层个人账户进行结算
风险性质:属于舞弊高风险信号, 资金可能被挪用、侵占或虚构。与“存在”认定的关联:
个人账户资金不被企业实际控制, 账面记录的存款可能不存在(虚增资产)。

实务案例(如康美药业)表明, 此类操作常伴随伪造凭证、虚构余额。

结论:最需关注“存在”认定风险, 因涉及资金真实性本质问题。D.货币资金收支与现金流量表经营活动现金流不匹配风险性质:可能反映收支分类错误或隐匿交易, 影响列报或准确性认定。与“存在”认定的关联:不匹配可能因记账错误, 但不直接证明资金不存在(如可能涉及其他活动现金流错报)

2. 注册会计师在审查甲公司编制的银行存款余额调节表时, 注意到以下事项: 甲公司已收但银行尚未入账的款项; 甲公司已付但银行尚未入账的款项; 银行已收但甲公司尚未入账的款项; 以及银行已代扣但甲公司尚未入账的费用。为了正确反映银行存款的实际余额, 注册会计师应该采取什么措施? (单选题 1分)

A. 根据银行对账单余额加上所有未到账项调整后的余额

B. 根据银行对账单余额加上所有未到账项调整后的余额

C. 将银行存款日记账余额与银行对账单余额直接相加平均

正确答案:B

试题解析:选择B

根据银行对账单余额加上所有未到账项调整后的余额。

3. 在对某公司进行现金监盘时, 注册会计师发现实际现金数额低于账面记录的金额。这种情况最可能违反了以下哪一项认定? (单选题 1分)

A. 完整性

B. 存在

C. 计价和分摊

D. 截止

正确答案:B

试题解析:货币资金的存在是指金额真实存在不虚假

4. 某公司在年底收到银行存款, 但由于会计人员的疏忽, 这笔款项未被记入当年的账目中。这种情况主要影响了哪项财务报表认定?

A. 存在

B. 权利和义务

C. 完整性

正确答案:C

D. 截止

试题解析:完整性认定确保所有应当记录的交易都被记录。在这个例子中, 由于遗漏了一笔已经发生的收入, 导致财务报表未能完整反映公司的财务状况。

第4课时 步骤4: 货币资金内控测试

4.1 货币资金内控审计流程

点击内控审计流程, 了解内控审计流程和各个流程的目的。

大数据审计实训平台

货币资金审计案例—安泰股份有限...

案例背景
案例背景
相关资料
思维导图

任务导航
货币资金内部控制测试

任务实训
1. 内控审计流程
2. 资金活动内部控制评价
3. 资金活动内部控制测试

任务描述:
了解内控审计流程和各个流程的目的

进行查看审计内控流程后，测试所学内容。

1. 下列关于货币资金访谈问卷作用的表述中，错误的是（单选题 1分）

- A. 了解被审计单位货币资金业务的流程、审批权限及关键控制点
- B. 发现货币资金科目余额的错报金额
- C. 评估内部控制设计的有效性及其执行的一贯性
- D. 识别潜在舞弊风险或异常交易模式

正确答案:B

试题解析:选项 A(了解流程与控制点):访谈话卷通过询问财务人员、业务人员等,可获取货币资金收支、盘点、对账等环节的具体操作流程及审批权限,帮助审计师理解业务逻辑,表述正确。

选项 B(直接发现错报金额):访谈问卷属于风险评估程序,目的是获取定性信息(如流程缺陷、控制弱点),而非直接量化错报金额。错报金额需通过实质性程序(如函证、盘点、重新计算)验证,因此“直接发现错报金额”是错误表述。

选项 C(评估内控设计与执行):问卷中设计的针对性问题(如“现金盘点是否由出纳以外人员监盘?”)可测试内部控制的设计合理性及实际执行情况,表述正确。

选项 D(识别舞弊风险):通过询问异常交易(如“是否存在大额现金收支无审批的情况?”)、员工动机(如“财务人员是否面临业绩压力?”)可发现潜在舞弊线索,表述正确。

2. 控制测试和风险评价的目的有所不同，以下关于二者目的的表述，正确的是（单选题 1分）

- A. 控制测试是为了识别和评估被审计单位重大错报风险，风险评价是为了证实被审计单位内部控制是否有效运行
- B. 控制测试是为了证实被审计单位内部控制是否有效运行，风险评价是为了识别和评估被审计单位重大错报风险
- C. 控制测试用于发现财务错报，风险评价用于预防安全事故
- D. 控制测试和风险评价的目的都是为了证实被审计单位内部控制是否有效运行

正确答案:B

试题解析:在风险导向审计模式下，风险评价是必须的审计程序，贯穿于整个审计过程，其工作目的是为了识别和评估被审计单位重大错报风险，所采用的程序包括询问、分析程序、观察和检查等。而控制测试并非必须的程序，审计人员可以选择执行，其目的是为了证实被审计单位内部控制是否有效运行，如果经过风险评估程序发现被审计单位内部控制根本就没有得到执行，则审计人员就可能放弃执行这一程序，但如果审计人员认为仅实施实质性程序获取的审计证据无法将认定层次重大错报风险降至可接受的低水平，就应当执行控制测试程序，其程序包括询问、观察、检查和穿行测试等，还包括重新执行。所以选项 B 正确，AD 错误。

C 错误，控制测试不直接发现财务错报(错报需通过实质性程序验证)风险评价适用于各类风险(财务、运营、安全等)，不限于安全事故。

3. 下列哪项不属于货币资金内控审计的重点关注内容？（单选题 1分）

- A. 银行账户管理
- B. 收入确认政策
- C. 现金与存款安全
- D. 收支合规性

正确答案:B

试题解析:B 属于收入循环审计，与货币资金内控的岗位分离、收支安全、对账监督等核心目标无关

4.2 资金活动内部控制评价

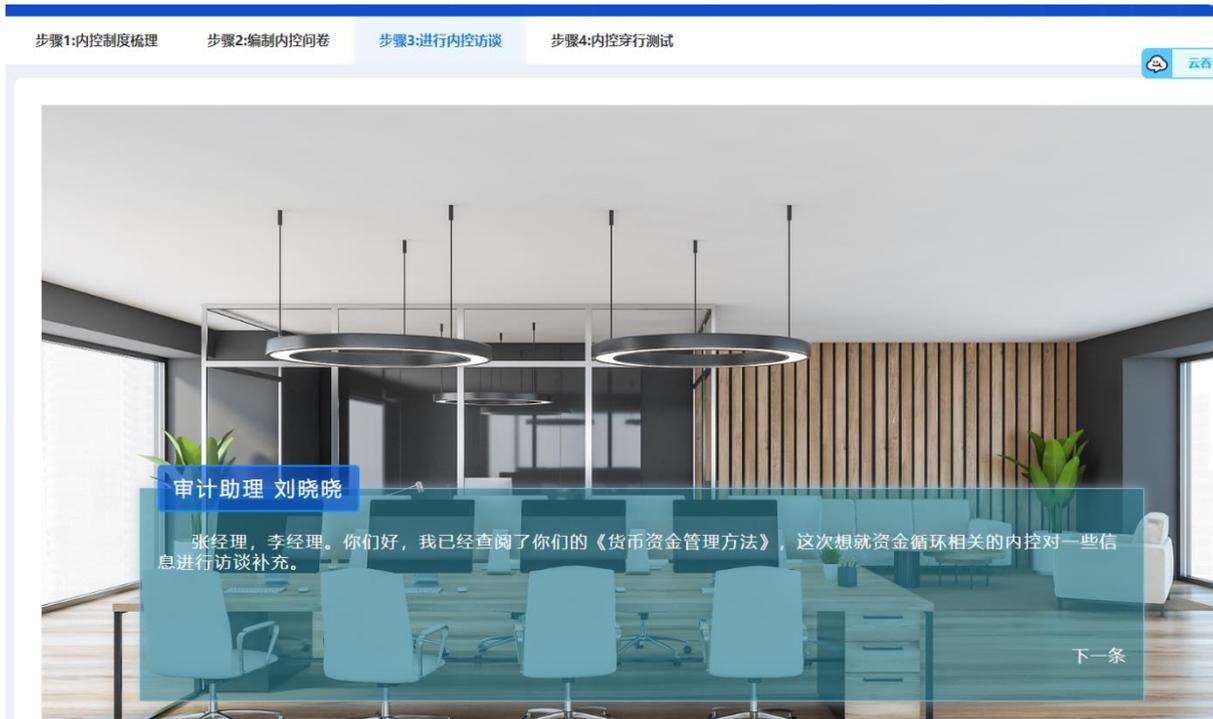
根据安泰思的货币资金管理制度；整理关键业务环节。



之后根据内控制度，整理资金循环的内控问卷。



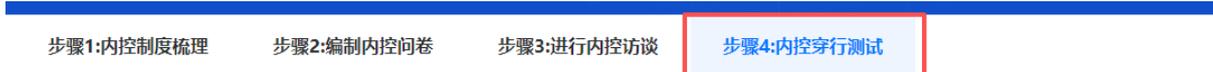
学生点击进行内控访谈，实地感受审计人员与安奈思财务部的访谈过程。



了解安奈思公司的现金盘点流程和逻辑。



之后进行内控穿行测试的底稿编制工作：



货币资金内控底稿-以库存现金盘点为例

1、根据步骤 1-3 了解到的内控业务流程信息，阅读已填写完成的内控评价底稿中【评价控制的设计并确定控制是否得到执行】表的【被审计单位的控制活动】及【相关内部控制制度】。(见文件链接)

任务描述

1、根据步骤 1-3 了解到的内控业务流程信息，阅读已填写完成的内控评价底稿中【评价控制的设计并确定控制是否得到执行】表的【被审计单位的控制活动】及【相关内部控制制度】。(见文件链接)

2、穿行测试是指追踪某笔交易从发生到最终被反映在财务报表中的整个处理过程。注册会计师在护行穿行测试时，通常需要综合运用询问、观察、检查相关文件及重新执行等程序。比如，在审计家公司的销售业务时，注册会计师会从客户下订单开始，一路追踪到订单的处理、发货、开票、记账等一系列环节，看这些环节是否按照既定的流程和规定进行操作。

3、阅读根据步骤 1-3 中所进行的审计程序，已填制完成的【穿行测试程序】表中【检

查的证据】和实施程序的记录。

检查下图由出纳提供的现金盘点表，判断是否由问题，并填制【穿行测试程序】表中【测试结果】，并在【控制活动是否得到执行】中填写 EXCEL 条件判断的函数，使【测试结果】是“已得到执行”或者“部分得到执行”时显示“是”

【评价控制的设计并确定控制是否得到执行】表：

确定控制是否得到执行(穿行测试等程序) ——货币资金循环									
被审计单位：安奈思								索引号： ZJL-3-1	
被审会计报表属期：2020年度								页次：	
编制人：								编制日期：	
复核人：								复核日期：	
6	序号	子流程	控制目标	被审计单位的控制活动	检查的证据	实施的程序记录	测试结果	索引号	控制活动是否得到执行
7	1	现金管理	现金账实相符	分为月度盘点和周盘点：每月末或者随机抽几周，财务主管指定除出纳以外的人员对现金进行盘点，编制库存现金盘点表，将盘点金额与现金日记账余额进行核对。盘点后盘点人签字，财务主管监督复核签字并盖章。	月度库存现金盘点表	· 询问会计主管和出纳的公司现金盘点制度 · 检查公司有关现金盘点的制度文件 · 选取一笔库存现金盘点表，检查是否经会计主管复核并盖章		ZJL-3-1	

进行穿行测试执行测试。检查现金盘点表。

福建安奈思时尚服饰股份有限公司库存现金盘点表

2020年 12月 31日 时点：17：55					
清点现金			核 对 账 目		
货币金额	张数	金 额	日期	项 目	合 计
100元	38	3,800.00	2020年12月31日	现金账面金额	3,897.00
50元	-	-		加：收入凭证未记账	
20元	-	-		减：付出凭证未记账	
10元	-	-		减：借条	
5元	19	95.00		加：票面已付，款未付	
2元	-	-		调整后现金余额	
1元	2	2.00	2020年12月31日		
5角	-	-			
2角				实点现金	3,897.00
1角	-	-		长款	-
2分				短款	
1分	-	-			
实点合计	59	3,897.00		备注：	

财务主管：张清雨

盘点人：张清雨

出纳：李晴

进行穿行测试的题目测试：

穿行测试的样本选择 (单选题 1.00分)

1. 下列穿行测试样本选取做法中, 恰当的是 ()

- A. 选取金额最大的交易, 确保覆盖高风险领域
- B. 选取一笔典型交易, 涵盖所有关键控制节点
- C. 每月固定选择应收账款、管理费用科目测试
- D. 由被审计单位指定样本以简化流程

正确答案:B

解析:

穿行测试通常选取 1 笔典型交易(如包含申请、审批、记录等全流程), 验证控制链完整性。选项 A 可能遗漏常规控制;选项 C 机械重复特定科目, 忽略业务多样性;选项 D 丧失独立性。

2. 在货币资金付款穿行测试中, 注册会计师需特别关注的控制点是 ()

- A. 现金盘点频率
- B. 付款申请是否经多级审批
- C. 银行存款利息计算准确性
- D. 总账与明细账余额核对

正确答案:B

解析:付款审批是防止资金挪用的关键控制点(如未经审批的付款可能涉及舞弊)。选项 A 属于现金管理控制;选项 C、D 属于事后核对, 非穿行测试重点。

穿行测试与控制测试的区别 (单选题 1.00分)

3. 关于穿行测试和控制测试的表述, 正确的是 ()

- A. 两者均需大量样本验证控制运行有效性
- B. 穿行测试关注控制设计, 控制测试关注运行有效性
- C. 穿行测试属于实质性程序, 控制测试属于风险评估程序
- D. 控制测试必须在穿行测试之前执行

正确答案:B

解析:穿行测试在风险评估阶段执行, 验证控制是否存在(样本量通常为 1 笔);控制测试在风险应对阶段执行, 验证控制是否持续有效(需根据业务频率选取样本)。选项 A 错误(穿行测试样本量小);选项 C 错误(两者均属内控测试)。选项 D 顺序颠倒。

4.3 资金活动内部控制测试

首先对出现的问题进行沟通交流。



同时, 由于现金盘点表出现未盖章的情况, 审计人员决定继续扩大控制测试样本量进行进一步的内控审计。

任务描述:

根据穿行测试发现的问题,扩大控制测试样本量进行进一步内控审计。

控制测试的范围		
注册会计师在测试控制的运行有效性时,应当在考虑与控制相关的风险基础上,确定测试的范围,即样本规模。		
注册会计师确定的测试范围,应当足以使其能够获取充分、适当的证据,为基准日内部控制是否存在重大缺陷提供合理保证。		
1. 对人工控制测试的最小样本规模		
对人工控制,如果采用重新执行程序,且预期偏差为零,测试的最小样本量区间参见下表:		
控制运行频率	控制运行的部数	测试的最小样本量区间
每年1次	1	1
每季1次	4	2
每月1次	12	2~5
每周1次	52	5~15
每天1次	250	20~40
每天多次	大于250次	25~60

确定样本总体的方法:

从业务发生的末端取得:即从财务系统中直接取得汇总的会计凭证或明细账信息,或通过明细账的借贷方、凭证摘要、或比较期初期末的差异进行分析获得样本总体。0 从业务发生的前端取得:考海连续编号的登记簿或汇总表,如支票登记簿、合同登记表等;或从信息系统中提取数据列表。采用该种技术时,需要事先做好计划,与公司「人员会谈,了解数据的形成和格式,便于正确完整地提取数据。

根据控制频率直接判断:如月度或季度的控制点,通常测试期间内只有一个样本,则全部抽取;部分控制可通过从测试期间抽取全部交易日(周)等方式作为样本总体。

点数法:即在测试期间内对样本总体进行实际点数。确定抽样的方法:抽样测试是目前首汽采用的烈试方式,主要包括随加抽样和任意抽样,测法人员在没有殖机款件支持的情况下应采取任意抽样(建议在考交易领率的基础上采取等源地样),并以大金额交易和临沂交易为调试重点。测试人员按设计的方法选取了样本后,不能安个人意原取舍,应保证样本的真实性。一旦确定了样本总体后,只能取样一次,不能在抽样的基础上进行二次抽样。例如,支票盘点、银行余额调节表等控制,通常选怪测时期间内的所有联户的样本,而不能先选定个别现试目份,再从有限的期间内再殖意抽取一定比例的联户进行测域,应独保取得一整秦的测法文档。再如季度利息预提控制的测法,可选择 2 个季度内的所有贷款利息预提事项;或选择测试期间内的所有预提事项作为总体再抽取 25 个样本进行测试。穿行测试的样本可作为一个样本用于控制测试,一般适用于每年、每季度、每月发生的控制。

判断控制测试失效的依据:

样本不满足要求:对于某个控制活动,一旦某一个样本不满足控制活动要求时,即可选择测试结果为“失效”。不过,对于每日数次发现 1 例失效样本,在追加 15 个样本测试未发现新不台格样本时,结论可为“有效”。反复失效;试验中出现反复失效的情况,每次失效均会计入失效数内。当失效数达到一定标准,如在可靠性试验中,试验至载尾时间,失效数不小于 3,则符合拒收的判决标准,可认为试样所代表的产品不能通过可靠性测试,即控制测试失效。

虽不计入失效数,但也反映出控制测试可能存在问题,一定程度上可作为控制测试失效从属失效情况:由于试样中其他失效直接或间接引起的奉连性失效,以及因操作不当和试验设备、仪器、仪表等失效而引起的失效的参考依据。

阅读完控制测试抽样情况后,进行课后作业练习:

1. 根据观察和询问，安奈思的库存现金盘点基本每月一次，控制测试应该样本量取多少个合适？

- A. 10
- B. 12
- C. 5
- D. 1

正确答案:C

试题解析:安奈思库存现金年度 12 次,控制运行的部数为 12,因此样本量 2-5 次最合适。

2. 安奈思货币资金的控制测试除了选取库存现金的盘点表检查内控外,还有哪些部分需要进行检查并留痕 (多选题 2分)

- A. 抽取库存现金的支付凭证,检查单据和审批流。
- B. 抽取银行存款的收支凭证,复核签章完整性。
- C. 抽取应付预付的采购凭证,检查是否合规。
- D. 抽取银行余额调节表,检查是否按时无误。

正确答案:ABD

试题解析:应付预付的采购抽样主要在采购和付款循环进行控制测试,而非货币资金循环的控制测试内容。

3. 下列控制测试样本选择做法中,会导致测试失效的有 () (多选题 2分)

- A. 仅选取金额最大的交易,忽略常规业务样本
- B. 抽样总体未包含高风险期间(如12月收入)
- C. 对无效单据(如作废收据)替换后测试
- D. 测试范围未覆盖穿行测试中发现的缺陷领域

正确答案:ABD

试题解析:

选项 A:样本缺乏代表性,无法推断整体控制有效性。

选项 B:遗漏关键期间样本(如年末)导致无法发现截止性错误。

选项 D:未针对已知缺陷扩大测试范围,测试不完整。

选项 C 属于正常替换,不影响测试有效性

第 5 课时 步骤 5: 货币资金实质性程序

5.1 库存现金科目审计

首先进行库存现金的实质性程序

案例背景

- 案例背景
- 相关资料
- 思维导图

任务导航

- 货币资金实质性程序

任务实训

- 1.库存现金科目审计
- 2.银行存款科目审计
- 3.其他货币资金科目审计

任务描述:

库存现金实质性审计程序

一、核对现金日记账与总账

目的: 确保账账一致, 为后续审计提供准确起点。

步骤:

- 1、核对现金日记账余额与总账余额是否相符。
- 2、检查非记账本位币现金的折算汇率及金额是否正确(如外币现金需按适用汇率折算)。

异常处理: 若存在差异, 需查明原因(如记账错误、跨期调整), 必要时建议调整分录。

核对现金日记账与总账

目的:确保账账一致, 为后续审计提供准确起点。

步骤:

- 1、核对现金日记账余额与总账余额是否相符
- 2、检查非记账本位币现金的折算汇率及金额是否正确(如外币现金需按适用汇率折算)

异常处理:若存在差异, 需查明原因(如记账错误、跨期调整), 必要时建议调整分录。

序号	币种	备注	审定数		审计调整		期末数		期初数	
			原币	本币	原币	本币	原币	本币	原币	本币
合计			***	-	***	-	***	-	***	-
1	人民币		账面金额与实际金额核对调整				取自科目余额表期初期末			
2	美元									

监盘库存现金(核心程序)-案例实操重点环节

目的:直接验证现金的“存在性”认定。

步骤:

- 1、制定计划:确定监盘时间与范围, 查阅现金日记账并与收付凭证核对(检查日期与金额一致性)。
- 2、现场清点:出纳清点现金, 注册会计师监督并记录实存数。编制《库存现金监盘表》, 分币种/面值列示金额。
- 3、账实核对:比对实存数与日记账余额, 差异需立即查明原因(如白条抵库、未报销凭证等)。
- 4、特殊问题处理:抵库凭证:借条、未提现支票等需在监盘表注明, 并提请调整。
- 5、非资产负债表日监盘:需倒推调整至资产负债表日金额(公式:资产负债表日余额 = 监盘日余额-期间收入 + 期间支出)

三、抽查大额现金收支

目的:检查交易真实性、合规性及账务处理准确性。

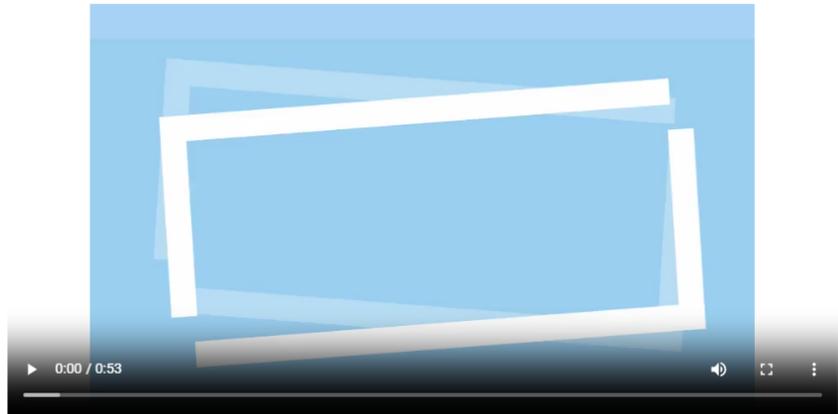
抽查重点:

- 1、原始凭证:检查内容完整性(如发票、合同)、审批手续(授权签字)、法律效力。
- 2、合规性:是否超现金使用范围(如单笔超 1000 元零星支出);是否存在“坐支现金”(收入未存银行直接支付)或超库存限额(一般 3-5 天日常开支)。
- 3、异常交易:关注与经营无关的收支、频繁存取、关联方资金拆借等舞弊迹象。

观看库存现金监盘内容:

任务描述:

2021年2月1日, 审计组到被审计单位财务室, 实施库存现金盘点。



观看完毕后, 学生进行测试题练习

1. 以下关于现金盘点的说法, 正确的有 () (多选题 2分)

- A. 库存现金盘点应在事先不通知被审计单位相关人员的情况下突击进行。
- B. 实施盘点时间应选择营业前(上午上班)或营业终了(下午下班)后。
- C. 当现金存放在单位不同地点的保险柜时, 应安排几个盘点小组同时对单位每一存放处的保险柜现金展开盘点, 或将每一存放处保险柜先做封存, 然后逐一盘点。
- D. 审计人员应邀请被审计单位会计主管、会计、出纳等一同到盘点现场参与盘点及监督工作, 盘点工作一般由被审计单位出纳在现场清点现金并作记录, 审计人员不直接参与盘点, 只是现场监督盘点。

正确答案:ABCD

试题解析:1、库存现金盘点应在事先不通知被审计单位相关人员的情况下突击进行。避免由于被审计单位相关人员了解到审计人员将要实施库存现金盘点, 而早有防备, 致使现金盘点达不到预期效果。

2、实施盘点 a1 时间应选择在营业前(上午上班)或营业终了(下午下班)后, 这样既可避免打扰被审计单位正常的经营业务, 又可防止被审计单位对盘点出的问题有这样或那样的解释, 影响现金盘点预期效果。

3、组织安排库存现金的清点工作前, 先要求出纳取出保险柜中混入属于私人个人的钱, 然后要求出纳将保险柜外所有单位公款全部放入保险柜, 最后封存保险柜, 避免由于单位钱和出纳个人钱混淆, 影响最后盘点数的认定。

4、当现金存放在单位不同地点的保险柜时, 应安排几个盘点小组同时对单位每一存放处的保险柜现金展开盘点, 或将每一存放处保险柜先做封存, 然后逐一盘点。避免被审计单位在现金实际盘点过程中, 采取拆东墙补西墙的办法应付盘点, 致使此次盘点失去原有的作用, 达不到预期的审计目标。

5、审计人员应邀请被审计单位会计主管、会计、出纳等一同到盘点现场参与盘点及监督工作, 盘点工作一般由被审计单位出纳在现场清点现金并作记录, 审计人员不直接参与盘点, 只是现场监督盘点。必要时审计人员可对盘点结果进行复查, 避免审计人员单独或亲自盘点产生一些不必要的麻烦。

最后进行编制库存现金盘点表的过程:

步骤1:库存现金实... 步骤2:库存现金监盘 **步骤3:编制库存现...**

任务描述 [🔗](#) [📄 操作指引](#)

根据前述现金盘点结果,并参考【文件列表】-【货币实质性程序】-【库存现金日记账】的内容,填写“库存现金盘点表”。

单选题 (单选题 1.00分)

1. 根据现金盘点结果,以下说法正确的是: ()

- A. 库存现金盘点是一种账表核对方法
- B. 即使企业期末没有库存现金,现金盘点表也需要被审计单位盖章确认
- C. 审计人员未在2020年12月31日及时盘点,属于审计程序执行不到位
- D. 2021年2月1日盘点的库存现金只需根据库存现金序时账倒推至2020年12月31日金额即可

正确答案:B

解析:现金盘点是必要的审计程序之一,因即使企业期末没有库存现金,现金盘点表也需要被审计单位盖章确认, B 正确;实物盘点法允许审计日迟于报表日,并根据多个账表口径账表倒推实物在报表日的数量, C 错误;还应结合库存现金从报表日到审计日的收支去向进行反向核对,如库存现金日记账发现有从银行账户转出记录,应查询银行存款序时账和流水是否发生该笔业务, D 错误。

盘点角色 (单选题 1.00分)

2. 库存现金盘点通常应该由谁来执行? ()

- A. 出纳员
- B. 会计
- C. 内部审计人员
- D. 以上都可以

正确答案:D

解析:

库存现金盘点可以由出纳员、会计或内部审计人员执行,以确保盘点的独立性和准确性。在某些情况下,可能需要多个角色共同参与,以提供交叉验证和增加盘点的可靠性。

现金盘点审计 (多选题 2.00分)

3. 盘点结束后, 审计人员应对盘点中出现的问题进一步审查, 分析库存现金盘盈或盘亏的原因, 以下处理正确的有 ()

- A. 对超限额保管的现金, 应建议及时送存银行
- B. 对出现的长、短款问题, 应查明原因和责任, 决不能因为是长款而放松审查
- C. 对尚未入账的符合财务制度规定的收、付凭证, 应按规定及时入账, 以保证会计信息的真实准确
- D. 对不符合财务制度规定的借条, 应要求追回款项。对白条开支, 应说明原因, 并要求其换取正式票据

正确答案: ABCD

5.2 银行存款科目审计

首先进行银行存款的实质性程序:

一、账表核对与余额验证

1、核对账表一致性:

获取银行存款余额明细表, 复核加计是否正确, 并与总账、日记账合计数核对、检查非记账本位币银行存款的折算汇率及金额是否正确

2、检查银行对账单与调节表
获取银行对账单:需加盖银行印章, 必要时亲自到银行获取或观察网银导出过程以验证真实性。核对调节表:检查未达账项(如“企付银未付”“银收企未收”), 确认期后是否入账, 关注长期未达账项是否存在挪用资金或舞弊。

二、函证程序(核心程序)

1、函证范围:

涵盖所有账户(含零余额、已注销账户)、借款、委托贷款及担保信息。需函证银行存款、借款余额及抵押/质押情况。

2、函证流程:

由会计师事务所直接发函并控制回函过程, 避免被审计单位干预。核对回函与账面记录, 差异需调查并记录。

三、发生额与交易测试

1、双向核对

将银行对账单与日记账逐笔核对, 双向验证(从对账单到日记账, 反之亦然)。使用数据分析技术检查大额、频繁或相同金额的异常流水

2、抽查大额收支
检查原始凭证(如合同、发票)是否齐全, 确认交易真实性及会计期间是否正确。关注非经营性资金转移(如关联方往来)。

四、特殊事项核查

1、受限存款

检查质押、冻结款项的证明文件(如质押合同), 确认是否恰当披露。核实境外存款的合规性及变现限制。

2、利息收入分析

计算应收利息与实际利息的差异, 评估合理性, 排查高息拆借或虚构存款。

3、截止测试

选取资产负债表日前后凭证, 检查跨期收支是否调整。

之后获取账面所有银行对账单及开户清单: 如下图所示



步骤1:银行存款实... 步骤2:获取银行对... 步骤3:账表核对与... 步骤4:函证程序 步骤5:发生额测试

任务描述:

为了验证工商银行余额的准确性, 你向财务主管获取了科目余额表, 其中银行存款相关的科目余额表金额如下:

科目名称	币别	期初借方余额	期初贷方余额	本期借方发生额	本期贷方发生额	本年借方累计	本年贷方累计	期末借方余额	期末贷方余额
银行存款		6,004,647.05		213,251,979.83	143,670,194.76	213,251,979.83	143,670,194.76	75,586,432.12	
中国工商银行	人民币	5,423,737.05		212,204,622.04	143,670,194.76	212,204,622.04	143,670,194.76	73,958,164.33	
中国建设银行	人民币	577,310.00		1,047,347.24		1,047,347.24		1,624,657.24	
中国农业银行	人民币	3,600.00		10.55		10.55		3,610.55	

- 1、据了解, 中国工商银行账户是安奈思的基本户, 其他银行账户是一般户, 据此, 填列银行存款明细表。
- 2、根据步骤2获取的银行对账单, 与科目余额表进行核对, 下载【银行存款明细表】, 并填写【银行存款明细表】对账单金额, 比较是否有差异。
- 3、根据填写明细表后的数据, 回答问题。

【银行存款明细表】: [银行存款明细表.xlsx](#)

查看银行存款明细表: 如下图所示:

序号	开户行	币种	账户性质	账号	期初金额		本期借方		本期贷方		期末金额		审计调整数		审定数		本期对账单金额	差异
					原币	本位币	原币	本位币	原币	本位币	原币	本位币	原币	本位币	原币	本位币		
合计					****	-	****	-	****	-	****	-	****	-	****	-	-	-
1	中国工商银行			6210252100338245959														
2	中国建设银行			6210365100985685627														
3	中国农业银行			6210368500125338628														
4																		

银行账户的币种和账户性质关系到本币转换核算和重分类调整。

若银行账户币种为外币，审计实质性程序中需要将外币余额通过审计年度期末的汇率转换为人民币进行核算。

若银行账户的账户性质为基本户和一般存款户，通过银行存款科目核算；若银行账户为保证金户等用途受限的银行账户（通常为企业向银行融资的保证金），则需要重分类调整至其他货币资金中。

2、银行存款明细填写

在明细表区域中，“期初余额”数字通常来源于企业提供的财务报表/科目余额表，“借/贷方发生额”则来自于序时账，不同的数据来源口径（企业方和银行方）进行交叉验证，最终通过底稿计算出期末余额是否和企业提供的“期末余额”一致，若有不一致、分类错误等问题则在“审计调整数”中进行调整，得出最终的“审定数”。

3、差异核对

由于审定数是经过企业口径、银行口径和审计师调整后的数据，需要和银行对账单的余额进行比较差异（不以企业的科目余额为基准是因为银行出具的对账单具有不可篡改性，第三方的证据可信度更高），出现差异需要说明差异来源。

货币资金实质性程序——填写银行存款明细表

银行存款明细表的作用是用于核对银行存款不同口径的期初期末余额、借方发生额和贷方发生额是否一致，以及确定银行账户的性质，判断是否需要进行调整。

1、银行账户性质确认

银行存款明细表			
序号	开户行	币种	账户性质
合计			
1	中国工商银行		
2	中国建设银行		
3	中国农业银行		
4			

银行账户的币种和账户性质关系到本币转换核算和重分类调整。

若银行账户币种为外币，审计实质性程序中需要将外币余额通过审计年度期末的汇率转换为人民币进行核算。

若银行账户的账户性质为基本户和一般存款户，通过银行存款科目核算；若银行账户为保证金户等用途受限的银行账户（通常为企业向银行融资的保证金），则需要重分类调整至其他货币资金中。

1. 安奈思工商银行账户：1、账户性质：___(1)___（根据【银行存款明细表】下拉选项填写）2、币种：___(2)___（根据【银行存款明细表】下拉选项填写）3、期末本位币金额：___(3)___（数值不带千分符，保留两位小数）4、本期对账单金额：___(4)___（数值不带千分符，保留两位小数）5、是否存在差异：___(5)___（是/否）（填空题 5）

正确答案：

1. 基本户
2. 人民币
3. 73958164.33
4. 73958164.33
5. 否

1. 安奈思建设银行账户：1、账户性质： ____ (1) ____ (根据【银行存款明细表】下拉选项填写) 2、币种： ____ (2) ____ (根据【银行存款明细表】下拉选项填写) 3、期末本位币金额： ____ (3) ____ (数值不带千分符，保留两位小数) 4、本期对账单金额： ____ (4) ____ (数值不带千分符，保留两位小数) 5、是否存在差异： ____ (5) ____ (是/否) (填空题 5)

正确答案：

- (1) 一般存款户
- (2) 人民币
- (3) 1624657.24
- (4) 1624657.24
- (5) 否

1. 安奈思农业银行账户：1、账户性质： ____ (1) ____ (根据【银行存款明细表】下拉选项填写) 2、币种： ____ (2) ____ (根据【银行存款明细表】下拉选项填写) 3、期末本位币金额： ____ (3) ____ (数值不带千分符，保留两位小数) 4、本期对账单金额： ____ (4) ____ (数值不带千分符，保留两位小数) 5、是否存在差异： ____ (5) ____ (是/否) (填空题 5)

正确答案：

- (1) 一般存款户
- (2) 人民币
- (3) 3610.55
- (4) 3610.55
- (5) 否

之后进行函证程序的调整查询：

步骤1:银行存款实... 步骤2:获取银行对... 步骤3:账表核对与... **步骤4:函证程序** 步骤5:发生额测试

任务描述：
银行函证程序

附 1
银行询证函 (格式一)
编号：

银行函证是审计中验证银行存款、借款等财务信息真实性的核心程序，其发函方式和注意事项需严格遵循监管要求及实务规范：

发函方式

1. 邮寄/快递发函

适用场景:外地银行、金额较小或偏远地区的银行。

操作要点:

注册会计师需亲自填写并寄出询证函，不可交由被审计单位代寄，保留快递单据作为底稿。

需核对银行官网公示的集中受理地址(如函证中心)，不可使用被审计单位提供的地址。

2. 审计人员跟函

适用场景:当地银行，需现场核实。

操作要点:

- 审计人员需全程控制函证过程,与企业财务人员共同至银行柜台办理,禁止由企业人员单独办理。
- 需携带身份证、介绍信等证件,对接银行对公柜台人员,避免与客户经理等非柜台人员接触。

3.电子函证

适用场景:支持数字化函证的银行(如通过区块链平台或银行网银)。

操作要点:

通过中国银行业协会区块链平台或银行指定电子渠道提交申请,需完成多重身份验证被审计单位需通过企业网银授权并支付手续费,银行在线生成回函。

注意事项

1.函证范围与内容

全覆盖原则:需函证所有账户(含零余额、已注销账户)、借款、担保及资金池业务。

格式选择:根据银行要求使用格式一或格式二(如交通银行仅接受格式二),不可留白或划斜线

2.印章与授权

预留印鉴:需加盖被审计单位公章、法人章及财务章(多页需骑缝章),并确认与银行预留印当。

授权控制:电子函证需被审计单位通过网银授权,避免代操作。

3.回函控制与验证

独立性:回函需直接寄至会计师事务所或由跟函审计人员带回,不可经企业转交

真实性核查:核对回函用章是否与银行公示一致,必要时联系银行确认。

4.时效与费用

回函时限:银行需在 10 个工作日内回复,若拒回需 3 日内通知事务所。

费用处理:确认扣款账户非外币或贷款户,避免因余额不足导致退函。

5.风险防范

异常交易排查:关注大额“一收一付”、长期未达账项,核实是否存在资金挪用或舞弊。D

历史案例警示:如康得新案中虚假回函问题,需强化对回函不符情形的追查。

三、特殊情形处理

资金池业务:需单独确认归集账户及调拨指令,核实余额是否包含归集资金。

注销账户:需函证期间内交易记录,避免遗漏未结清负债。

上图是银行函证的格式一标准模板,请回答下列问题:

1. 下列关于安奈思银行函证的发函程序,不正确的是 (单选题 1分)
- A. 发函之前,需核对开户清单中的账户是否都在账面列示,避免漏函。
 - B. 安奈思农行账户本年的银行流水发生额较小,风险低,可以不函证。
 - C. 安奈思的工行账户可以通过电子函证来验证余额的真实性。
 - D. 纸质函证在发函之前需到官网复核发函地址是否准确,并全程监督发函过程。

正确答案:B

试题解析:银行函证是银行存款科目审计的关键审计程序,需要函证所有银行账户并且本年注销的账户也需要函证,因此 B 错误。

2. 在函证安奈思的工商银行账户过程中, 下列做法正确的是 (单选题 1分)

- A. 函证地址使用安奈思财务主管提供的银行地址, 函证联系方式根据财务主管提供的资料填列。
- B. 对于工商银行账户无需函证的的项目留白即可。
- C. 工商银行回函地址误填至安奈思大厦导致回函邮寄至客户手上, 与客户联系取得函证以避免二次发函耽误时间。
- D. 工商银行的纸质函证采用一级分行集中受理的模式, 一份函证只列示一个函证基准日。

正确答案:D

试题解析:A 错误, 银行函证函证地址应需核对银行官网公示的集中受理地址(如函证中心), 不可使用被审计单位提供的地址。B:错误。对于工商银行账户无需函证的的项目, 要用斜线划掉, 不应留白。C:错误。银行函证回函需直接寄至会计师事务所或由跟函审计人员带回, 不可经企业转交。D:正确, 在银行函证资料中可以看出工商银行的纸质函证采用一级分行集中受理的模式, 一份函证只列示一个函证基准日。

3. 注册会计师执行银行函证程序时, 其无法被其他审计程序替代的本质功能是 (单选题 1分)

- A. 提供独立第三方对资金状况的直接确认
- B. 验证企业内部控制设计的合理性
- C. 发现未记录的负债或担保事项
- D. 确认银行存款账户余额的准确性

正确答案:A

试题解析:

选项 A(核心作用):银行函证的不可替代性在于其直接获取独立第三方(银行)的书面确认, 形成外部证据, 这是其他程序(如检查内部凭证、分析性复核)无法实现的。独立性是审计证据可靠性的基石, 而银行作为第三方提供的回函具有最高级别的客观性, 能直接支撑存在性、权利与义务等认定。

选项 B(干扰项):内部控制验证可通过穿行测试或控制测试完成, 无需依赖函证。

选项 C(次要作用):发现未记录负债虽是函证的作用之一, 但可通过检查合同、信用报告等替代程序实现, 非其独有功能。

选项 D(表面作用):余额准确性可通过对账单、调节表等内部证据验证, 但此类证据易被篡改, 函证的价值在于通过独立来源突破企业内部证据链的局限性。

最后进行由代码编写运行的的发生额测试:

银行存款审计需要对银行流水和序时账进行账账核对, 银行流水通常是按月出具, 账账核对需要先将各月流水合并为一张 excel, 再和序时账逐笔核对

步骤1:银行存款实...

步骤2:获取银行对...

步骤3:账表核对与...

步骤4:函证程序

步骤5:发生额测试

任务描述

银行存款审计需要对银行流水和序时账进行账账核对, 银行流水通常是按月出具, 账账核对需要先将各月流水合并为一张excel, 再和序时账逐笔核对

实例代码如下:

```

import pandas as pd

# 定义函数来处理数据
def extract_rows_by_keywords(input_file, output_file, column_index):
    # 定义包含所有月份名称的列表
    month_names = [f"{i}月" for i in range(1, 13)] # ['1月', '2月', ...,
'12月']

    # 创建一个空的列表用于存储符合条件的数据
    extracted_data = []

    # 遍历每个月份名称, 读取对应的 sheet 并根据关键词提取行
    with pd.ExcelFile(input_file) as excel:
        for month in month_names:
            if month in excel.sheet_names: # 检查是否存在该月份的 sheet
                sheet = pd.read_excel(excel, sheet_name=month)

                # 假设我们不知道列名, 但我们知道列的位置 (column_index)
                # 注意: pandas 的列索引是从 0 开始的
                try:
                    # 使用 str.contains 方法检查指定列中的每一行是否包含关键词
                    mask = sheet.iloc[:,
column_index].astype(str).str.contains('转账|现金')
                    filtered_sheet = sheet[mask]

                    if not filtered_sheet.empty:
                        filtered_sheet['Month'] = month # 添加月份信息
                        extracted_data.append(filtered_sheet)
                    except IndexError:
                        print(f"警告: 工作表 '{month}' 中没有第 {column_index
+ 1} 列。")
                    except Exception as e:
                        print(f"在工作表 '{month}' 中发生错误: {e}")
                else:
                    print(f"警告: 没有找到名为 '{month}' 的工作表。")

    # 合并所有符合条件的数据到一个新的 DataFrame
    if extracted_data:
        merged_data = pd.concat(extracted_data, ignore_index=True)
    else:
        print("没有找到任何符合条件的行。")
        return

```

```

if len(new_headers) != len(merged_data.columns):
    print("新表头数量与数据列数不匹配，请检查表头定义。")
    return

# 设置新表头
merged_data.columns = new_headers
merged_data.to_excel(output_file, index=False)

# 使用定义的函数
input_excel_path = r"[DataFolderPath]/工商银行对账单.xlsx" # 输入 Excel 文件路径
output_excel_path = r"[UserFolderPath]/工商银行对账单合并.xlsx" # 输出 Excel 文件路径
column_index = 1 # 要检查的列的位置（假设是第二列）
new_headers = ['日期',
               '业务产品种类',
               '对方户名',
               '摘要',
               '借方发生额',
               '贷方发生额',
               '余额',
               '月份']

extract_rows_by_keywords(input_excel_path, output_excel_path,
                        column_index)

```

5.3 其他货币资金科目审计

进行其他货币资金的实质性程序审计：

The screenshot shows a software interface for auditing. The top navigation bar has two tabs: 'Step 1: Other Monetary Funds' (highlighted) and 'Step 2: Third Party Payments...'. The left sidebar contains a 'Task Navigation' section with a dropdown menu set to 'Monetary Funds Substantive Procedures'. Below it, 'Task Practice' includes three items: '1. Audit of Cash and Cash Equivalents', '2. Audit of Bank Deposit Accounts', and '3. Audit of Other Monetary Funds' (highlighted with a red box). The main content area shows the 'Task Description' for 'Other Monetary Funds Substantive Audit Procedures'. It states that substantive audit procedures should be designed based on specific types (e.g., guarantee deposits, investment deposits, foreign deposits) to ensure the authenticity, integrity, and compliance of the funds. The first step is '1. Audit of Guarantee Deposits (such as letters of credit, bank acceptance bill guarantees)', which includes sub-steps: '1. Check agreements and approval documents' and '2. Check accounts and transactions'. A note mentions that if the guarantee ratio is abnormal (e.g., 100% pledge), an investigation is needed to see if there is hidden collateral or fund occupation.

实质性审计程序需根据其具体类型(如保证金存款、存出投资款、外埠存款等)设计差异化的审计步骤，以确保资金真实性、完整性及合规性。保证金存款(如信用证、银行承兑汇票保证金)

1、检查协议与审批文件

获取开立银行承兑汇票或信用证的协议、银行授信审批文件，核对保证金比例是否符合合同约定异常关注：若保证金比例异常(如 100%质押)，需调查是否存在隐性担保或资金占用 2、核对账户与交易

将保证金账户对账单与应付票据规模对比，验证保证金金额合理性检查保证金与对应债务的匹配性，特别关注“有保证金无对应保证事项”的异常情形 3、函证与受限核查

函证保证金余额及受限状态，确认是否被质押或冻结

对已质押保证金，检查存单复印件与质押合同的一致性，并确认借款是否入账二、存出投资款(如证券、基金投资)

1、追踪资金流向

获取董事会决议、开户资料及操作授权文件，核实投资款的审批流程结合金融资产审计，核对证券账户户名是否与被审计单位一致

2、验证交易真实性

获取证券公司交易流水，抽查大额收支，核对账面记录与流水是否相符风险提示：若资金未专户管理或亏损严重，可能需重分类为“其他应收款”并计提减值。

三、外埠存款(异地临时账户)

1、核实开户与用途

检查外埠存款的审批文件及用途说明，确认与经营需求是否匹配函证账户余额及交易流水，关注资金划转是否与采购/工程等业务相关

2、检查原始凭证

抽查大额支出凭证(如采购合同、发票)，验证资金使用的合规性

四、信用卡存款与第三方支付平台资金

1、信用卡存款

核对信用卡持有人身份(如是否为员工)，获取持卡人书面确认检查信用卡对账单，验证消费记录与业务活动的关联性

2、第三方支付平台(如支付宝、微信)-案例获取平台协议及账户信息，验证余额及交易流水真实性替代函证：若平台不支持函证，需观察被审计单位人员操作导出数据的过程，核对界面信息与账面记录五、定期存款(非质押类)

1、检查原件与监盘

对未质押定期存款，检查开户证实书原件(非复印件)，核对存款人、金额、期限监盘定期存单凭据，特别关注资产负债表日的大额存单 2、利息收入分析

测算应计利息与实际入账利息的差异，识别是否存在体外循环或虚构存款。

了解其他货币资金实质性程序审计后，学习第三方支付过程审计：

任务描述:

观察安奈思支付宝流水并回答以下问题



阅读材料回答如下问题:

1. 审计过程中你发现, 安奈思的支付宝账户由员工个人名义注册但用于公司业务, 你应优先 (单选题 1分)
- A. 直接认可该账户为公司资产
 - B. 要求企业提供书面授权证明
 - C. 视为舞弊迹象并扩大审计范围
 - D. 忽略该账户的审计

正确答案:C

试题解析:个人账户用于公司业务可能涉及资金挪用或隐匿交易, 需核查背景并扩大审计程序。

2. 以下哪项不属于支付宝审计的必备程序? (单选题 1分)
- A. 获取支付宝账户的开户协议
 - B. 函证支付宝账户余额及交易流水
 - C. 检查支付宝账户的理财产品收益
 - D. 核对支付宝流水与账面记录

正确答案:C

试题解析:理财产品收益属于投资活动, 与支付宝账户本身的审计无直接关系, 除非支付宝余额直接用于购买理财

3. 观察安奈思的支付宝流水,可以发现其2020年期末余额为 () 元。 (单选题 1分)

- A. 4723256.65
- B. 5240550.56
- C. 6661700.56
- D. 5335550.56

正确答案:B

试题解析:由支付宝收支明细可以看出,2020/12/31 安奈思支付主期末余额为 5240550.56 元

日期	金额	摘要	对方户名	余额	币种
2020/12/23 21:55:20	290,000.00	提现		5,656,225.56	支付宝
2020/12/23 21:55:20		提现	500,000.00	5,156,225.56	支付宝
2020/12/25 9:01:06		提现	500,000.00	4,656,225.56	支付宝
2020/12/26 12:15:13	350,000.00	提现		5,006,225.56	支付宝
2020/12/27 13:29:25	310,000.00	提现		5,316,225.56	支付宝
2020/12/29 10:09:03		提现	300,000.00	5,016,225.56	支付宝
2020/12/30 18:03:54	524,325.00	提现		5,540,550.56	支付宝
2020/12/31 16:25:32		提现	300,000.00	5,240,550.56	支付宝
2021/01/01 8:59:32	95,000.00	提现		5,335,550.56	支付宝

第 6 课时 步骤 6: 审计结论

6.1 分析性程序

货币资金--银行存款分析性程序

根据前面通过 python 合并后生成的工商银行流水,对安奈思工商银行余额和发生额进行分析性程序:

- ①通过折线图关联日期和余额,并命名为《2020 年安奈思工商银行每日余额变动趋势图》分析 2020 年的余额波动情况和趋势
- ②通过折线图关联贷方发生额,并命名为《2020 年安奈思工商银行支出结构图》,分析 2020 年的流水支出的对方户名构成。

货币资金——银行存款分析性程序

[操作指引](#)

根据前面通过python合并后生成的工商银行流水,对安奈思工商银行余额和发生额进行分析性程序:

- ①通过折线图关联日期和余额,并命名为《2020年安奈思工商银行每日余额变动趋势图》分析2020年的余额波动情况和趋势
- ②通过折线图关联贷方发生额,并命名为《2020年安奈思工商银行支出结构图》,分析2020年的流水支出的对方户名构成

操作指引如下:

明确目的:此次使用 BI 来进行数据分析的目的是:

直观体现安奈思工商银行余额 2020 年整体的变动趋势和支出的构成结构

分析要素:

变动趋势:折线图,支出构成:饼图

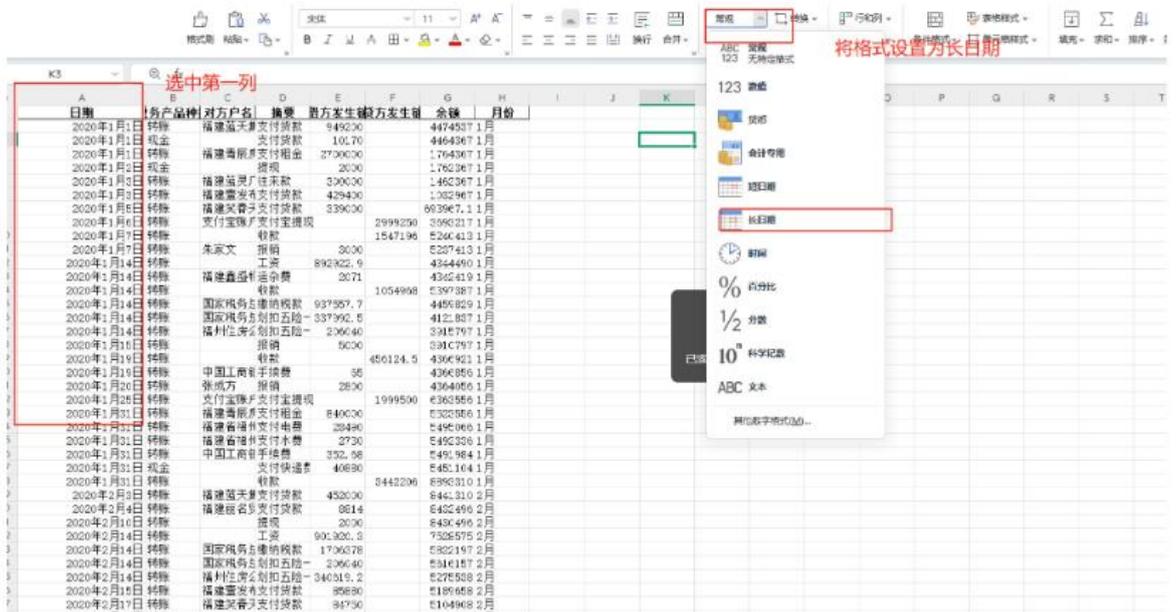
细节思考:

- 1、将原数据数据透视会将某日含多个余额的数据加总，并非我们想要的每日余额表，因此需
要将原数据加工成一个能直接看出每日余额的数据。
- 2、为了直观体现当日数据余额增长的原因，折线图的辅助 y 轴可以新增一个贷方发生额，直观
体现某日余额增长的原因。
- 3、为了整体可视化效果，需建立表格让数据随图形变化，字体颜色格式需进行调整。

每日余额

目标: 将【原数据】加工为呈现每日余额的表格

- 1、将 A 列的格式改为长日期的格式



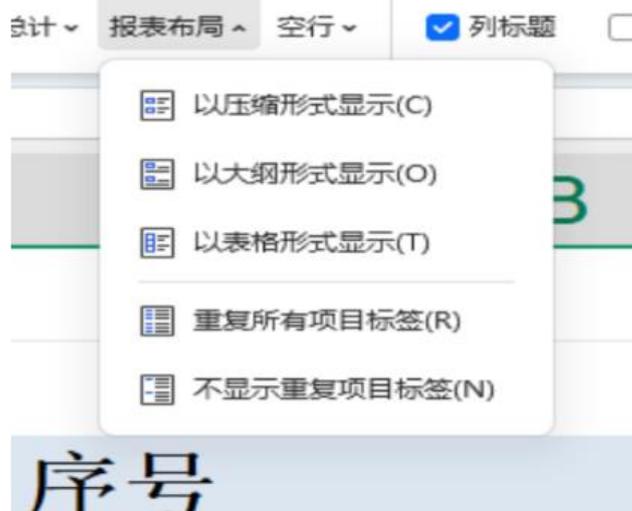
- 2、将【序号】列插入第一列，给每行排列序号，即每日的最后一行是当日余额

序号	日期	业务产品种类	对方户名	摘要	借方发生额	贷方发生额	余额	月份
1	2020年1月1日	转账	福建富天集团有限公司	支付贷款	949,200.00		4474537.05	1月
2	2020年1月1日	现金		支付贷款	10,170.00		4464367.05	1月
3	2020年1月1日	转账	福建青辰房地产经纪有限公司	支付租金	2,700,000.00		1764367.05	1月
4	2020年1月2日	现金		提现	2,000.00		1762367.05	1月
5	2020年1月3日	转账	福建益民广告有限公司	往来款	300,000.00		1462367.05	1月
6	2020年1月3日	转账	福建晋发药业有限公司	支付贷款	429,400.00		1032967.05	1月
7	2020年1月5日	转账	福建晋发药业有限公司	支付贷款	339,000.00		693967.05	1月
8	2020年1月6日	转账	支付宝账户	支付宝提现		2,999,250.00	3693217.05	1月
9	2020年1月7日	转账		收款		1,547,196.00	5240413.05	1月
10	2020年1月7日	转账	朱家文	报销	3,000.00		5237413.05	1月
11	2020年1月14日	转账		工资	892,922.88		4344490.19	1月
12	2020年1月14日	转账	福建鑫康物业管理有限公司	违约金	2,071.00		4342419.19	1月
13	2020年1月14日	转账		收款		1,054,968.00	5397387.19	1月
14	2020年1月14日	转账	国家税务总局福州市仓山区税务局	缴纳税款	937,557.72		4459829.47	1月
15	2020年1月14日	转账	国家税务总局福州市仓山区税务局	划扣五险一金	337,992.51		4121836.96	1月
16	2020年1月14日	转账	福州住房公积金管理中心	划扣五险一金	206,040.00		3915796.96	1月
17	2020年1月15日	转账		报损	5,000.00		3910796.96	1月
18	2020年1月19日	转账		收款		456,124.50	4366921.46	1月
19	2020年1月19日	转账	中国工商银行对公账户	手续费	65.00		4366856.46	1月
20	2020年1月20日	转账	张佩方	报销	2,800.00		4364056.46	1月
21	2020年1月25日	转账	支付宝账户	支付宝提现		1,999,500.00	6363556.46	1月
22	2020年1月31日	转账	福建青辰房地产经纪有限公司	支付租金	840,000.00		5523556.46	1月
23	2020年1月31日	转账	福建省福州市供电有限公司	支付电费	28,490.00		5495066.46	1月
24	2020年1月31日	转账	福建省福州市自来水公司	支付水费	2,730.00		5492336.46	1月
25	2020年1月31日	转账	中国工商银行对公账户	手续费	382.68		5491953.78	1月
26	2020年1月31日	现金		支付快递费	40,880.00		5451073.78	1月
27	2020年1月31日	转账		合计		2,447,706.00	8098779.78	1月

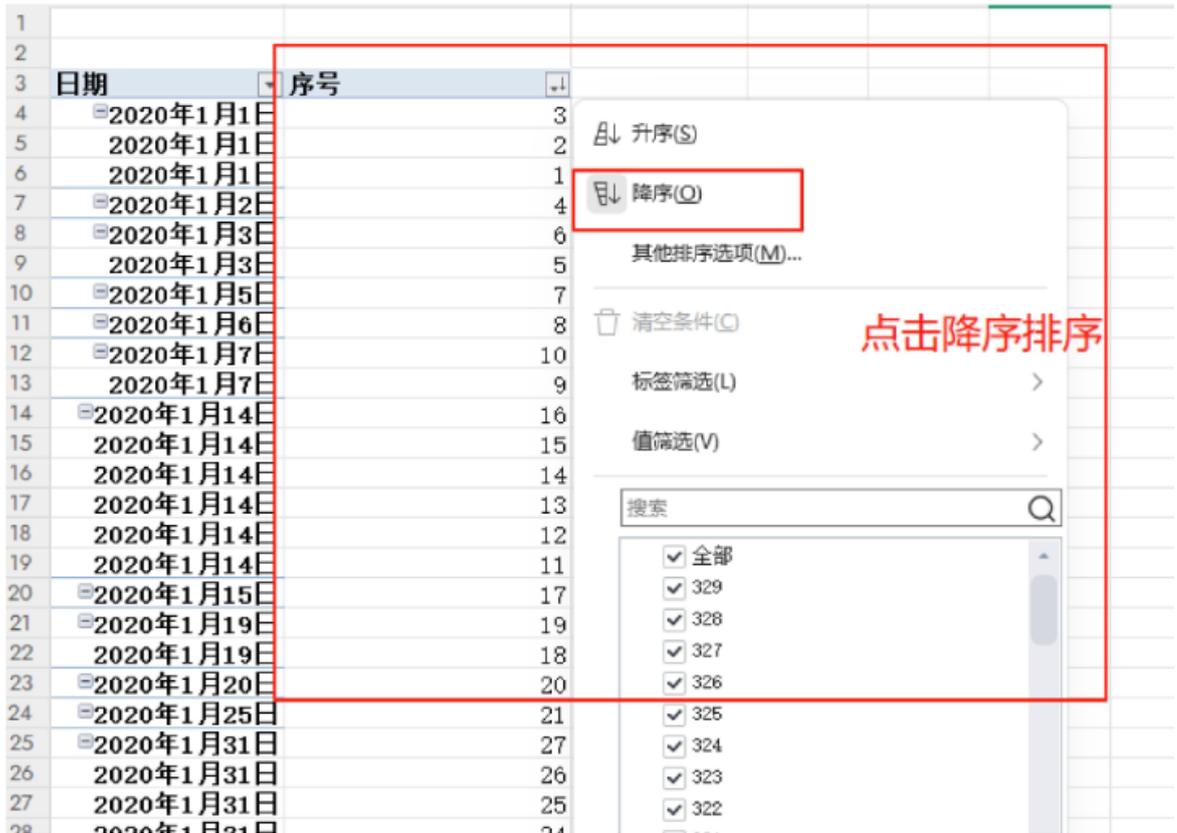
3、对加工后的表格进行数据透视，将数据透视后的表命名为【数据透视】，并将【日期】【序号】添加至行。

日期	序号
2020年1月1日	3
2020年1月1日	2
2020年1月1日	1
2020年1月2日	4
2020年1月3日	6
2020年1月3日	5
2020年1月5日	7
2020年1月6日	8
2020年1月7日	10
2020年1月7日	9
2020年1月14日	16
2020年1月14日	15

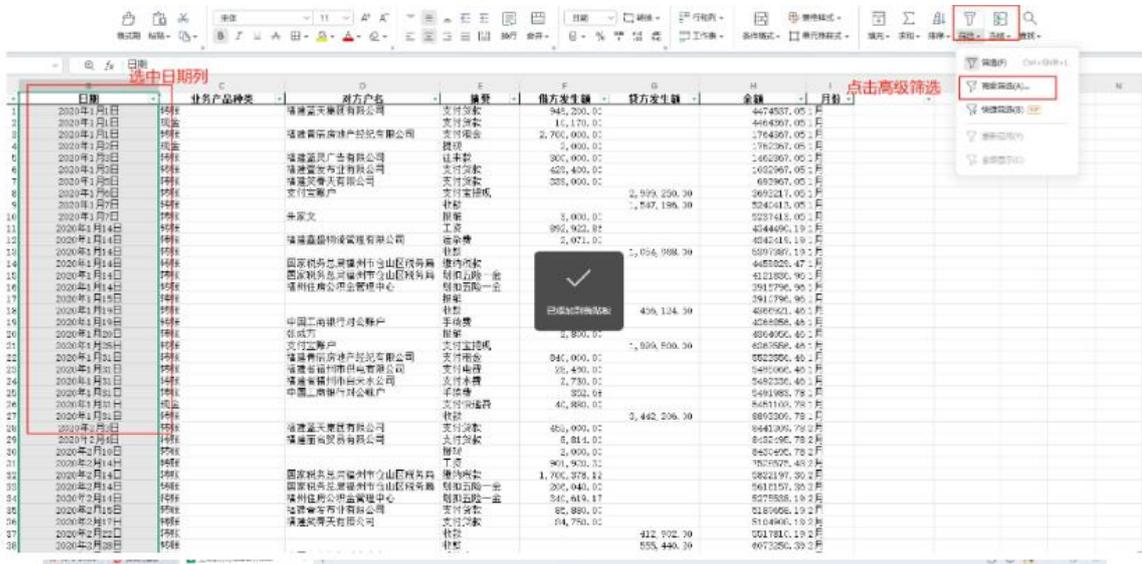
4、点击【设计】，对数据透视报表布局进行调整，选择以【表格形式显示】和【重复所有项目标签】。

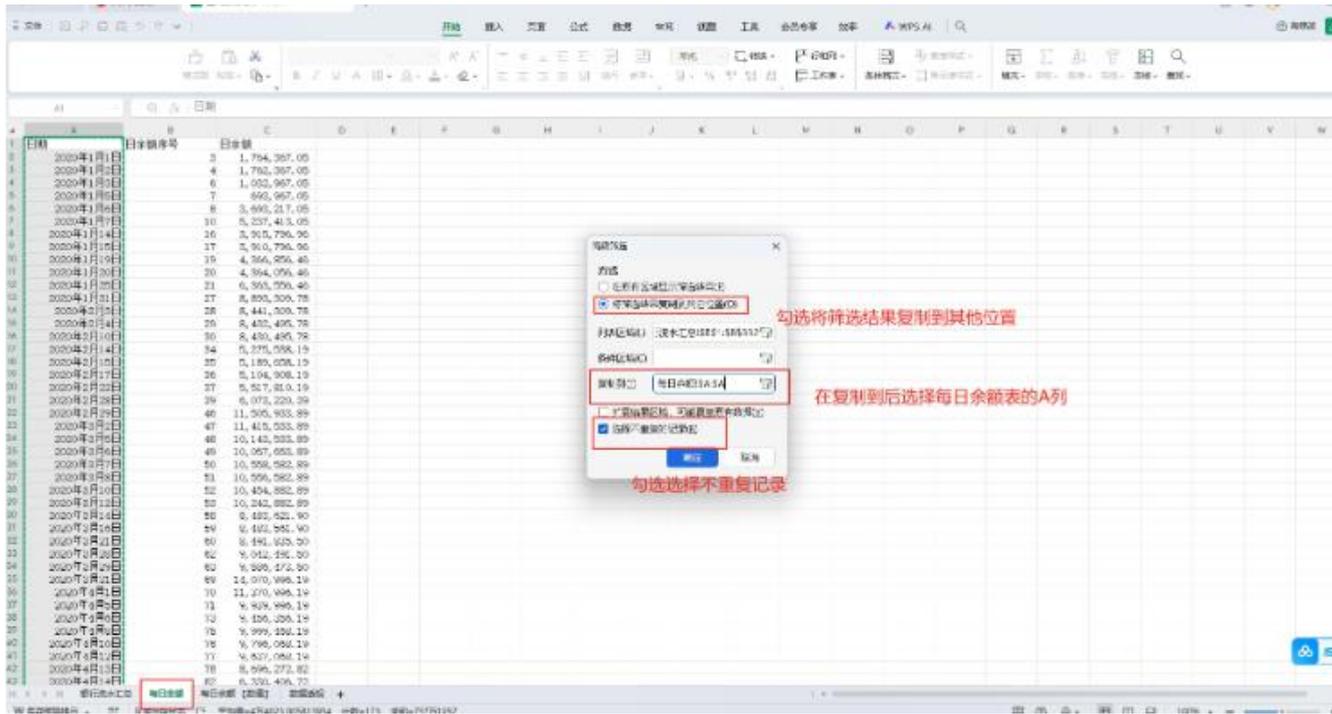


5、对数据透视表的序号列进行降序排序

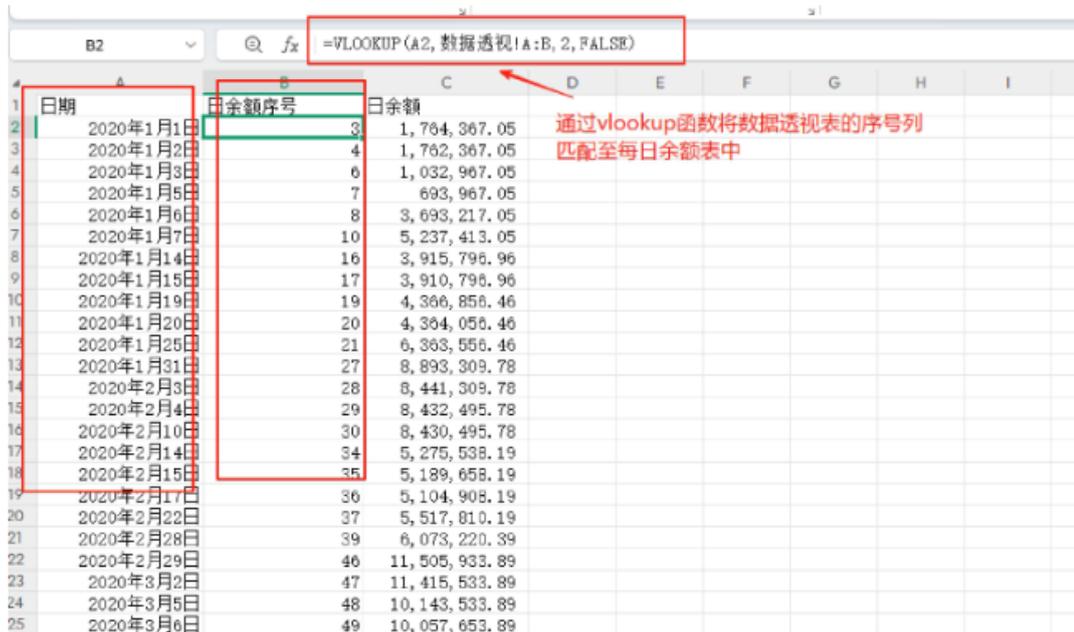


6、新建 sheet 并命名为【每日余额】，将【原数据】中的日期去除重复值并粘贴至【每日余额】表中。





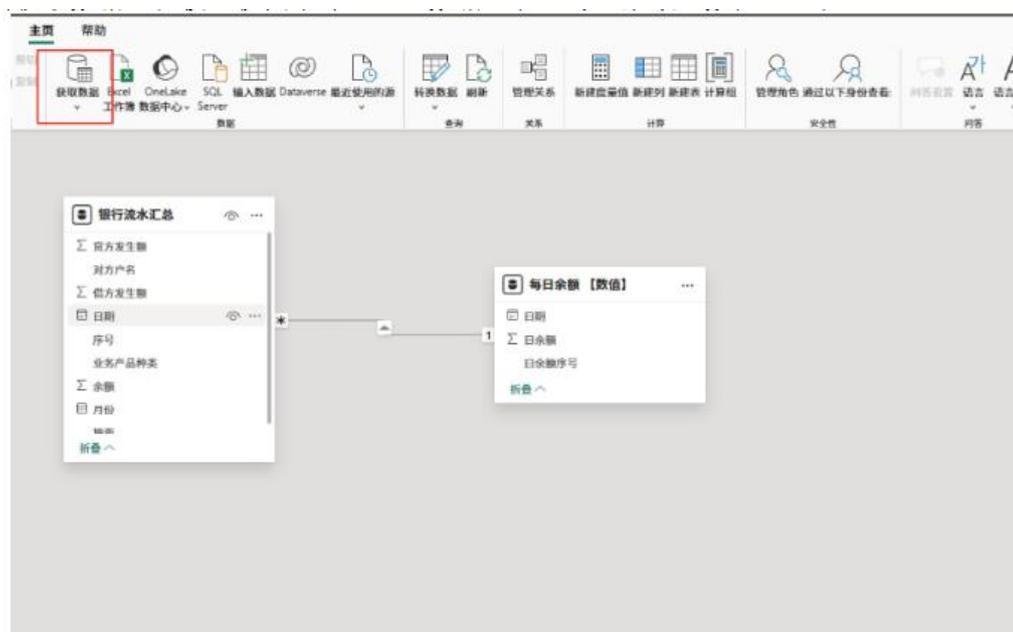
7、新建【日余额序号】列，通过 vlookup 函数以【日期】为匹配项，将【数据透视】表的【序号】匹配至【每日余额】表的【日余额序号】列。



8、新建【日余额】列，通过 vlookup 函数以【日余额序号】为匹配项，将【原数据】表的【余额】匹配至【每日余额】表的【日余额】列。

BI 可视化

1、点击获取数据，在模型视图中导入【原数据】和【每日余额【数值】】表。



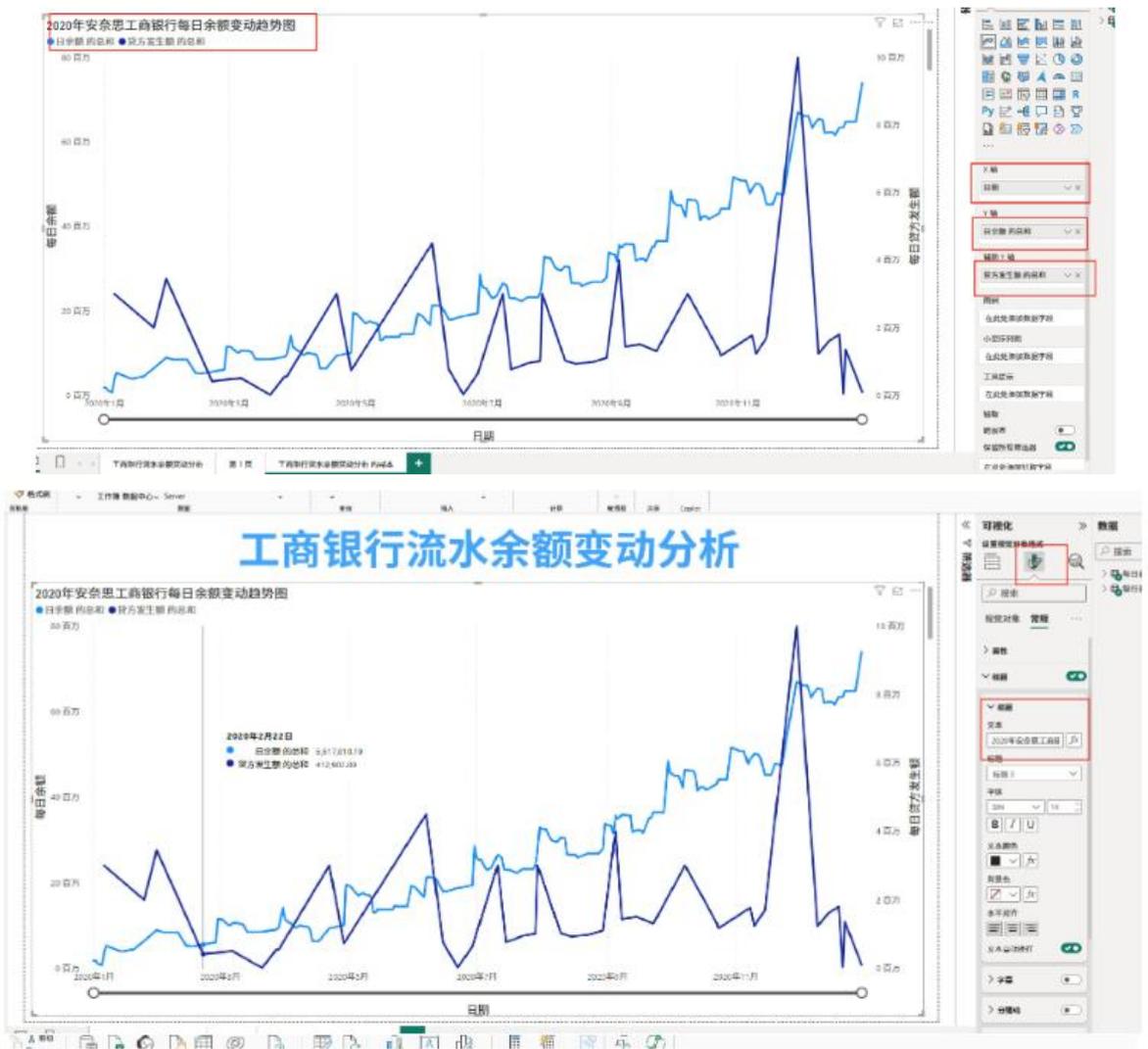
2、点击管理关系，新关系，以日期为纽带建立新关系

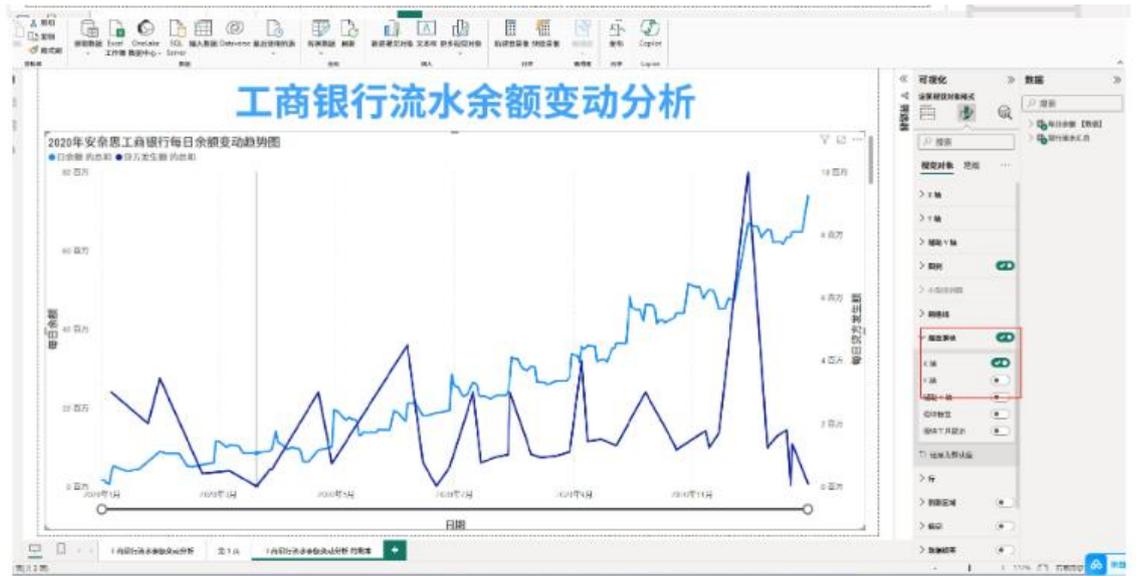


3、切换至表格视图，检查表格视图中的两个表的【日期】列格式是否都是日期格式，如果不是日期格式，需要手动调回日期格式。

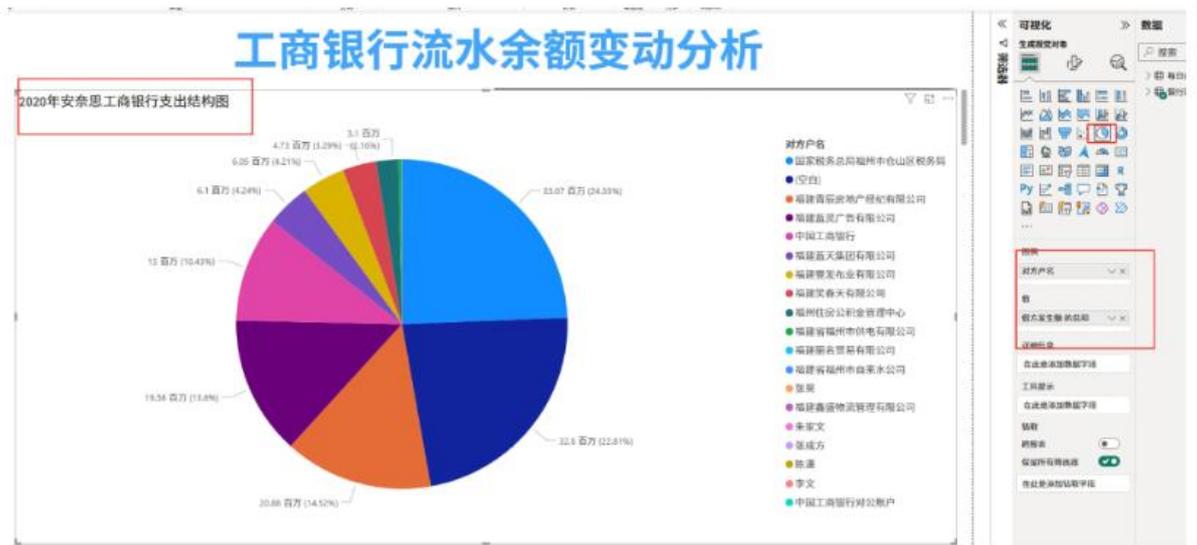
序号	日期	业务产品种类	对方户名	摘要	借方发生额	贷方发生额	余额	月份
1	2020年7月1日	转账	福建益天美鞋鞋有限公司	支付货款	948200		4474537.05	2020年7月
2	2020年7月1日	转账	福建益天美鞋鞋有限公司	支付租金	2700000		1764687.05	2020年7月
3	2020年7月1日	转账	福建益天美鞋鞋有限公司	往来款	300000		1462687.05	2020年7月
4	2020年7月1日	转账	福建益天美鞋鞋有限公司	支付货款	428400		1032967.05	2020年7月
5	2020年7月1日	转账	福建益天美鞋鞋有限公司	支付货款	338000		694967.05	2020年7月
6	2020年7月1日	转账	李发文	货款	3000		527433.05	2020年7月
7	2020年7月4日	转账	工业	货款	651922.85		494480.19	2020年7月
8	2020年7月4日	转账	福建益天美鞋鞋有限公司	运费	2071		492409.19	2020年7月
9	2020年7月4日	转账	国家税务总局福州台山区税务局	缴纳税金	997537.22		4459623.47	2020年7月
10	2020年7月4日	转账	国家税务总局福州台山区税务局	缴纳五险一金	307982.51		4121836.96	2020年7月
11	2020年7月4日	转账	福州住房公积金管理中心	缴纳五险一金	208240		3913796.36	2020年7月
12	2020年7月5日	转账	货款	货款	5000		3910796.36	2020年7月
13	2020年7月9日	转账	中国工商银行对公账户	手续费	45		498885.46	2020年7月
14	2020年7月9日	转账	货款	货款	2000		496405.46	2020年7月
15	2020年7月9日	转账	福建益天美鞋鞋有限公司	支付租金	840000		532355.46	2020年7月
16	2020年7月9日	转账	福建益天美鞋鞋有限公司	支付电费	28480		5489385.46	2020年7月
17	2020年7月9日	转账	福建益天美鞋鞋有限公司	支付电费	2730		5482236.46	2020年7月
18	2020年7月9日	转账	中国工商银行对公账户	手续费	352.68		5481883.78	2020年7月
19	2020年7月9日	转账	福建益天美鞋鞋有限公司	支付货款	452000		8441208.78	2020年7月
20	2020年7月9日	转账	福建益天美鞋鞋有限公司	支付货款	3000		8444208.78	2020年7月

4、新建折线图并添加相关字段，更正标题，可以增加缩放滑块方便查看





5、新建饼状图并添加相关字段，更正标题



6、新建表格，并调整样式预设，让整体更加美观

冰余额变动分析

日期	日余额	贷方发生额 的总和	借方发生额 的总和
2020年1月1日	1,764,367.05		2,700,000.00
2020年1月2日	1,762,367.05		2,000.00
2020年1月3日	1,032,967.05		429,400.00
2020年1月5日	693,967.05		339,000.00
2020年1月6日	3,693,217.05	2,999,250.00	
2020年1月7日	5,237,413.05		3,000.00
2020年1月14日	3,915,796.96		206,040.00
2020年1月15日	3,910,796.96		5,000.00
2020年1月19日	4,366,856.46		65.00
2020年1月20日	4,364,056.46		2,800.00
2020年1月25日	6,363,556.46	1,999,500.00	
2020年1月31日	8,893,309.78	3,442,206.00	
2020年2月3日	8,441,309.78		452,000.00
2020年2月4日	8,432,495.78		8,814.00
2020年2月10日	8,430,495.78		2,000.00
2020年2月14日	5,275,538.19		340,619.17
2020年2月15日	5,189,658.19		85,880.00
2020年2月17日	5,104,908.19		84,750.00
2020年2月22日	5,517,810.19	412,902.00	
2020年2月28日	6,073,220.39		30.00
2020年2月29日	11,505,933.89		182.50
2020年3月2日	11,415,533.89		90,400.00
总计		69,632,636.99	77,679,539.08

可视化
数据

生成视觉对象

搜索

每日

银行

列

- 日期
- 日余额
- 贷方发生额 的总和
- 借方发生额 的总和

钻取

跨报表

保留所有筛选器

在此处添加钻取字段

冰余额变动分析

日期	日余额	贷方发生额 的总和	借方发生额 的总和
2020年1月1日	1,764,367.05		2,700,000.00
2020年1月2日	1,762,367.05		2,000.00
2020年1月3日	1,032,967.05		429,400.00
2020年1月5日	693,967.05		339,000.00
2020年1月6日	3,693,217.05	2,999,250.00	
2020年1月7日	5,237,413.05		3,000.00
2020年1月14日	3,915,796.96		206,040.00
2020年1月15日	3,910,796.96		5,000.00
2020年1月19日	4,366,856.46		65.00
2020年1月20日	4,364,056.46		2,800.00
2020年1月25日	6,363,556.46	1,999,500.00	
2020年1月31日	8,893,309.78	3,442,206.00	
2020年2月3日	8,441,309.78		452,000.00
2020年2月4日	8,432,495.78		8,814.00
2020年2月10日	8,430,495.78		2,000.00
2020年2月14日	5,275,538.19		340,619.17
2020年2月15日	5,189,658.19		85,880.00
2020年2月17日	5,104,908.19		84,750.00
2020年2月22日	5,517,810.19	412,902.00	
2020年2月28日	6,073,220.39		30.00
2020年2月29日	11,505,933.89		182.50
2020年3月2日	11,415,533.89		90,400.00
总计		69,632,636.99	77,679,539.08

可视化
数据

搜索

每日余额【数值】

银行流水汇总

视觉对象 常规

样式库

样式

添加标题的右格式行

- 默认值
- 无
- 差异最小
- 加粗标题
- 交替行
- 具有对比度的交替行
- 右格式行
- 带加粗标题的右格式行
- 稀疏

单元格元素

URL 链接

图像大小

7、插入文本框在上方，并调整字体颜色和粗细，使可视化数据面板更加美观



8、调整面板布局，使整体呈现更加直观



评分标准如下：

评分标准	
操作步骤	分值
①通过数据加工获得每日余额作为BI可视化数据源	6
②可视化：创建折线图，标题为：《2020年安奈思工商银行每日余额变动趋势图》	3
②可视化：创建饼图，标题为：《2020年安奈思工商银行支出结构图》	3
③建立表格让数据随图形变化，调整字体颜色和粗细	3
总计	15

参考格式如下：



进行 BI 的实质性分析后，根据完成内容做题：

银行流水分析性程序 (单选题 1.00分)

1. 根据安奈思2020年的工商银行流水余额变动情况，下面说法错误的是

- A. 安奈思2020年日余额呈现波动上升的趋势
- B. 安奈思2020年11月30日银行余额出现大幅增长是由于当日流入1千万的银行流水
- C. 安奈思工商银行的贷方发生额表示银行存款金额减少，借方表述银行存款的增加
- D. 安奈思2020年工商银行支出构成中对国家税务总局福州市仓山区税务局的支出占比最高

正确答案:C

解析:

银行流水的贷方发生额表示银行存款金额增加，借方表述银行存款的减少，与账面相反

2. BI工具中的哪种可视化图表最适合展示“某账户月度资金流入流出趋势”？

- A. 饼图
- B. 双轴折线图
- C. 散点图
- D. 热力图

正确答案:B

解析:双轴折线图可清晰对比不同时间点的流入/流出金额及波动趋势(如突增突减提示异常)。饼图(选项A)适用于比例分析;散点图(选项C)用于相关性分析;热力图(选项D)适合地域分布展示。

3. 验证企业“存贷双高”合理性时, BI工具相比SQL的不可替代性是

- A. 执行数据库查询更快
- B. 动态展示资金与利息收益趋势对比
- C. 灵活修正原始财务数据
- D. 生成会计分录凭证

正确答案:B

解析:

BI工具(如Power BI)可联动分析存款规模贷款规模及利息收入趋势(如双轴折线图)直观暴露“高存款低收益”的舞弊迹象;SQL仅能输出表格数据,缺乏可视化能力。

6.2 各环节审计结论

款低收益”的舞弊迹象;SQL仅能输出表格数据,缺乏可视化能力。



根据之前的审计过程发现:

安奈思公司在内控方面, 月底部分审批流程未完成, 没有盖章。实质性程序部分, 安奈思支付宝账户以老板个人名义注册用于公司业务, 可能涉及舞弊风险, 已经要求企业提供支

付宝账户的完整交易流水与账面逐笔核对，并检查支付宝账户与公司对公账户、其他关联账户之间的转账记录，要求企业提供相关交易凭证核实交易真实性。



在最后进行审计意见时要都写入管理建议书中。

6.3 审计小结-管理建议书的撰写



结合整个审计流程所学知识以及财务相关内容，填写对安奈思 2020 年内控管理建议。

下载答题卡如下：

管理建议书		
<p>我们接受委托，对【安奈思股份有限公司】（以下简称“【安奈思】”）及下属企业2020年度会计报表进行审计。在审计过程中，我们调查测试了贵公司及下属企业与年度审计相关的内部控制的有效性。我们的测试和评价遵循了中国注册会计师执业准则的有关规定。贵公司及下属企业对所提供的会计报表负责，我们的责任是对会计报表发表审计意见，而不是对内部控制整体发表意见和提供保证。我们提供管理建议书的目的，是针对我们在审计过程中注意到的、可能导致会计报表发生重大错报和漏报的内部控制重大缺陷提出改进建议。管理建议书中提到的内部控制存在不足的地方，仅是我们注意到的，不应被视为对内部控制发表的鉴证意见，所提建议不具有强制性和公正性，仅供贵公司及下属企业管理当局完善内部控制、加强相关内部管理参考之用。</p>		
科目	具体内容	管理建议
货币资金	月底部分审批流程未完成，签章不全。	
	支付宝账户以股东个人名义注册用于公司业务。	

做完建议书的撰写后回答有关管理建议书的问题：

管理建议书的性质特征 （单选题） 1.00分

1. 管理建议书的性质特征

- A. 具有法律强制效力的鉴证文件
- B. 审计过程中发现的内部控制缺陷的改进建议
- C. 对被审计单位财务报表的补充说明
- D. 注册会计师的法定审计责任证明

正确答案:B

解析:管理建议书本质是审计的“副产品”，仅针对内控薄弱环节提出建议，不具强制性和鉴证功能。选项 A、D 混淆了其与审计报告的法律属性;选项 C 错误(与财务报表无直接关联)。

2. 关于管理建议书的使用范围, 正确的是 ()。

- A. 可作为投资者决策依据
- B. 仅供被审计单位管理层内部参考
- C. 需向监管部门备案
- D. 具有对外公证效力

正确答案:B

解析:管理建议书明确限走为内部参考不对外报送, 且需注明“因使用不当造成的后果与注册会计师无关”。选项 A、C、D 均违背其非公开、非鉴证的特性。

管理建议书对注册会计师的价值

(单选题 1.00分)

3. 注册会计师借助管理建议书可间接实现 ()。

- A. 减少审计程序工作量
- B. 降低自身法律责任风险
- C. 替代内部控制测试
- D. 强制企业整改缺陷

正确答案:B

解析:

管理建议书通过提前揭示内控缺陷, 证明注册会计师已履行提示义务, 从而减少因企业后续舞弊导致的连带责任。选项 A 错误(不减少工作量);选项 C、D 违背其非强制性本质。

产业案例 审计法治与监督效能

任务一：初步业务活动

步骤一：了解被审计单位背景

1.1.1 审前调查了解

单击步骤导航栏中的“了解被审计单位背景”，点击左上方“步骤 1：审前调查了解”，观看场景对话，获取相关信息，如图 2-1-1 所示。

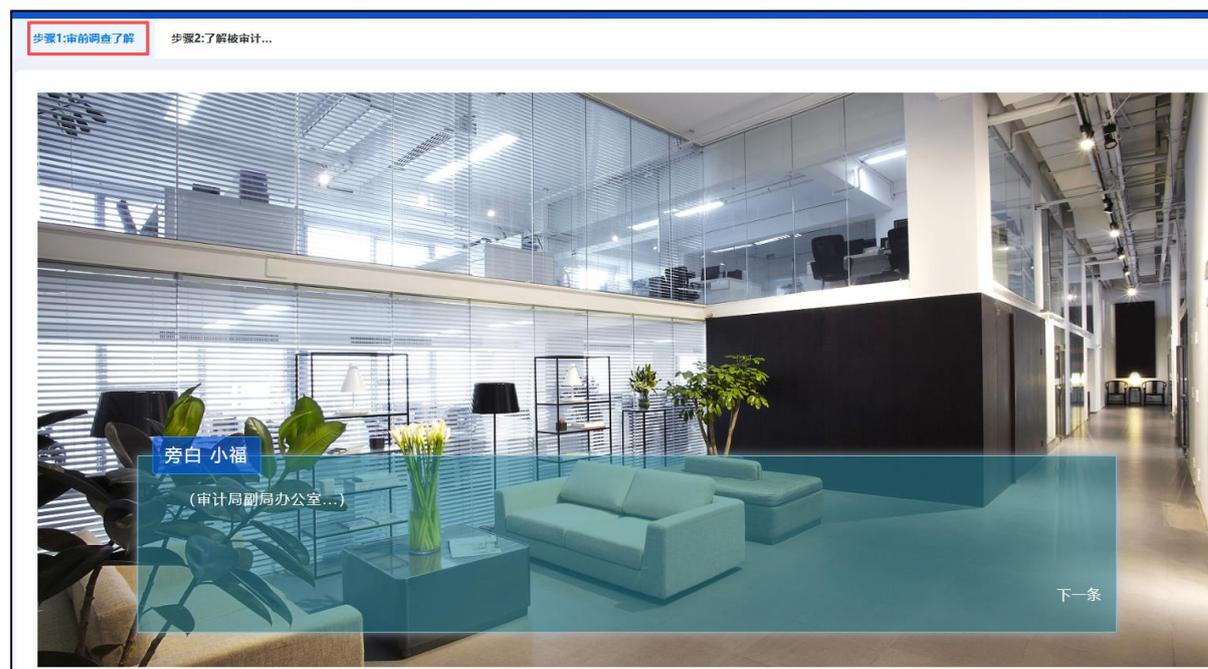


图 2-1-1 了解被审计单位及其环境

根据对话内容，以下是对被审计单位——大连天源贝类养殖股份有限公司及其环境的了解：

1.经济环境

扇贝养殖行业养殖产量逐渐增加，海水养殖成为主要方式；贝类产值不断增加，养殖收入逐步提高；品种结构基本形成，主产品产量持续增长。

2.被审计企业的行业状况、法律环境、监管环境以及其他外部因素

贝类产业在供应优质蛋白、优化居民饮食结构、提高渔民收入、增加外汇收入与改善近海环境等方面发挥了重要作用；贝类养殖也面临着空间制约、水域污染、资源衰退、食品安全等诸多问题；各项监管措施加大力度科学调整养殖产业结构，修复养殖海域生态环境。

3.被审计企业的所有权结构、治理结构和组织结构

大连天源贝类养殖股份有限公司由大连市人民政府国有资产监督管理委员会 100%控股；设股东大会、董事会、监事会，监事会下设总经理及各管理部门，董事会下设审计委员会；设有总经办、采购部、销售部、行政部、财务部、仓储部门、养殖部门、质量控制部门。

4.被审计企业财政、财务隶属关系或者国有资产监督管理关系

大连市人民政府国有资产监督管理委员会下属企业。

5.被审计企业的战略目标和相关经营

2020 年战略目标是提高扇贝产量、提升生产能力，经济收入持续增长

6.被审计企业的经营范围、业务性质、经营情况和关联方交易

经营范围包括许可项目：水产苗种生产，水产养殖，渔业捕捞，食品经营，食品互联网销

售（销售预包装食品），食品经营（销售散装食品）；主营产品为虾夷扇贝。

7.财务会计机构及其工作情况

设有财务部门，各部门分岗位有序开展工作。部门现有财务工作人员 35 人，其中：高级职称 3 人，中级职称 18 人，初级职称 12 人。

8.相关内部控制及其执行情况

设有内部控制相关制度，且按内控制度开展日常工作活动。

9.重大会计政策选用及变动情况

2020 年度未有重大会计政策变动。

10.以往接受审计的情况

以往每年都会接受专业机构进行的审计。

11.重大经营活动和其他重要情况

主要从事扇贝养殖与销售。

12.被审计企业的领导人员基本情况

法人：刘万全，董事长：夏圣文，总经理：罗宁，财务负责人：冯秋媛。

1.1.2 了解被审计单位背景

点击“了解被审计单位背景”，了解被审计单位基本情况，将以答题卡中的信息归类到对应的调查内容中。如表 2-1-1 所示。

<p>任务描述：了解被审计单位基本情况，将以下信息归类到对应的调查内容中。</p> <p>养殖产量逐渐增加，海水养殖成为主要方式。 法人：刘万全，董事长：夏圣文，总经理：罗宁，财务负责人：冯秋媛。 贝类产品不断增加，养殖收入逐步提高。 品种结构基本形成，主产品产量持续增长。 贝类产品在提供优质蛋白、优化居民饮食结构、提高渔民收入、增加外汇收入与改善近海环境等方面发挥了重要作用。 贝类养殖也面临着空间制约、水域污染、资源衰退、食品安全等诸多问题。</p> <p>扇贝养殖与销售。 由大连市人民政府国有资产监督管理委员会100%控股。 设股东大会、董事会、监事会，监事会下设总经理及各管理部门，董事会下设审计委员会。 设有总经办、采购部、销售部、行政部、财务部、仓储部门、养殖部门、质量控制部门。 是大连市人民政府国有资产监督管理委员会下属企业。 2020年战略目标是提高扇贝产量、提升生产能力，经济收入持续增长。 设有财务部门，部门分岗位有序开展工作。 部门现有财务工作人员35人，其中：高级职称3人，中级职称18人，初级职称12人。 设有内部控制相关制度，且按内控制度开展日常工作活动。 2020年度未有重大会计政策变动。 各项监管措施加大力度科学调整养殖产业结构，修复养殖海域生态环境。 以往每年都会接受专业机构进行审计。 经营范围包括许可项目：水产苗种生产，水产养殖，渔业捕捞，食品经营，食品互联网销售（销售预包装食品），食品经营（销售散装食品）；主营产品为虾夷扇贝。</p>	
了解被审计单位基本环境	
调查方面	具体内容（单选或多选）
被审计企业所处的经济环境	<input type="checkbox"/>
被审计企业的行业状况、法律环境、监管环境以及其他外部因素	<input type="checkbox"/>
被审计企业的所有权结构、治理结构和组织结构	<input type="checkbox"/>
被审计企业财政、财务隶属关系或者国有资产监督管理关系	<input type="checkbox"/>
被审计企业的战略目标和相关经营	<input type="checkbox"/>
被审计企业的经营范围、业务性质、经营情况和关联方交易	<input type="checkbox"/>
财务会计机构及其工作情况	<input type="checkbox"/>
相关内部控制及其执行情况	<input checked="" type="checkbox"/>
重大会计政策选用及变动情况	<input type="checkbox"/>
以往接受审计的情况	<input type="checkbox"/>
重大经营活动和其他重要情况	<input type="checkbox"/>
被审计企业的领导人员基本情况	<input type="checkbox"/>

表 2-1-1 了解被审计单位基本情况题目

答案如表 2-1-2 所示：

调查方面	具体内容
被审计企业所处的经济环境	养殖产量逐渐增加，海水养殖成为主要方式。贝类产值不断增加，养殖收入逐步提高。品种结构基本形成，主产品产量持续增长。
被审计企业的行业状况、法律环境、监管环境以及其他外部因素	贝类产业在供应优质蛋白、优化居民饮食结构、提高渔民收入、增加外汇收入与改善近海环境等方面发挥了重要作用。贝类养殖也面临着空间制约、水域污染、资源衰退、食品安全等诸多问题。各项监管措施加大力度科学调整养殖产业结构，修复养殖海域生态环境。
被审计企业的所有权结构、治理结构和组织结构	由大连市人民政府国有资产监督管理委员会100%控股。设股东大会、董事会、监事会，监事会下设总经理及各管理部门，董事会下设审计委员会。设有总经办、采购部、销售部、行政部、财务部、仓储部门、养殖部门、质量控制部门。
被审计企业财政、财务隶属关系或者国有资产监督管理关系	是大连市人民政府国有资产监督管理委员会下属企业。
被审计企业的战略目标和相关经营	2020年战略目标是提高扇贝产量、提升生产能力，经济收入持续增长。
被审计企业的经营范围、业务性质、经营情况和关联方交易	经营范围包括许可项目：水产苗种生产，水产养殖，渔业捕捞，食品经营，食品互联网销售（销售预包装食品），食品经营（销售散装食品）；主营产品为虾夷扇贝。
财务会计机构及其工作情况	设有财务部门，部门分岗位有序开展工作的。部门现有财务工作人员35人，其中：高级职称3人，中级职称18人，初级职称12人。
相关内部控制及其执行情况	设有内部控制相关制度，且按内控制度开展日常工作活动。
重大会计政策选用及变动情况	2020年度未有重大会计政策变动。
以往接受审计的情况	以往每年度都会接受专业机构进行审计。
重大经营活动和其他重要情况	扇贝养殖与销售。
被审计企业的领导人员基本情况	法人：刘万全，董事长：夏圣文，总经理：罗宁，财务负责人：冯秋媛。

表 2-1-2 了解被审计单位基本情况题目答案

解析：该表格中的问题在上一段视频中的对话中均有详细描述，同学们只需要认真观看即可提取答案。

点击右侧的步骤分析回答相关选择题

问题1
问题2

经济环境（单选题 1.00分）

1. 对被审计企业所处的经济环境的描述，以下正确的是（ ）

- A. 扇贝养殖行业养殖产量逐渐减少
- B. 海水养殖成为主要方式
- C. 贝类产值不断减少
- D. 主产品产量持续下降

正确答案：B

解析：由对话可知，经济环境：扇贝养殖行业养殖产量逐渐增加，海水养殖成为主要方

式；贝类产值不断增加，养殖收入逐步提高；品种结构基本形成，主产品产量持续增长。

步骤分析

问题1 问题2

行业分析 (多选题 1.00分)

2. 贝类养殖行业面临的挑战包括 ()

- A. 空间制约
- B. 水域污染
- C. 资源衰退
- D. 食品安全

正确答案：ABCD

解析：对话中提到，贝类养殖行业面临的挑战包括空间制约、水域污染、资源衰退和食品安全等问题。故选择 A,B, C, D 。

步骤二：确定审计工作重难点

1.2.1 编制审计方案

点击“确定审计工作重难点”左上方的“编制审计方案”，观看动画，如图 2-1-2 所示，了解本次审计工作的重难点，进一步认识审计相关注意事项及流程。

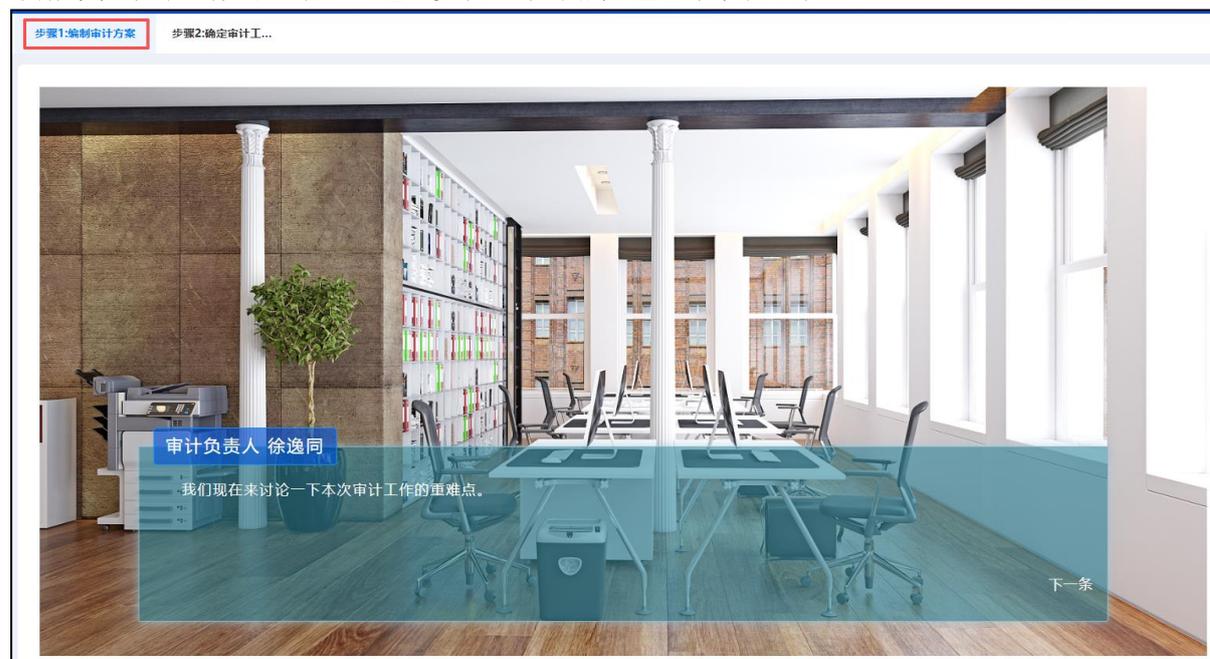


图 2-1-2 编制审计方案视频对话场景

1.2.2 确定审计工作重难点

点击“确定审计工作重难点”左上方的“”，结合上一步骤“编制审计方案”中的动画对话内容，进行答题，如图 2-1-3 至图 2-1-4 所示。

1. 下列关于本次审计工作的重点，说法正确的有（ ）（多选题 1分） ?

- A. 审查扇贝减产因素是否真实存在及扇贝是否减产
- B. 审查账面存货的真实性
- C. 对存货进行盘点
- D. 还原企业扇贝的捕捞区域、船只捕捞记录

图 2-1-3 确定审计工作重难点第 1 题

答案：ABCD

解析：通过前面的对话，可知本次审计工作的重点包括：（1）审查扇贝减产因素是否真实存在及扇贝是否减产；（2）审查账面存货的真实性；（3）对存货进行盘点；（4）还原企业扇贝的捕捞区域、船只捕捞记录。

2. 下列关于审计难点的说法，正确的有（ ）（多选题 1分） ?

- A. 向农户个人购买扇贝种苗时无发票，对追溯存货采购存在困难
- B. 存货资产存在重大不确定性，由于存货资产对所处环境敏感性较高，环境异常变动对其生长影响较大，影响资产存活率
- C. 存货资产盘点困难，由于存货资产生长在海里，区域广泛且数量庞大，无法靠人工进行准确盘点
- D. 海洋水产养殖业，虾夷扇贝库存难以肉眼观测、采捕作业海域难以事后追溯

图 2-1-4 确定审计工作重难点第 2 题

答案：ABCD

解析：通过前面的对话可知本次审计工作的难点包括：（1）向农户个人购买扇贝种苗时无发票，对追溯存货采购存在困难；（2）存货资产存在重大不确定性，由于存货资产对所处环境敏感性较高，环境异常变动对其生长影响较大，影响资产存活率；（3）存货资产盘点困难，由于存货资产生长在海里，区域广泛且数量庞大，无法靠人工进行准确盘点；（4）海洋水产养殖业，虾夷扇贝库存难以肉眼观测、采捕作业海域难以事后追溯。

步骤三：确定审计假设

1.3.1 审计组沟通

单击步骤导航栏中的“确定审计假设”左上角的审计组沟通，观看动画，了解本次审计工作的相关假设，如图 2-1-5 所示。并根据相应对话，回答相应场景问题，如图 2-1-6 至 2-1-8 所示。

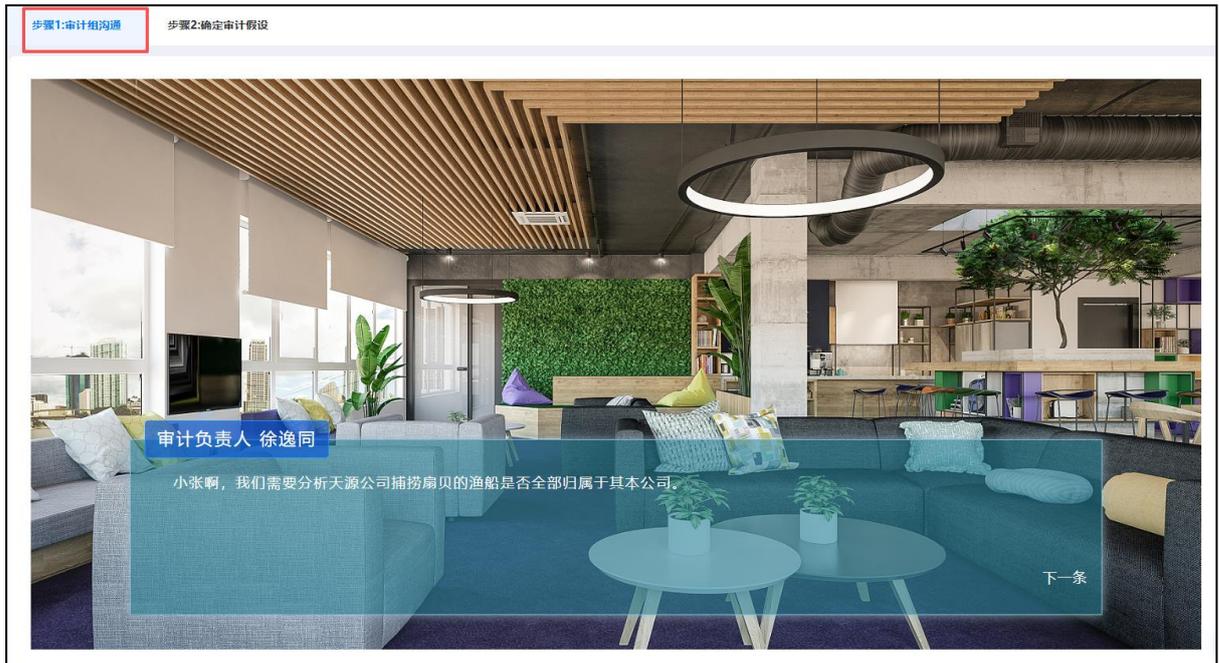


图 2-1-5 审计假设对话场景

1.3.2 确定审计假设

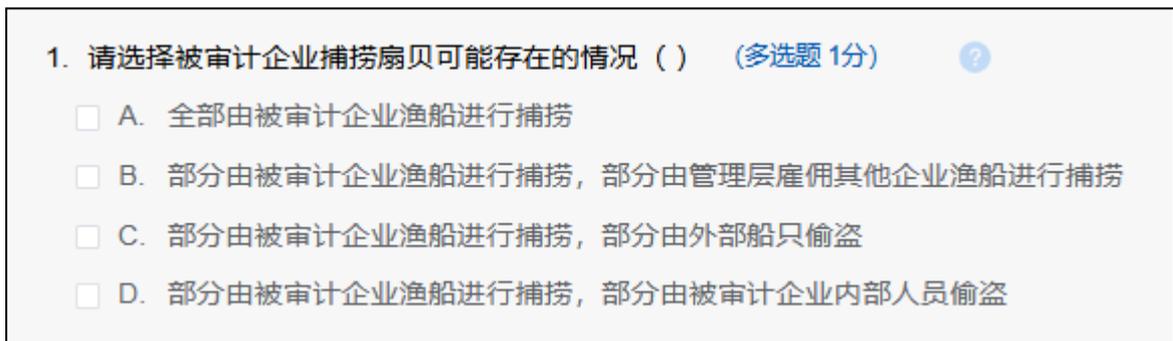


图 2-1-6 审计假设第 1 题

答案：ABCD

解析：通过本步骤对话场景可知，天源贝类养殖股份有限公司捕捞扇贝的渔船供应充足，不存在渔船供应紧张的情况，如果扇贝没有减产，可能存在捕捞渔船不全部归属于天源公司的情况，因此四个答案选项均正确。

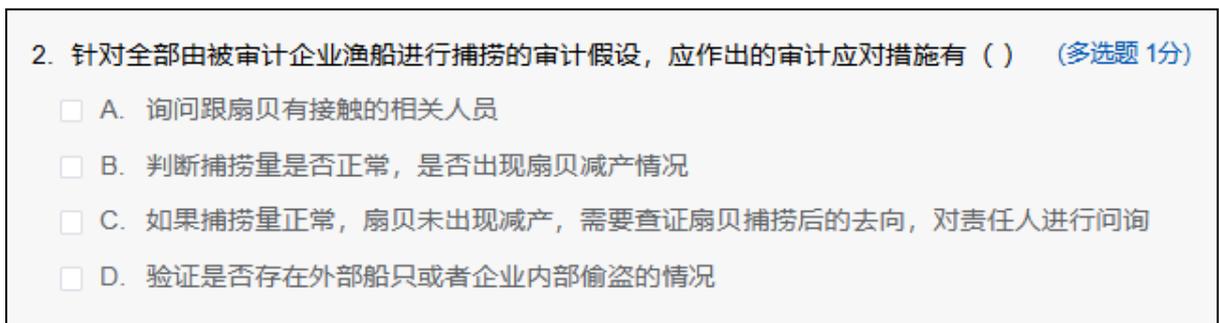


图 2-1-7 审计假设第 2 题

答案：ABCD

解析：通过本步骤对话场景可知，针对全部由被审计企业渔船进行捕捞的审计假设，应作出的审计应对措施有：（1）询问跟扇贝有接触的相关人员；（2）判断捕捞量是否正常，是否出现扇贝减产情况；（3）如果捕捞量正常，扇贝未出现减产，需要查证扇贝捕捞后的去向，对责任人进行问询；（4）验证是否存在外部船只或者企业内部偷盗的情况。

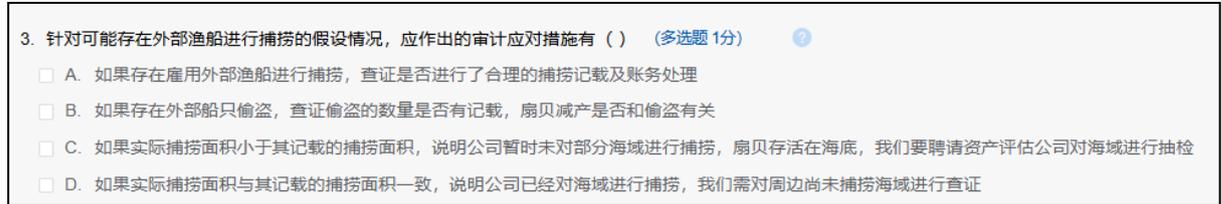


图 2-1-8 审计假设第 3 题

答案：ABCD

解析：通过本步骤对话场景可知，针对可能存在外部渔船进行捕捞的审计假设，应作出的审计应对措施有：（1）如果存在雇用外部渔船进行捕捞，查证是否进行了合理的捕捞记载及账务处理；（2）如果存在外部船只偷盗，查证偷盗的数量是否有记载，扇贝减产是否和偷盗有关；（3）如果实际捕捞面积小于其记载的捕捞面积，说明公司暂时未对部分海域进行捕捞，扇贝存活在海底，我们要聘请资产评估公司对海域进行抽检；（4）如果实际捕捞面积与其记载的捕捞面积一致，说明公司已经对海域进行捕捞，我们需对周边尚未捕捞海域进行查证。

步骤四：函证获取数据

1.4.1 完善函证内容

单击步骤导航栏中的“函证获取数据”下的“完善函证内容”，下载答题卡填写答案，如图 2-1-9 所示。



答案：养殖海域、捕捞海域等相关资料。

解析：审计天源贝类养殖股份有限公司需要协助提供养殖海域和捕捞海域资料，主要是为了核实其生物资产（如扇贝等海产品）的真实性、数量和价值。审计机构必须通过独立验证海域权属、实际使用情况以及捕捞记录，来确认存货是否存在、计量是否准确，并排除虚构交易或虚增资产的可能性。这些外部资料（如政府颁发的海域使用权证、海洋监管部门备案的养殖区域图、第三方捕捞日志等）能够提供客观证据，弥补企业自行提供数据的可信度不足问题，是发现存货异常的关键审计证据。

国家海洋局大连海洋环境监测中心站：											
我局准备对大连天源贝类养殖股份有限公司开展审计工作，需要单位协助提供渤海湾海域2016-2020年的海水温度数据统计，包括_____。											
专此函达，请答复。											
											

答案：经度、维度、深度和温度等相关资料。

解析：审计天源贝类需要协助单位提供养殖海域的经度、纬度、深度和温度等相关资料，主要是为了验证其贝类养殖业务的真实性和财务数据的准确性。这些关键环境数据能够客观反映养殖海域的实际状况，帮助审计人员核实贝类生长的合理性和产量数据的可信度。通过比对历史环境数据与养殖记录，可以排查是否存在虚构养殖区域或虚报产量的风险。同时，深度和温度等参数直接影响贝类生长周期和存活率，是评估存货价值的重要依据。获取这些第三方数据能够弥补企业自行提供信息的局限性，为存货审计提供独立、可靠的证据支持，有效防范类似獐子岛“扇贝跑路”的财务舞弊风险。

北斗卫星导航系统有限公司											
我局准备对大连天源贝类养殖股份有限公司开展审计工作，需要单位协助提供大连天源贝类养殖股份有限公司2020年度所有渔船的卫星定位数据以及大连天源贝类养殖股份有限公司养殖海域内所有船只的卫星定位数据，卫星定位数据包括_____。											
专此函达，请答复。											
											

答案：经度、纬度、捕捞船只和其他船只等相关资料。

解析：审计天源贝类需要捕捞海域的经度、纬度、捕捞船只和其他船只等相关资料，主要是为了验证其捕捞作业的真实性和贝类存货的准确性。这些数据能够客观记录捕捞活动的具体位置和时间，通过比对 GPS 定位数据与申报的捕捞区域，可以核实是否存在虚报捕捞范围或产量的情况。捕捞船只的识别信息和其他船只的监控记录，则有助于确认捕捞作业的合规性，排查是否存在借用外部船只虚构捕捞量的风险。这些第三方提供的航行日志、船舶 AIS 轨迹等资料，能够为存货审计提供独立证据，有效识别潜在的财务舞弊行为，确保财务报表中贝类存货数据的真实可靠。

点击右侧步骤分析，完成数据收集的相关选择题

步骤分析	
问题1	
收集数据 (多选题 1.00分)	
1. 下列审计数据中，需要发函获取的有 ()。	
<input type="checkbox"/>	A. 海域归属权
<input type="checkbox"/>	B. 海洋数据
<input type="checkbox"/>	C. 卫星定位数据
<input type="checkbox"/>	D. 财务数据

答案：ABC

解析：在审计中，是否需要发函取决于数据的可验证性和来源：海域归属权（A）必须发函，因其需政府机构确认；海洋数据（B）和卫星走位数据（C）若来自外部机构也需发函。财务数据（D）若能从账套直接验证（如原始凭证齐全）可不发函。

1.4.2 查看函证结果

单击步骤导航栏中的“函证获取数据”下的“查看函证结果”，查看回函内容。

<p>任务描述：</p> <p>查看回函结果。</p>
<p>回函1</p> <div style="border: 1px solid #ccc; padding: 10px; margin: 10px 0;"> <p>大连市审计局：</p> <p>大连天源贝类养殖股份有限公司渤海湾海域使用权61.63万亩，主要用于水产养殖。</p> <p>大连市人民政府国有资产监督管理委员会 (盖章)</p> <p>2021年_3_月_11_日</p> </div>
<p>回函2</p> <div style="border: 1px solid #ccc; padding: 10px; margin: 10px 0;"> <p>大连市审计局：</p> <p>我局网站“国家海洋科学数据中心”记载了历年各海域的海水温度数据，请自行前往下载。</p> <p>国家海洋局大连海洋环境监测中心站 (盖章)</p> </div>

步骤五：下达审计通知书

点击左侧导航栏中的“下达审计通知书”，查看向被审计单位下达的审计通知书。



任务二：风险评估

单击步骤导航栏中的“案例资料-报表数据”，查看利润表、资产负债表和趋势分析

The screenshot displays a financial statement analysis interface. The main table is titled "利润表" (Profit Statement) and shows data for the period from 2019/12/31 to 2020/8/31. The table includes columns for the period and rows for various financial items. The data is as follows:

序号	项目	2019/12/31	2020/1/31	2020/2/28	2020/3/31	2020/4/30	2020/5/31	2020/6/30	2020/7/31	2020/8/31
10	自用固定资产 (损失以“-”号填列)									
17	资产减值损失 (损失以“-”号填列)									
18	资产处置收益 (损失以“-”号填列)									
19	二、营业利润 (亏损以“-”号填列)	7,373,500.56	7,195,070.41	7,961,572.17	54,735,621.50	52,527,910.70	6,143,349.10	6,750,835.72	6,626,457.81	6,537,912.90
20	加：营业外收入									
21	减：营业外支出									
22	三、利润总额 (亏损总额以“-”号填列)	7,373,500.56	7,195,070.41	7,961,572.17	54,735,621.50	52,527,910.70	6,143,349.10	6,750,835.72	6,626,457.81	6,537,912.90
23	减：所得税	921,687.57	899,383.80	995,196.52	6,841,952.69	6,565,988.84	767,918.64	843,854.47	828,307.23	817,239.12
24	四、净利润 (净亏损以“-”号填列)	6,451,812.99	6,295,686.61	6,966,375.65	47,893,668.81	45,961,921.86	5,375,430.46	5,906,981.26	5,798,150.59	5,720,673.79
25	(一) 持续经营净利润 (净亏损以“-”号填列)									
26	(二) 终止经营净利润 (净亏损以“-”号填列)									
27	五、其他综合收益的税后净额									
28	(一) 不能重分类进损益的其他综合收益									
29	1. 重新计量设定受益计划变动额									
30	2. 权益法下不能转损益的其他综合收益									
31	3. 其他权益工具投资公允价值变动									
32	4. 企业自身信用风险公允价值变动									
33									
34	(二) 将重分类进损益的其他综合收益									
35	1. 权益法下可转损益的其他综合收益									

步骤一：扇贝生长因素研究

单击步骤导航栏中的“扇贝生长因素研究”下的“扇贝生长因素研究”，点击右上角【数据中心】，选择“天源贝类-研究报告”，进入扇贝研究报告分析页面。请对影响虾夷扇贝成长、存活率等因素进行分析，并完成相应选择题。

任务描述:

点击右上角【数据中心】，选择“天源贝类-研究报告”，进入扇贝研究报告分析页面。请对影响虾夷扇贝成长、存活率等因素进行分析。

扇贝生长因素分析

1. 下列关于扇贝生长的因素，正确的有（ ） (多选题 1分)

- A. 海水温度
- B. 溶解氧
- C. 养殖密度
- D. 苗种质量

答案：ABCD。

解析：查阅扇贝研究报告可知，海水温度、溶解氧、养殖密度和苗种质量均对扇贝生长有影响。

步骤二：财务指标分析

2.2.1 业务说明

单击步骤导航栏中的“财务指标分析”左上角的“业务说明”，根据任务描述查看相关数据。

任务描述:

点击右上角【数据中心】，选择“天源贝类-财报数据”，进入上市公司财务指标爬取页面，爬取上市公司2016-2021年主要指标、资产负债表、利润表、现金流量表等财务数据。



2.2.2 采集财务指标数据

单击步骤导航栏中的“财务指标分析”左上角的“采集财务指标数据”，完善代码运行。



根据前面步骤中得出的海水温度分析结果，再依次点击“平均温度计算——技术需求转化——需求实现——执行并显示结果”，输入相关指令进行数据进一步处理。

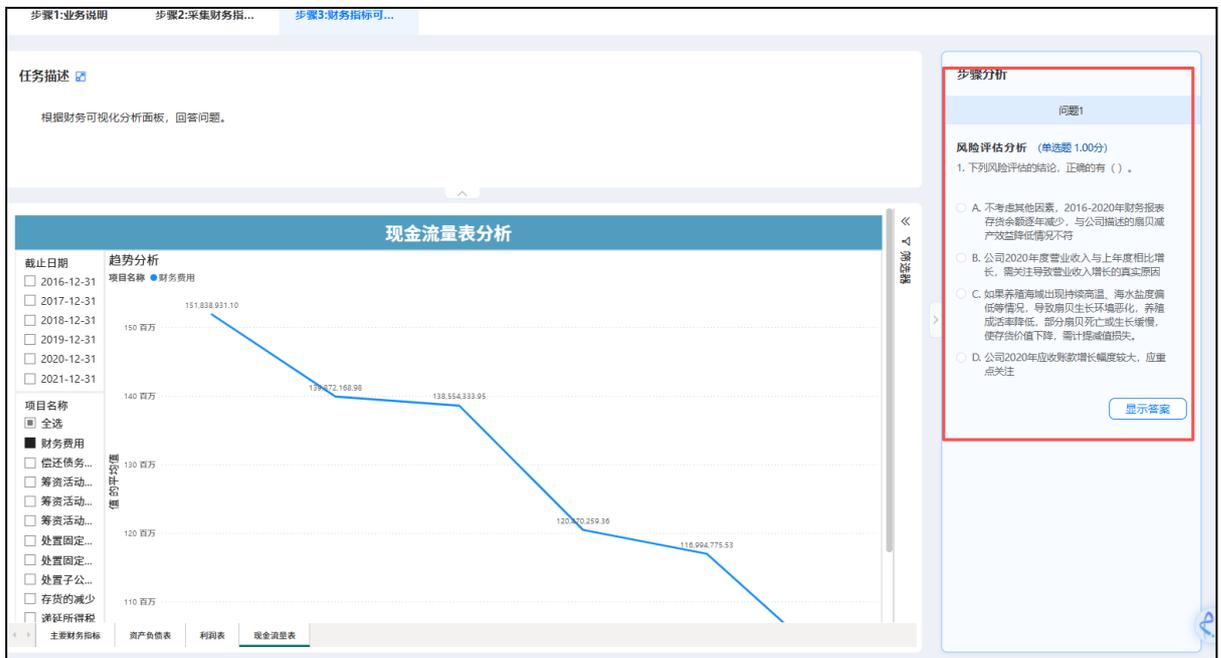
(1) 主要指标

(2) 2021-12-31

(3) <http://43.136.80.101:9000/fstxskzhxxkj/DataCenter/3065/5/index.html>

2.2.3 财务指标可视化分析

单击步骤导航栏中的“财务指标分析”左上角的“财务指标可视化分析”，根据财务可视化分析面板，回答问题。



正确答案：C

解析：A、扇贝减产-其他因素情况不变的情况下-存货余额减少-效益降低，与公司描述的情况相符，错误。 B、公司 2020 年度营业收入与上年度相比减少，需关注导致营业收入减少的真实原因，错误。 C、如果海水环境问题导致的扇贝质量下降和价值降低，需计提减脂减

值准备，正确。D、公司应收账款减少，选项内容相反，错误。

步骤三：海水温度数据分析

2.3.1 海水温度数据采集

单击步骤导航栏中的“海水温度数据分析”左上角的“海水温度数据采集”，查看并保存相关数据信息。

The screenshot displays the 'Data Collection' page on the National Marine Science Data Center (NMDSC) website. The page is titled '数据采集' (Data Collection) and includes a breadcrumb trail: '分析预报数据 > 实况分析数据 > 实况分析数据-温度'. The main content area is divided into two columns: '基本信息' (Basic Information) and '数据块摘要' (Data Block Summary). The '基本信息' section provides details such as the data name 'Real-time analysis data-temperature', time period '2016-2021', sharing level '完全公开' (Fully Open), and update frequency '每日' (Daily). The '数据块摘要' section describes the data as '西北太平洋实况分析产品的海面温度为99°E-150°E, 10°S-52°N, 水平分辨率0.125°, 垂向为标量层, 产品要素包括: 三维温度、盐度、密度、声速和地转流。' (Northwest Pacific real-time analysis product surface temperature data for 99°E-150°E, 10°S-52°N, with a horizontal resolution of 0.125°, vertical as scalar layer, and product elements including 3D temperature, salinity, density, sound speed, and geostrophic flow). Below the information sections are two '查询条件' (Query Conditions) sections, each with a search input field and '查询' (Query) and '清除' (Clear) buttons.

2.3.2 数据筛选

单击步骤导航栏中的“海水温度数据分析”左上角的“数据筛选”，根据任务描述补全程序编码。

任务描述

121.125	38.875	-20
121.125	38.75	-20

海水温度筛选 (7.00分)

程序编码补全

```
1 import csv
2 import os
3
4
5 # 处理原始数据,筛选指定坐标点
6 def write_csv():
7     lon = ['120.875,38.75,-20', '(1)请补全代码', '120.875,39.0,-20', '121.125,39.0,-20', '121.0,38.75,-20',
8           '121.125,38.875,-20', '120.875,38.875,-20', '(2)请补全代码', '121.0,38.875,-20']
9     csv_path = r'[UserFolderPath]/海水温度(2016_2020).csv'
10    csv_path2 = r'[UserFolderPath]/海水温度-筛选后.csv'
11    data_list = []
12    if os.path.exists(csv_path):
13        with open('(3)请补全代码', 'r') as fp:
14            data = fp.read().split('\n')
15            for da in data:
16                if any(lo in da for lo in lon):
17                    data_list.append(da)
18        with open(csv_path2, 'w', newline='') as fp:
```

步骤分析

问题1

数据筛选 (填空题 1.00分)

1. 筛选后的样本数据一共 (1) 条, 涵盖 (2) 天的温度数据, 其中有 (3) 条无效数据。无效数据来源于坐标 (4) 、 (5) 。 (按经度、维度、深度的顺序填写, 用“,”分开)

(1) :

(2) :

(3) :

(4) :

(5) :

(1) 121.125,38.75,-20

(2) 121.0,39.0,-20

(3) csv_path

根据材料回答右侧步骤分析中的计算填空题

步骤分析

问题1

数据筛选 (填空题 1.00分)

1. 筛选后的样本数据一共 (1) 条, 涵盖 (2) 天的温度数据, 其中有 (3) 条无效数据。无效数据来源于坐标 (4) 、 (5) 。 (按经度、维度、深度的顺序填写, 用“,”分开)

(1) :

(2) :

(3) :

(4) :

(5) :

正确答案:

- (1) 8523
- (2) 947
- (3) 2039
- (4) 121.125,39,-20
- (5) 121.125,38.875,-20

2.3.3 计算海水平均温度

单击步骤导航栏中的“海水温度数据分析”左上角的“计算海水平均温度”，根据任务描述补全程序编码。

The screenshot shows a learning management system interface. At the top, there are four steps: '步骤1:海水温度数据...', '步骤2:数据筛选', '步骤3:计算海水平均...', and '步骤4:海水平均温...'. The '步骤3' step is active. Below the steps, there is a '任务描述' (Task Description) section with the text '完善代码, 计算海水平均温度。'. The main area is titled '计算海水平均温度 (7.00分)'. It contains a '程序编码补全' (Code Completion) section with a code editor. The code is as follows:

```
1 import pandas as pd # 导入pandas数据分析模块
2
3
4 # 对筛选后的数据进行处理,计算平均温度
5 def pd_clean():
6     csv_path = r'(UserFolderPath)/海水温度(平均温度).csv'
7     csv_path2 = r'(1)请补全代码
8     da_df = pd.read_csv((2)请补全代码, encoding='gbk') # 读取待处理数据文件
9     #print(da_df)
10    data = da_df.groupby(['日期']).agg({'温度': '(3)请补全代码'}) # 根据数据列名进行分组求平均值
11    data['深度'] = -20 # 添加新列并进行赋值
12    data.rename(columns={'温度': '平均温度'}, inplace=True) # 修改列名为指定名称
13    data.insert(0, '深度', data.pop('深度')) # 根据索引插入数据
14    data.to_csv(csv_path, encoding='utf-8-sig') # 导出处理后的数据文件
15    #print(da_df)
16
17
18 pd_clean()
19
```

On the right side, there is a '步骤分析' (Step Analysis) section. It contains '问题1' (Question 1) titled '海水温度分析 (多选题 1.00分)'. The question is: '1. 下列关于渤海湾海水温度数据的分析结论, 正确的有 ()。'. The options are:

- A. 2020年各月海水温度数据与往年大体一致, 未发现异常
- B. 未发现2020年有异常偏低或者异常偏高的海水温度
- C. 2020年8月存在异常偏高的海水温度
- D. 2020年1月存在异常偏低的海水温度

根据材料回答右侧步骤分析中的选择题

This is a detailed view of the '步骤分析' (Step Analysis) section. It shows '问题1' (Question 1) titled '海水温度分析 (多选题 1.00分)'. The question is: '1. 下列关于渤海湾海水温度数据的分析结论, 正确的有 ()。'. The options are:

- A. 2020年各月海水温度数据与往年大体一致, 未发现异常
- B. 未发现2020年有异常偏低或者异常偏高的海水温度
- C. 2020年8月存在异常偏高的海水温度
- D. 2020年1月存在异常偏低的海水温度

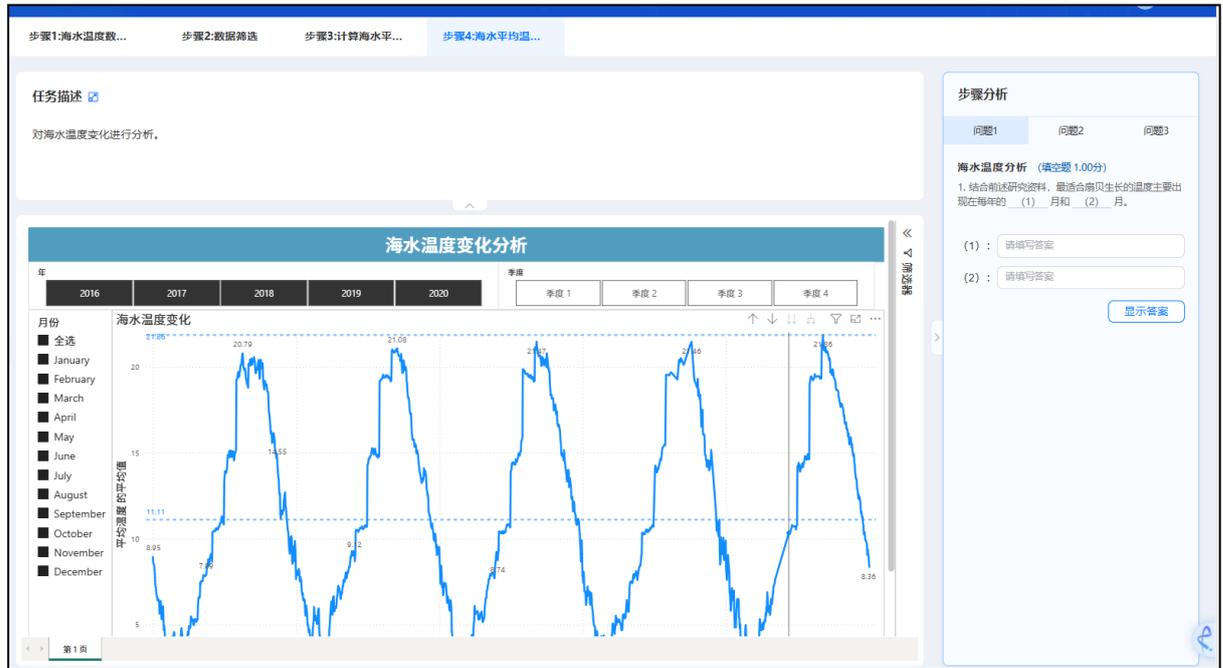
答案: AB。

解析: 根据海域平均温度趋势图。可以看到每年的季节性波动都是正常一致的, 并没有

偏高或者偏低。

2.3.3 海水平均温度分析

单击步骤导航栏中的“海水温度数据分析”左上角的“海水平均温度分析”，根据图表完成右侧相关的题目。



问题 1

步骤分析

问题1 问题2 问题3

海水温度分析 (填空题 1.00分)

1. 结合前述研究资料, 最适合扇贝生长的温度主要出现在每年的 (1) 月和 (2) 月。

(1): 请填写答案

(2): 请填写答案

显示答案

正确答案: (1) 7(2) 11

解析: 参阅数据中心-研究报告-虾夷扇贝对温度的-饥饿胁迫的应答机制可知 5~23℃ 为虾夷扇贝生长和发育的适宜温度范围, 15℃ 左右为最适生长温度, 此时, 虾夷扇贝的摄食率达到峰值。7、11月海水温度为 15℃。

问题 2:

步骤分析

问题1	问题2	问题3

海水温度分析 (单选题 1.00分)

2. 观察可视化面板, 结合上文材料信息, 关于海水温度的变化, 以下说法, 错误的是 ()

- A. 2020年6月到7月和7月到8月交界处, 温度会大幅变化, 可能导致扇贝会消耗更多的能量, 发生扇贝大规模死亡。
- B. 2020年平均温度是近5年最低的, 低温会影响扇贝摄食率
- C. 2020年海水温度整体呈现先降低后升高再降低的变化趋势
- D. 2020年达到了近五年来海水温度最高值, 可能受全球变暖影响

正确答案: B

解析:

由图可知, 2019年平均温度 9.4℃, 是近五年最低的平均温度, B 错误

问题 3:

步骤分析

问题1	问题2	问题3

海水温度分析 (单选题 1.00分)

3. 对比往年, 2020年海水温度是否存在异常变化 ()

- A. 存在, 海水温度整体偏低, 温度超出扇贝生长的适宜温度。
- B. 存在, 海水温度整体偏高, 温度超出扇贝生长的适宜温度。
- C. 不存在, 海水温度与往年相比变化不大, 依然在适合扇贝生长的温度范围。
- D. 存在, 海水温度在下半年相比于往年温度整体偏低, 不在适合扇贝生长的温度范围。

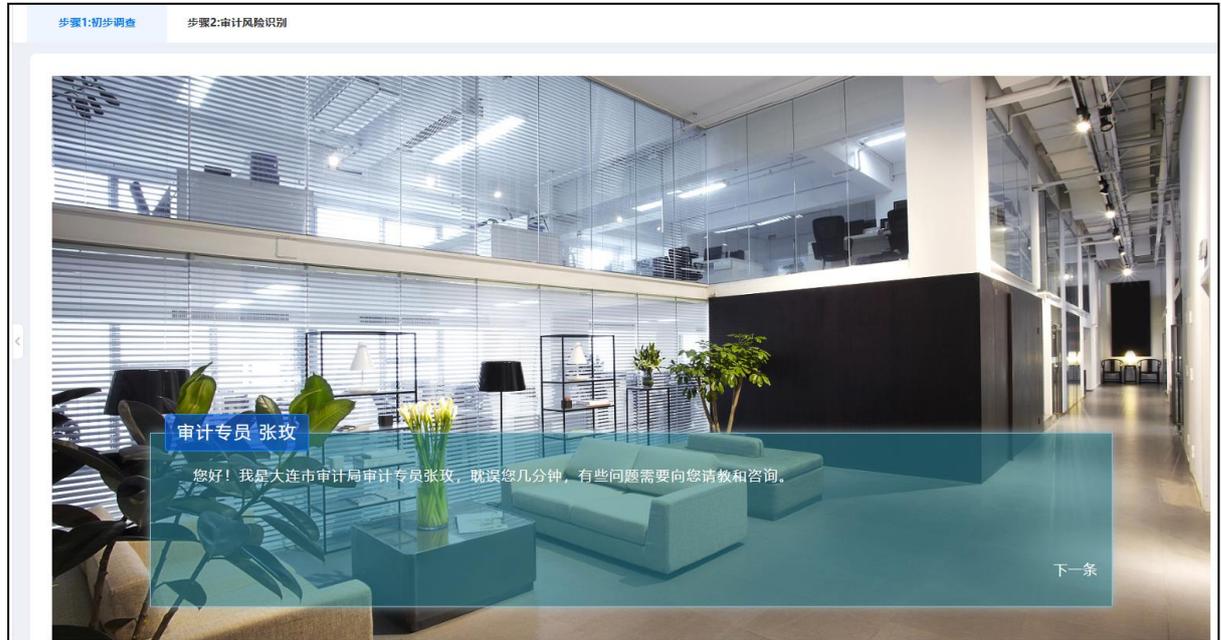
正确答案: C

解析: C 正确, 由图可知, 每年年的季节性波动都是正常一致的, 并没有偏高或者偏低, 且气温都在 5~23℃ 之间, 在适合生长的范围。ABC 的描述中均存在不适合的扇贝温度范围描述, 因此错误。

步骤四: 风险评估分析

2.4.1 初步调查

单击步骤导航栏中的“风险评估分析”左上角的“初步调查”，根据动画获取信息。



2.4.2 审计风险识别

单击步骤导航栏中的“风险评估分析”左上角的“审计风险识别”，根据相关资料识别审计风险。完成选择题。

审计风险分析

1. 根据情景对话进行分析判断，下列说法中错误的是（ ）（单选题 1分）

- A. 2020年公司的养殖海域遭受了几次冷水团袭击，海水温度异常偏冷，但大数据分析未发现在2020年存在异常海水温度的情况
- B. 公司2020年度的存货量比上一年度有所减少，审计人员可以导出收发存明细表核实具体数据
- C. 据了解，2020年度基本上对贝龄成熟的海域都进行了捕捞，有些海域未捕捞是因为冷水团不利于船只出海
- D. 可以通过比对相同捕捞海域同行业的相关数据再核实冷水团袭击是否真实存在

正确答案：C

试题解析：C，错误。根据对话可知，2020年度基本上对贝龄成熟的海域都进行了捕捞，有些海域未捕捞是因为扇贝处于未长大的状态。其他选项正确。

2. 下列风险评估的结论，正确的有（ ）（多选题 1分）

- A. 应重点关注公司记载的2020年度捕捞海域是否真实
- B. 未发现2020年度存在导致扇贝减产的因素，应重点关注公司2020年度亩产量下降是否真实
- C. 2020年度海水温度异常偏冷，存在扇贝减产的因素
- D. 公司可能存在管理层舞弊现象，需重点关注

正确答案：ABD

任务三：内部控制测试

步骤一：检查分析

3.1.1 采购流程分析

根据指引点击到下一任务模块：“任务三：内部控制测试”——再点击步骤一的第一步“采购流程分析”。如下图所示！

任务描述：
根据下文提供的采购申请单和验收单相关资料，对采购流程进行分析，然后回答问题。

任务名称：
任务三：内部控制测试

任务实例：
步骤一：检查分析

1 采购流程分析

2 采购流程分析

3 销售流程分析

步骤二：内部控制

步骤三：内部控制分析

名称	规格	单位	数量	单价	金额	供应商
虾夷扇贝苗	2<H<3 (H-扇贝壳高)	枚	1746089700	0.06	104765382	大连叁乔海产有限公司
申请人：张玉	采购部经理：蔡伟	财务经理：冯秋媛	总经理：罗宁			
日期：2017年12月17日	日期：2017年12月17日	日期：2017年12月18日	日期：2017年12月18日			

查看资料框，查看若干份资料，回答后面的问题。

任务描述：
根据下文提供的采购申请单和验收单相关资料，对采购流程进行分析，然后回答问题。

总金额	大写：人民币柒仟贰佰肆拾捌万叁仟伍佰捌拾捌元整	小写：	72483588.00
采购单位验收意见	验收意见：(✓)合格；()不合格 验收人：王娜 验收时间：2018年12月4日		

采购验收单

采购公司：大连叁乔海产有限公司 (盖章) 采购日期：2019年12月13日 验收单号：1206256125352

货物名称	规格型号	数量及单位	单价	金额
虾夷扇贝苗	2≤H<3 (H-扇贝壳高)	2017261700 枚	0.06	121035702.00

问题回答：

1. 下列关于采购与付款内控测试，错误的有（ ）。（单选题 1分）
- A. 格式采购合同（或采购订单）需合同订立人填写《采购合同》或“采购申请单”，经部门经理确认，公司总经理审核，财务负责人审核即可。
 - B. 采购验收单、采购订单需与采购记录表互相核对检查，确保资料的完整性和准确性
 - C. 采购验收单为财务部门办理付款、入库等手续提供凭证，确保资金支付与验收结果一致
 - D. 观察采购验收单发现大连叁乔海产有限公司是虾夷扇贝苗的唯一供应商

正确答案:D

试题解析:D、观察发现大连市丰缪扇贝养殖专业合作社和大连叁乔海产有限公司都是虾夷扇贝苗的供应商，D 错误

3.1.2 养殖流程分析

按照下图指引，点击相关按钮，回答右侧步骤分析的问题。

任务描述

扇贝采用底播的方式，先把扇贝幼苗投入到海里进行养殖，等扇贝长大成熟后进行捕捞。查看被审计单位养殖过程数据，回答问题。

查看任务描述

1. 采购流程分析

2. 养殖流程分析

3. 销售流程分析

步骤二：走访调查

步骤三：内部控制分析

正在加载数据...

步骤分析

问题1

问题2

养殖情况（单选题 1.00分）

1. 饲料投资率为（ ）

- A. 间隔1个月
- B. 间隔2个月
- C. 间隔15天
- D. 间隔1.5个月

显示答案

任务描述

扇贝采用底播养殖的方式，先要把扇贝幼苗投入到底底进行养殖，等扇贝长大成熟后进行捕捞。查看被审计单位养殖过程数据，回答问题。

饲料投放情况

投放面积 (亩) (按年和月份)

年	华奥岛1区	华奥岛2区	华奥岛3区	华奥岛4区	总计
2020	1338237	1330438	1314018	1347268	53299
季度 1	446132	284269	295083	297438	13229
January	153363	133654	142364		4313
February	139617	150615		149547	4397
March	151152		152719	147891	4517
季度 2	303329	459275	283554	302180	13483
April		153215	143651	153619	4504
May	152069	155791	139903		4477
June	151260	150269		148561	4500
季度 3	297134	291119	446061	295607	13299
July	147519		153189	142691	4433
August		142603	140269	152916	4357
September	149615	148516	152603		4507
季度 4	291642	295775	289320	452043	13287
October	137891	147615		150615	4361
November	153751		136951	152915	4436
总计	1338237	1330438	1314018	1347268	53299

日期	投放地点	投放面积 (亩)	投放组	监督人员	备注
2020/1/15	华奥岛2区	133654	投放二组 (组长:李斌 组员:张明、王东东、罗管有、肖文艺、李建国、张科)	孙月涛	无异常
2020/3/5	华奥岛3区	152719	投放二组 (组长:李斌 组员:张明、王东东、罗管有、肖文艺、李建国、张科)	孙月涛	无异常
2020/4/25	华奥岛4区	153619	投放二组 (组长:李斌 组员:张明、王东东、罗管有、肖文艺、李建国、张科)	孙月涛	无异常
2020/6/15	华奥岛1区	151260	投放二组 (组长:李斌 组员:张明、王东东、罗管有、肖文艺、李建国、张科)	孙月涛	无异常
2020/8/5	华奥岛2区	142603	投放二组 (组长:李斌 组员:张明、王东东、罗管有、肖文艺、李建国、张科)	孙月涛	无异常
总计		5329961			

饲料投放区域

查看完相关资料，回答右侧问题：

步骤分析

问题1
问题2

养殖情况 (单选题 1.00分)

1. 饲料投放频率为 ()

A. 间隔1个月
 B. 间隔2个月
 C. 间隔15天
 D. 间隔1.5个月

关闭答案

步骤分析

问题1
问题2

养殖过程分析 (多选题 1.00分)

2. 经过分析，可以得到如下结论 ()

A. 海域安全巡逻基本保证每月4-5次，扇贝生长关键月份外来船只较多，但都有效管控。
 B. 扇贝捕捞主要集中在3、4月份，尤其是3月居多。
 C. 扇贝播种主要集中在12月份，这个月的温度是最适合扇贝生长的温度
 D. 目前四个养殖区的饲料投放面积相差不多，各月份饲料投放量相差不多

关闭答案

问题 1：养殖情况(单选题 1.00 分)

1.饲料投放频率为()

- A. 间隔 1 个月
- B.间隔 2 个月
- C. 间隔 15 天
- D.间隔 1.5 个月

正确答案:B

解析:由可视化视图可知, 饲料大致每隔两个月投放一次

问题 2: 养殖过程分析(多选题 1.00 分)

2.经过分析, 可以得到如下结论()

- A.海域安全巡逻基本保证每月 4-5 次, 扇贝生长关键月份外来船只较多, 但都有效管控。
- B.扇贝捕捞主要集中在 3、4 月份, 尤其是 3 月居多。
- C.扇贝播种主要集中在 12 月份, 这个月的温度是最适合扇贝生长的温度
- D.目前四个养殖区的饲料投放面积相差不大, 各月份饲料投放量相差不大

正确答案:ABD

解析:

C 错误, 最适合扇贝生长的温度主要集中在 6、7、11 月份

3.1.3 销售流程分析

根据下图操作, 然后回答右侧步骤分析的问题

任务描述 ②查看任务描述

查看被审计单位销售过程数据, 回答问题。

日期	客户名称	商品名称	数量	单位	单价	金额	发货人	出库单号
2020/9/22	除尔滨家华海洋食品有限公司	南澳虾夷扇贝	450000	kg	18	8100000	万元站	20200322143
2020/9/19	南京露露达贸易有限公司	南澳虾夷扇贝	400000	kg	18	7200000	万元站	20200319264
2020/9/25	大连鼎龙海产品有限公司	南澳虾夷扇贝	400000	kg	18	7200000	万元站	20200325621
2020/11/22	除尔滨家华海洋食品有限公司	南澳虾夷扇贝	390000	kg	18	7020000	万元站	20201122623
2020/4/13	南京露露达贸易有限公司	南澳虾夷扇贝	395000	kg	18	7110000	万元站	20200413029
2020/11/20	南京露露达贸易有限公司	南澳虾夷扇贝	395000	kg	18	7110000	万元站	20201120649
2020/9/16	长春露木食品有限公司	南澳虾夷扇贝	390000	kg	18	7020000	万元站	20200916316
2020/4/9	沈阳华丰海味有限公司	南澳虾夷扇贝	390000	kg	18	7020000	万元站	20200409019
2020/4/10	长春露木食品有限公司	南澳虾夷扇贝	390000	kg	18	7020000	万元站	20200410013
2020/10/16	长春露木食品有限公司	南澳虾夷扇贝	390000	kg	18	7020000	万元站	20201016158
2020/10/23	除尔滨家华海洋食品有限公司	南澳虾夷扇贝	390000	kg	18	7020000	万元站	20201023019
2020/11/27	晋源石福源有限公司	南澳虾夷扇贝	389500	kg	18	7011000	万元站	20201127640
2020/10/20	南京露露达贸易有限公司	南澳虾夷扇贝	385000	kg	18	6930000	万元站	20201020394
2020/9/22	除尔滨家华海洋食品有限公司	南澳虾夷扇贝	380000	kg	18	6840000	万元站	20200922012
2020/4/9	沈阳华丰海味有限公司	南澳虾夷扇贝	380000	kg	18	6840000	万元站	20200409012
2020/4/17	除尔滨家华海洋食品有限公司	南澳虾夷扇贝	380000	kg	18	6840000	万元站	20200417038
2020/11/8	大连瑞升海洋食品有限公司	南澳虾夷扇贝	380000	kg	18	6840000	万元站	20201108587
2020/11/28	晋源石福源有限公司	南澳虾夷扇贝	369800	kg	18	6656400	万元站	20201128640
2020/10/28	晋源石福源有限公司	南澳虾夷扇贝	360000	kg	18	6480000	万元站	20201028012
2020/10/9	大连瑞升海洋食品有限公司	南澳虾夷扇贝	360000	kg	18	6480000	万元站	20201009678
2020/11/5	大连露达食品有限公司	南澳虾夷扇贝	361100	kg	18	6499800	万元站	20201105619
2020/9/6	大连露达食品有限公司	南澳虾夷扇贝	360000	kg	18	6480000	万元站	20200906004
2020/4/25	大连鼎龙海产品有限公司	南澳虾夷扇贝	350000	kg	18	6400000	万元站	20200425028
2020/9/19	南京露露达贸易有限公司	南澳虾夷扇贝	350000	kg	18	6300000	万元站	20200919005
2020/4/10	长春露木食品有限公司	南澳虾夷扇贝	350000	kg	18	6300000	万元站	20200410016
2020/10/22	除尔滨家华海洋食品有限公司	南澳虾夷扇贝	350000	kg	18	6300000	万元站	20201022068
2020/11/25	大连鼎龙海产品有限公司	南澳虾夷扇贝	350000	kg	18	6300000	万元站	20201125694
2020/11/25	大连鼎龙海产品有限公司	南澳虾夷扇贝	350000	kg	18	6300000	万元站	20201125636
2020/11/12	大连露达食品有限公司	南澳虾夷扇贝	340000	kg	18	6120000	万元站	20201102327
总计			38931691		3130	695923740		3292712642300

① 查看各表格

问题回答:

步骤分析

问题1 问题2 问题3

销售与出库情况 (判断题 3.00分)

1. 对比出库和销售情况, 可以发现基本保持一致。

正确 错误

关闭答案

正确答案: 正确

解析:
波动情况基本一致

问题1 问题2 问题3

销售情况分析 (单选题 1.00分)

2. 维持被审计单位全年销售收入的产品主要是()

A. 单冻半壳虾夷扇贝
 B. 底播虾夷扇贝
 C. 浮筏虾夷扇贝
 D. 底播虾夷扇贝和浮筏虾夷扇贝

关闭答案

正确答案: A

解析:
单冻半壳虾夷扇贝销售额最高

步骤分析

问题1 问题2 问题3

结算方式分析 (单选题 1.00分)

3. 被审计单位采用现款现货结算方式, 这可能导致哪些审计风险()

A. 现金流量的真实性难以确认
 B. 存货的计价不准确
 C. 客户信用风险显著增加
 D. 财务报表的透明度提高

关闭答案

正确答案: A

解析:
现款现货结算方式要求立即收取现金, 这可能会导致现金流入记录的不准确。如果财务人员未能及时或准确记录现金交易, 可能会造成财务报表中现金流量的失真, 这种情况下, 审计师需要仔细核对现金收入与相关单据, 以确保现金流的真实性。

查看并回答问题 1、2、3。为方便查看, 具体题目与答案解析如下:

问题 1: 销售与出库情况(判断题 3.00 分)

1.对比出库和销售情况, 可以发现基本保持一致。

- 正确
- 错误

正确答案:正确

解析:

波动情况基本一致

问题 2: 销售情况分析(单选题 1.00 分)

2.维持被审计单位全年销售收入的产品主要是()

- A.单冻半壳虾夷扇贝
- B.底播虾夷扇贝
- C.浮筏虾夷扇贝
- D.底播虾夷扇贝和浮筏虾夷扇贝

正确答案:A

解析:

单冻半壳虾夷扇贝销售额最高

问题 3: 结算方式分析(单选题 1.00 分)

3.被审计单位采用现款现货结算方式, 这可能导致哪些审计风险()

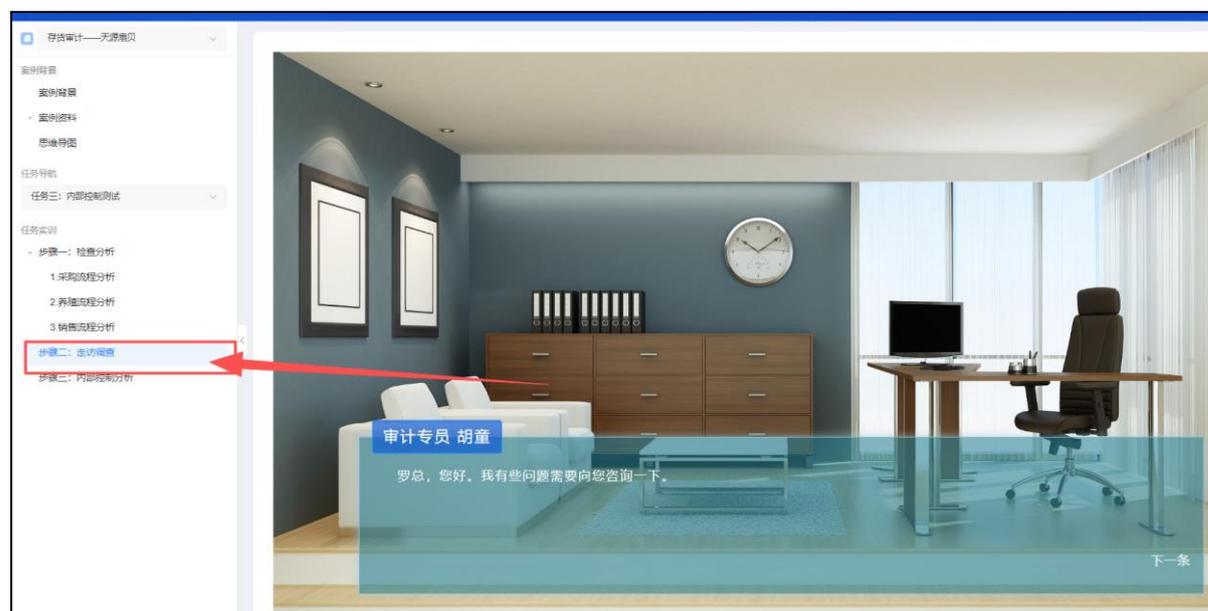
- A.现金流量的真实性难以确认
- B.存货的计价不准确
- C. 客户信用风险显著增加
- D.财务报表的透明度提高

正确答案:A

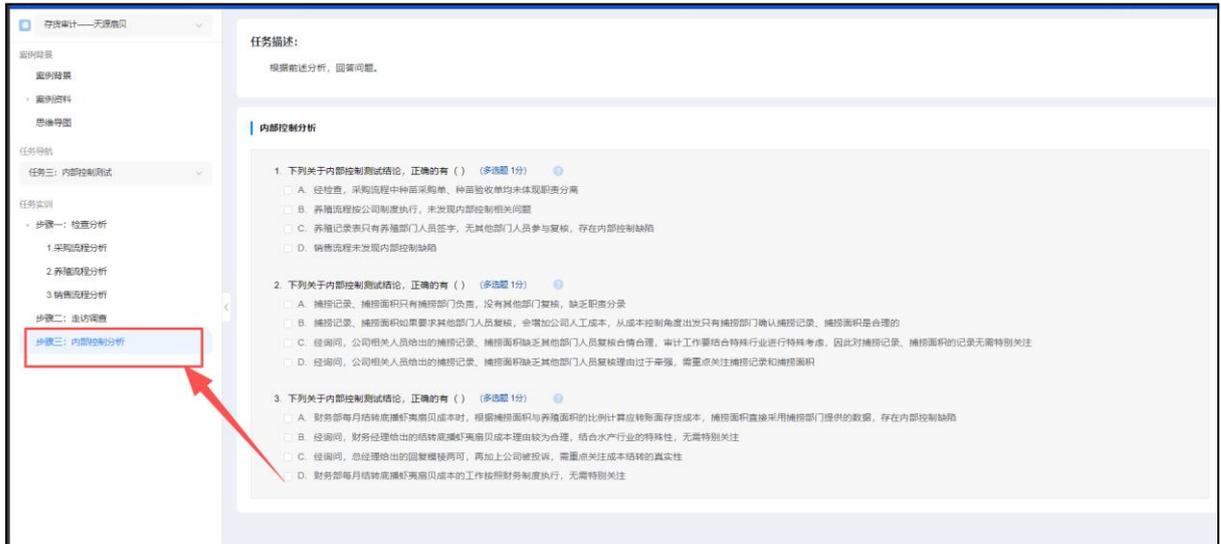
解析:现款现货结算方式要求立即收取现金, 这可能会导致现金流入记录的不准确。如果财务人员未能及时或准确记录现金交易, 可能会造成财务报表中现金流量的失真, 这种情况下, 审计师需要仔细核对现金收入与相关单据, 以确保现金流的真实性。

步骤二: 走访检查

根据图片指引查看“步骤二: 走访检查”。查看相关审计访谈对话, 据此回答步骤三的问题。



步骤三: 内部控制分析



查看任务描述并回答问题。为方便查看，具体题目与答案解析如下：

1. 下列关于内部控制测试结论，正确的有（）（多选题 1分）

- A. 经检查，采购流程中种苗采购单、种苗验收单均未体现职责分离
- B. 养殖流程按公司制度执行，未发现内部控制相关问题
- C. 养殖记录表只有养殖部门人员签字，无其他部门人员参与复核，存在内部控制缺陷
- D. 销售流程未发现内部控制缺陷

2. 下列关于内部控制测试结论，正确的有（）（多选题 1分）

- A. 捕捞记录、捕捞面积只有捕捞部门负责，没有其他部门复核，缺乏职责分离
- B. 捕捞记录、捕捞面积如果要求其他部门人员复核，会增加公司人工成本，从成本控制角度出发只有捕捞部门确认捕捞记录、捕捞面积是合理的
- C. 经询问，公司相关人员给出的捕捞记录、捕捞面积缺乏其他部门人员复核合情合理，审计工作要结合特殊行业进行特殊考虑，因此对捕捞记录、捕捞面积的记录无需特别关注
- D. 经询问，公司相关人员给出的捕捞记录、捕捞面积缺乏其他部门人员复核理由过于牵强，需重点关注捕捞记录和捕捞面积

3. 下列关于内部控制测试结论，正确的有（）（多选题 1分）

- A. 财务部每月结转底播虾夷扇贝成本时，根据捕捞面积与养殖面积的比例计算应转账面存货成本，捕捞面积直接采用捕捞部门提供的数据，存在内部控制缺陷
- B. 经询问，财务经理给出的结转底播虾夷扇贝成本理由较为合理，结合水产行业的特殊性，无需特别关注
- C. 经询问，总经理给出的回复模棱两可，再加上公司被投诉，需重点关注成本结转的真实性
- D. 财务部每月结转底播虾夷扇贝成本的工作按照财务制度执行，无需特别关注

1、下列关于内部控制测试结论，正确的有() (多选题 1分)

- A、经检查，采购流程中种苗采购单、种苗验收单均未体现职责分离。
- B.养殖流程按公司制度执行，未发现内部控制相关问题。
- C、养殖记录表只有养殖部门人员签字，无其他部门人员参与复核，存在内部控制缺陷。
- D、销售流程未发现内部控制缺陷。

正确答案:CD

试题解析:通过实地调查得知被审计单位采购和销售流程没有问题，能体现职责分离:但是养殖流程中缺乏职责分离存在内部控制缺陷。

2.下列关于内部控制测试结论，正确的有() (多选题 1分)

- A.捕捞记录、捕捞面积只有捕捞部门负责，没有其他部门复核，缺乏职责分离。
- B.捕捞记录、捕捞面积如果要求其他部门人员复核，会增加公司人工成本，从成本控制角度出发只有捕捞部门确认捕捞记录、捕捞面积是合理的。
- C.经询问，公司相关人员给出的捕捞记录、捕捞面积缺乏其他部门人员复核合情合理，审计工作要结合特殊行业进行特殊考虑，因此对捕捞记录、捕捞面积的记录无需特别关注。

D.经询问，公司相关人员给出的捕捞记录、捕捞面积缺乏其他部门人员复核理由过于牵强，需重点关注捕捞记录和捕捞面积。

正确答案:AD

试题解析:选项 A 和 D 是正确的。选项 A 指出“捕捞记录、捕捞面积只有捕捞部门负责,没有其他部门复核,缺乏职责分离这属于典型的内部控制缺陷。职责分离是内部控制的基本原则,若仅由单一部门完成记录和确认,可能导致数据错误或舞弊风险增加。选项 D 进一步强调“公司相关人员给出的缺乏复核的理由过于勉强,需重点关注捕捞记录和捕捞面积”,这表明审计人员应保持职业怀疑态度,对不合规的内部控制流程采取进一步审计程序。选项 B 认为”如果要求其他部门复核会增加人工成本,因此仅由捕捞部门确认是合理的”,这一观点错误,因为成本控制不能以牺牲内部控制有效性为代价否则可能导致更大的财务或合规风险。选项 C 主张”无需特别关注捕捞记录和捕捞面积”,这与审计的基本要求相违背,审计人员应对高风险领域保持关注,而非轻易接受管理层解释。因此,AD 是符合内部控制测试要求的正确结论。

3.下列关于内部控制测试结论，正确的有()(多选题 1 分)

A、财务部每月结转底播虾夷扇贝成本时，根据捕捞面积与养殖面积的比例计算应转账面存货成本，捕捞面积直接采用捕捞部门提供的数据，存在内部控制缺陷

B、经询问，财务经理给出的结转底播虾夷扇贝成本理由较为合理，结合水产行业的特殊性，无需特别关注。

C、经询问，总经理给出的回复模棱两可，再加上公司被投诉，需重点关注成本结转的真实性

D、财务部每月结转底播虾夷扇贝成本的工作按照财务制度执行，无需特别关注

正确答案:AC

试题解析:在评估内部控制测试结论时，选项 A 和 C 是正确的。选项 A 指出了数据来源单一且缺乏独立验证的问题。良好的内部控制应确保关键数据的准确性，仅依赖单一部门提供的数据可能导致成本结转不准确，甚至存在舞弊风险。选项 C 提这表明审计人员应保持职业怀疑态度，尤其在管理层解释不清晰或存在外部投诉的情况下，必须对成本结转的合理性进行深入核查。选项 B 认为“财务经理的解释合理，结合行业特殊性无需特别关注”，这一结论过于轻率，行业特殊性不能成为忽视内部控制缺陷的理由，审计人员仍需确保数据的可靠性和流程的合规性，有可能成本计提过程中存在舞弊行为。选项 D 称“财务部按制度执行，无需特别关注”，但制度本身若存在缺陷(如缺乏复核机制)，仍可能导致重大错报风险，因此不能简单认定其合规性。综上，AC 是符合审计逻辑的正确结论，强调了内部控制缺陷的识别和风险应对的必要性。

任务四：实质性测试

前导

根据下图指引，点击到任务导航的任务四模块，点击案例资料，查看相关资料。



步骤一：对采购与付款实施实质性测试

根据下图的操作指引，点进进入步骤一“对采购与付款实施实质性测试”，查看任务描述与两个表格，回答右侧步骤分析问题。



问题回答：

查看任务描述并回答问题。为方便查看，具体题目与答案解析如下：

步骤分析		步骤分析	
问题1	问题2	问题1	问题2
采购分析 (单选题 1.00分) 1. 扇贝企业在采购中供应商较为单一，这种集中采购模式可能引发哪些潜在风险？请分析并选择最合适的选项（） <input type="radio"/> A. 采购流程效率降低，审批环节增多 <input checked="" type="radio"/> B. 供应商议价能力增强，企业采购成本上升 <input type="radio"/> C. 仓储管理难度加大，库存积压风险增加 <input type="radio"/> D. 采购人员操作规范性不足，舞弊风险上升 关闭答案		分析性复核 (多选题 1.00分) 2. 下列关于采购与付款的实质性测试结论，正确的有（） <input checked="" type="checkbox"/> A. 种苗采购记录表、种苗付款明细表均有一致的记录，未发现异常 <input type="checkbox"/> B. 种苗采购记录表、种苗付款明细表存在不一致的记录，需重点关注 <input type="checkbox"/> C. 种苗采购的供应商未出现在供应商清单中，需重点关注 <input checked="" type="checkbox"/> D. 种苗采购的月份与其播种的月份一致，未发现异常 关闭答案	

采购分析(单选题 1.00 分)

1.扇贝企业在采购中供应商较为单一，这种集中采购模式可能引发哪些潜在风险?请分析并选择最合适的选项()

- A.采购流程效率降低，审批环节增多
- B.供应商议价能力增强，企业采购成本上双保
- C.仓储管理难度加大，库存积压风险增加
- D.采购人员操作规范性不足，舞弊风险上升

正确答案:B

解析:

选项 B(正确):供应商单一时，企业依赖少数甚至单一来源，供应商在谈判中占据主动地位，可能通过抬高价格、降低服务质量等方式增加企业采购成本(如芯片行业若仅依赖单一芯片供应商，可能面临高价垄断)。这是集中采购模式下供应商单一的最直接、典型风险。选项 A(错误):单一供应商通常简化了供应商筛选和比价流程，反而可能提高采购效率(如减少多家比选的耗时)，而非降低效率，选项 C(错误):仓储管理难度和库存积压主要与采购计划、需求预测相关，与供应商数量无直接必然联系(单一供应商也可能精准供货，多家供应商也可能导致库存混乱)。选项 D(错误):舞弊风险更多源于内部控制缺陷(如采购审批不透明、缺乏监督)，而非供应商单一本身(即使供应商更多，若内控失效仍可能舞弊)。总结:供应商单一的核心风险是议价权失衡，导致企业成本上升或供应稳定性受到威胁(如供应商断供)，本题最直接体现为成本风险(选项 B)

分析性复核(多选题 1.00 分)

2.下列关于采购与付款的实质性测试结论，正确的有

- A.种苗采购记录表、种苗付款明细表均有一致的记录，未发现异常

- B.种苗采购记录表、种苗付款明细表存在不致的记录，需重点关注
- C.种苗采购的供应商未出现在供应商清单中，需重点关注
- D.种苗采购的月份与其播种的月份一致，未发现异常

关闭答案

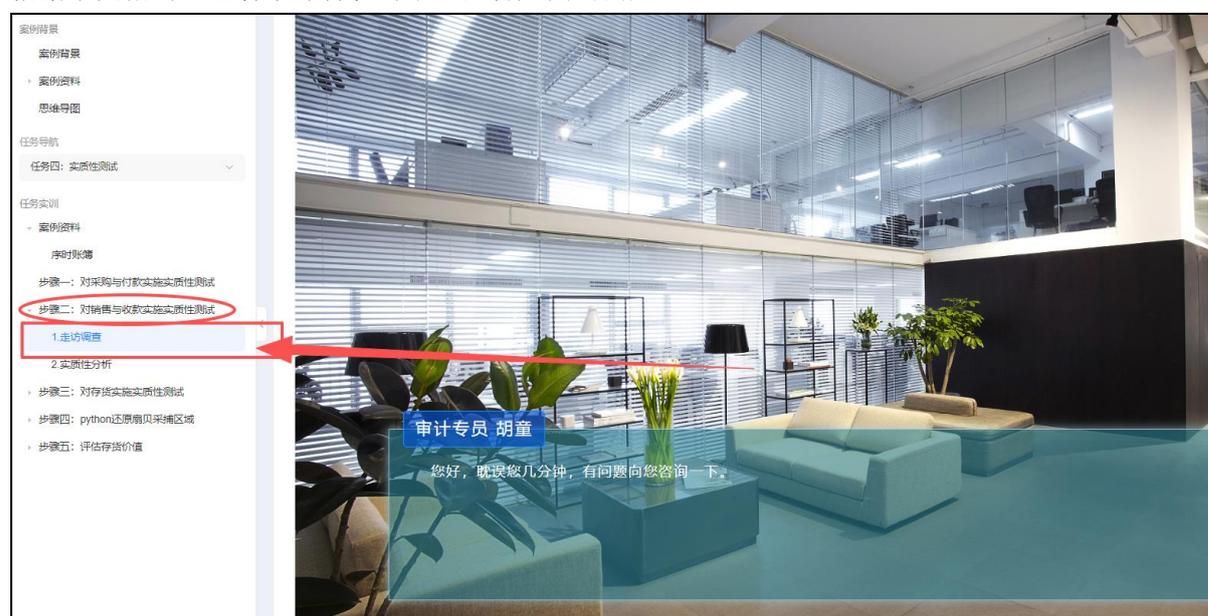
正确答案:AD

解析:以采购单号为唯一值，关联种苗的采购记录表和付款记录表，发现两者有一致记录，未发现异常，A 正确，B 错误。采购供应商在供应商清单中，B 错误。任务三:内控测试中养殖过程分析的养殖记录表中底播月份在 12 月，与采购记录表的月份一致，未发现异常，D 正确。

步骤二：对销售与收款实施实质性测试

1. 走访调查、

根据下图指引，查看审计访谈对话，回答后面的问题。



2. 实质性分析

根据下图指引，点击机内步骤二的第二部分“实质性分析”，查看问题描述，与两个表格。回答右侧步骤分析的问题。

任务描述

查看被审计单位销售与收款情况，回答问题，完成分析。

②查看问题描述

主要客户明细表

序号	客户名称	开始合作日期	法人姓名	关键联系人职位及姓名	备注
1	大连益达食品有限公司	2017年3月	张柯	总经理	重点客户
2	大连神州海洋食品有限公司	2017年5月	李娜	总经理	重点客户
3	大连宇丰食品有限公司	2017年8月	程文	财务总监	重点客户
4	长春霖大食品有限公司	2017年10月	何伟	张琳 (财务总监)	重点客户
5	南京益达食品有限公司	2018年2月	李慧娟	李光强 (总经理)	重点客户
6	烟台益达食品有限公司	2018年5月	张雷	张琳 (财务总监)	重点客户
7	大连益达食品销售有限公司	2018年10月	刘国恩	李光强 (总经理)	重点客户
8	烟台益达食品销售有限公司	2018年11月	李娜	张琳 (财务总监)	重点客户
9	大连益达食品销售有限公司	2018年12月	李娜	张琳 (财务总监)	重点客户
10	烟台益达食品销售有限公司	2019年3月	李娜	张琳 (财务总监)	重点客户
11	长沙益达食品销售有限公司	2019年9月	张琳	张琳 (财务总监)	重点客户
12	烟台益达食品销售有限公司	2019年12月	李娜	张琳 (财务总监)	重点客户
13	烟台益达食品销售有限公司	2020年1月	李娜	张琳 (财务总监)	重点客户
14	西安益达食品销售有限公司	2020年2月	王江涛	张琳 (财务总监)	重点客户
15	西安益达食品销售有限公司	2020年4月	李娜	张琳 (财务总监)	重点客户
16	烟台益达食品销售有限公司	2020年10月	李娜	张琳 (财务总监)	重点客户
17	大连益达食品销售有限公司	2020年11月	王城	张琳 (财务总监)	重点客户

客户销售对比

2019: 27088800
2020: 12434360
2018: 20971960
2017: 20067600

开始合作日期

2017: 42.05%
2018: 42.24%
2019: 8.20%
2020: 8.51%

实质性测试 (多选题 1.00分)

1. 下列关于销售与收款的实质性测试结论，正确的有 ()

A. 经审计，公司在3、4、10、11月的2019和2020两期变动比率最大，说明扇贝的减产对旺季的影响更显著

B. 经审计，公司只有在3、4、10、11月才会销售鲜活底播虾夷扇贝、鲜活浮筏虾夷扇贝，结合公司被举报的背景，需重点关注是否存在异常

C. 经审计，公司捕捞月份为3、4、10、11月，且只有这4个月销售鲜活底播虾夷扇贝、鲜活浮筏虾夷扇贝，其它月份只销售单冻半壳虾夷扇贝，符合虾夷扇贝的行业环境

D. 经询问，销售经理给出的公司在3、4、10、11月销售额比其他月份多的理由符合虾夷扇贝的行业环境

③查看两个表格

问题回答：

步骤分析

问题1 问题2

实质性测试 (多选题 1.00分)

1. 下列关于销售与收款的实质性测试结论，正确的有 ()

A. 经审计，公司在3、4、10、11月的2019和2020两期变动比率最大，说明扇贝的减产对旺季的影响更显著

B. 经审计，公司只有在3、4、10、11月才会销售鲜活底播虾夷扇贝、鲜活浮筏虾夷扇贝，结合公司被举报的背景，需重点关注是否存在异常

C. 经审计，公司捕捞月份为3、4、10、11月，且只有这4个月销售鲜活底播虾夷扇贝、鲜活浮筏虾夷扇贝，其它月份只销售单冻半壳虾夷扇贝，符合虾夷扇贝的行业环境

D. 经询问，销售经理给出的公司在3、4、10、11月销售额比其他月份多的理由符合虾夷扇贝的行业环境

[关闭答案](#)

步骤分析

问题1 问题2

实质性分析 (多选题 1.00分)

2. 经分析，可以得到如下结论 ()

A. 相比于上年同期，被审计单位各月份销售收入均有降低，且降低幅度在6%-25%之间

B. 从每年扇贝分销商新增数量来看，2019、2020年变化不大

C. 目前被审计单位主要收入来源于老客户，新客户的贡献较低，整体不足全年收入的10%

D. 新增客户与被审计单位存在关联关系，可能存在收入舞弊风险

[关闭答案](#)

查看任务描述与资料，回答问题。为方便查看，具体题目与答案解析如下：

问题 1 实质性测试(多选题 1.00 分)

1. 下列关于销售与收款的实质性测试结论，正确的有

A. 经审计，公司在 3、4、10、11 月的 2019A. 和 2020 两期变动比率最大，说明扇贝的减产对旺季的影响更显著

B.经审计，公司只有在 3、4、10、11 月才会销售鲜活底播虾夷扇贝、鲜活浮筏虾夷扇贝，结合公司被举报的背景，需重点关注是否存在异常

经审计，公司捕捞月份为 3、4、10、11 月，且只有这 4 个月销售鲜活底播虾夷扇贝、鲜活浮筏虾夷扇贝，其他月份只销售单冻半壳虾夷扇贝，符合虾夷扇贝的行业环境

D.经询问，销售经理给出的公司在 3、4、10、11 月销售额比其它月份多的理由符合虾夷扇贝的行业环境

正确答案:ACD

解析:

B，并不是 3、4、10、11 月才销售，而是这四个月为销售的旺季，错误。其他正确

问题 2 实质性分析 (多选题 1.00 分)

2.经分析，可以得到如下结论()

A.相比于上年同期，被审计单位各月份销售收入均有降低，且降低幅度在 6%-25%之间

B.从每年扇贝分销商新增数量来看，2019.2020 年变化不大

C.目前被审计单位主要收入来源于老客户，新客户的贡献较低，整体不足全年收入的 10%

D.新增客户与被审计单位存在关联关系，可能存在收入舞弊风险

正确答案:ABC

A、从销售变动情况可知，被审计单位的各月份销售收入都有所降低，且变动比率都在 6%-35%，正确。B、开始合作时间在 2020 年、2019 年的分销商较少，可见 2019、2020 年新增的分销商数量较少，正确。C、由客户销售额对比图可知，销售额贡献较多的分销商主要是 2018、2017 年合作的客户，正确。D、并没有明确的证据表明存在关联关系，D 错误

步骤三：对存货实施实质性测试

1. 走访调查

根据下图指引，点击步骤三的第一部分，查看相关对话。

The screenshot displays a software interface for an audit task. On the left, a sidebar menu lists various tasks under '任务实训' (Task Practice). The selected task is '步骤三：对存货实施实质性测试' (Step 3: Substantive testing of inventory), with its first sub-task '1. 走访调查' (1. Field investigation) highlighted. A red arrow points from this sub-task to a 3D rendered office environment. In the office scene, a blue speech bubble contains the text '审计专员 胡童 您好，又要打扰您了。' (Audit specialist Hu Tong: Hello, I'm disturbing you again). A '下一条' (Next) button is located in the bottom right corner of the office scene.

2. 实质性分析

根据下图指引，点击下一个步骤实质性分析，并且查看任务描述与两个表格，回答右侧步骤分析的问题。

The screenshot shows an audit software interface with the following components:

- Task Description:** 查看存货管理情况，然后回答问题。
- Table 1: 主营业务成本变动表**

月份	2019年	2020年	变动额	变动比率
1	17,865,609.68	13,996,215.27	-3,869,394.41	-21.66%
2	16,156,886.73	14,026,493.26	-2,130,393.47	-13.19%
3	71,891,582.83	58,647,887.01	-13,243,705.82	-18.41%
4	70,594,873.28	55,643,558.80	-14,951,314.38	-21.18%
5	15,365,204.12	12,456,498.06	-2,908,706.06	-18.93%
6	15,594,616.89	12,605,486.60	-2,989,130.29	-19.17%
7	16,065,694.31	12,464,478.18	-3,601,216.13	-22.42%
8	15,581,696.31	12,456,478.18	-3,125,218.13	-19.99%
9	15,055,696.31	13,303,674.82	-1,752,021.49	-11.64%
10	73,846,897.29	60,986,012.80	-12,860,884.49	-17.29%
11	75,694,781.09	58,687,932.02	-16,806,849.07	-22.20%
12	16,713,705.16	14,176,130.29	-2,537,574.87	-15.18%
- Table 2: 主营业务成本明细表**

月份	商品名称	单位	单价	销售数量	成本结转金额
2020年1月	单冻半壳虾夷扇贝	kg	13.16	1376000	13,996,215.27
2020年2月	单冻半壳虾夷扇贝	kg	9.81	1431000	14,026,493.26
2020年3月	单冻半壳虾夷扇贝	kg	10.02	1279900	12,816,244.29
2020年4月	单冻半壳虾夷扇贝	kg	6.63	2656400	22,541,124.61
2020年5月	冷冻扇贝	kg	6.07	2568500	20,890,510.37
2020年6月	单冻半壳虾夷扇贝	kg	10.63	1163400	12,363,013.09
2020年7月	冷冻扇贝	kg	8.24	2666400	22,153,373.45
2020年8月	冷冻扇贝	kg	7.95	2659500	21,144,172.30
2020年9月	单冻半壳虾夷扇贝	kg	9.94	1233400	12,456,486.06
2020年10月	单冻半壳虾夷扇贝	kg	9.66	1278400	12,405,466.80
2020年11月	单冻半壳虾夷扇贝	kg	10.00	1209700	12,094,596.16
2020年12月	单冻半壳虾夷扇贝	kg	9.66	1263200	12,456,478.18
2020年1月	单冻半壳虾夷扇贝	kg	9.83	1339700	13,303,674.82
2020年10月	单冻半壳虾夷扇贝	kg	10.99	1219700	13,402,674.62
2020年10月	冷冻扇贝	kg	6.82	2664700	23,199,487.25
2020年10月	冷冻扇贝	kg	8.71	2719100	23,691,851.03
2020年11月	单冻半壳虾夷扇贝	kg	10.65	1247900	13,284,519.25
2020年11月	冷冻扇贝	kg	8.48	2675500	22,697,663.59
2020年11月	冷冻扇贝	kg	8.47	2703500	22,895,549.72
2020年12月	单冻半壳虾夷扇贝	kg	9.90	1422900	14,176,130.29
- Step Analysis Question:** 问题1: 实质性测试 (单选题 1.00分) 1. 下列关于存货的实质性测试结论，正确的有 ()

问题回答:

查看任务描述与资料，回答问题。为方便查看，具体题目与答案解析如下:

步骤分析

问题1

实质性测试 (单选题 1.00分)

1. 下列关于存货的实质性测试结论，正确的有 ()

- A. 单冻半壳虾夷扇贝的成本最低，销售数量最少
- B. 每次出海的捕捞时间和捕捞面积差异不大，2020年每次出海捕捞的扇贝量与2019年差不多，每次都可以把船舱装满，不足以支撑2020年扇贝减产1/3的说法
- C. 经审计，存货捕捞记录表都有相关负责人签字，无需特别关注
- D. 经询问各负责人，未发现异常

关闭答案

实质性测试(单选题 1.00 分)

1. 下列关于存货的实质性测试结论，正确的有()

A. 单冻半壳虾夷扇贝的成本最低，销售数量最少

- B.每次出海的捕捞时间和捕捞面积差异不大，2020 年每次出海捕捞的扇贝量与 2019 年差不多，每次都可以把船舱装满，不足以支撑 2020 年扇贝减产 1/3 的说法
- C.经审计，存货捕捞记录表都有相关负责人签字，无需特别关注
- D.经询问各负责人，未发现异常

关闭答案

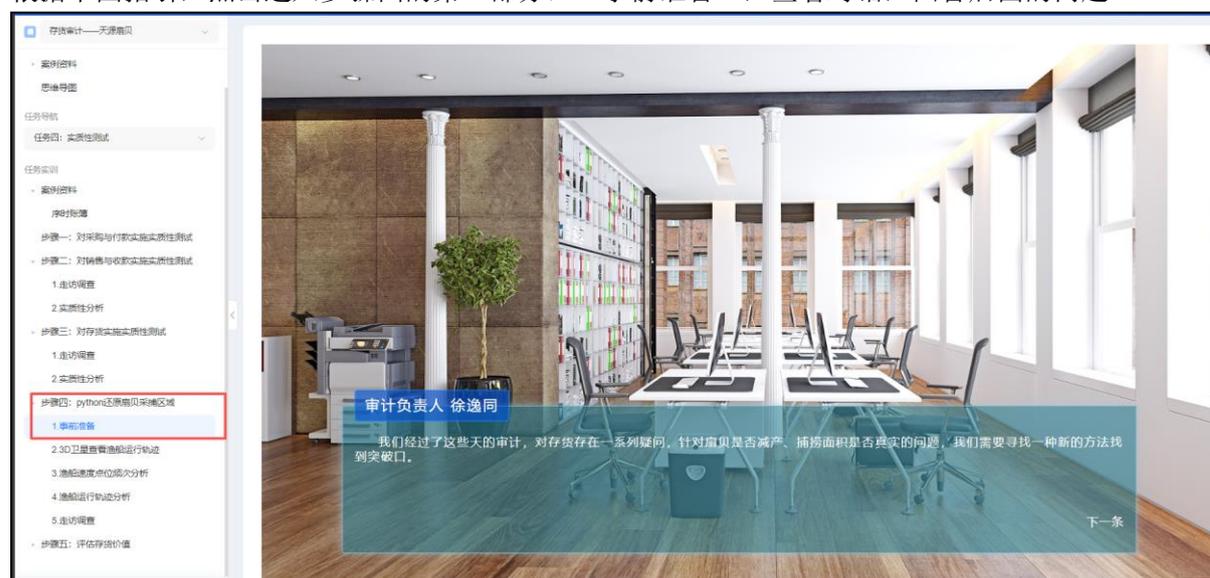
正确答案:B

解析:A、错误，单冻半壳虾夷扇贝的成本最高，销售数量最多。C、D 错误，有相关负责人签字但相关 负责人对实际情况并不具体了解，内控存在缺陷。

步骤四：python 还原扇贝采捕区域

1.事前准备

根据下图指引，点击进入步骤四的第一部分：“事前准备”，查看对话，回答后面的问题。



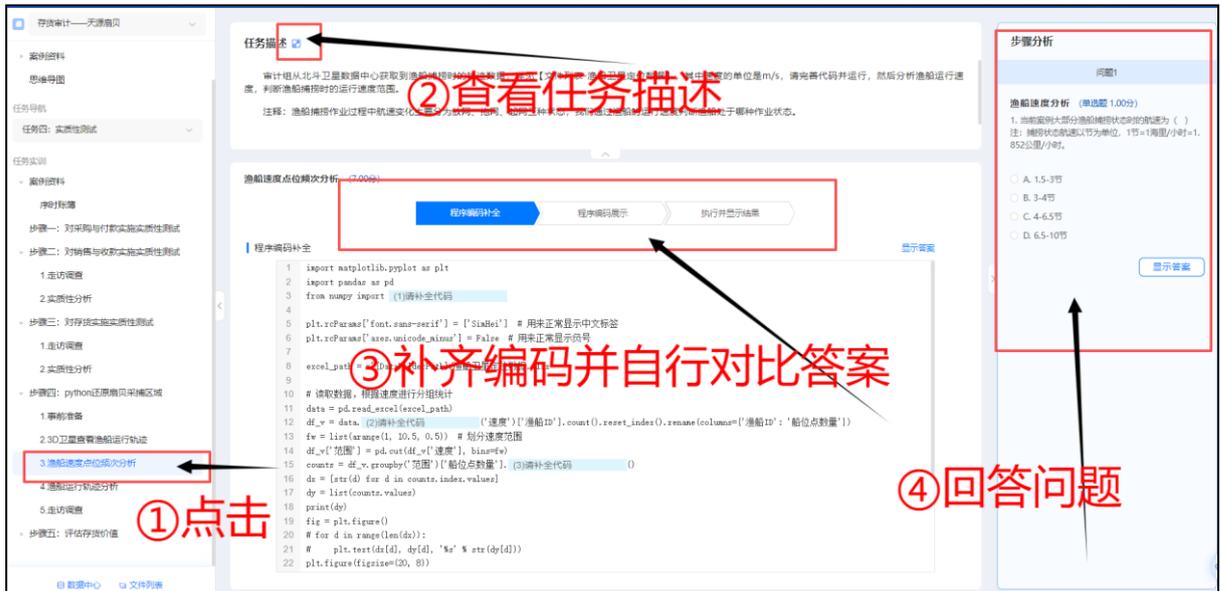
2.3D 卫星查看渔船运行轨迹

根据下图指引，点击下一个部分“3D 卫星查看渔船运行轨迹”，根据下图操作。



3. 渔船速度点位频次分析

根据下图指引，根据题目描述，补齐编码并自行对比答案，同时回答右侧步骤分析的问题。



问题回答：

查看任务描述与资料，回答问题。为方便查看，具体题目与答案解析如下：

步骤分析

问题1

渔船速度分析 (单选题 1.00分)

1. 当前案例大部分渔船捕捞状态时的航速为 ()
注：捕捞状态航速以节为单位，1节=1海里/小时=1.852公里/小时。

A. 1.5-3节

B. 3-4节

C. 4-6.5节

D. 6.5-10节

[关闭答案](#)

(单选题 1.00 分)渔船速度分析

1.当前案例大部分渔船捕捞状态时的航速为()注:捕捞状态航速以节为单位,1节=1海里/小时=1.852公里/小时。

- A. 1.5-3 节
- B. 3-4 节
- C. 4-6.5 节
- D.6.5-10 节

关闭答案

正确答案:B

解析:由图可知,速度范围在 6.5m/s-7m/s 的船只点位最多, $6.5/1.852=3.51$ 节, $7/1.852=3.78$ 节, 由此可知, 此案例大部分渔船捕捞状态的航速为 3-4 节, B。

4.渔船运行轨迹分析

根据下图指引, 点击下一任务“渔船运行轨迹分析”, 查看任务描述, 并且补齐代码, 回答右侧步骤分析问题。

任务描述

审计组从被审计单位获取到某国捕捞区域信息如下：

区域	区域1	区域2	区域3	区域4
经度	119.35	120.00	121.00	121.90
纬度	119.35	120.20	121.00	121.90

② 查看任务描述

程序代码补全

```

1 import pandas as pd # 导入pandas数据科学库
2 import matplotlib.pyplot as plt # 导入matplotlib画图模块
3
4 plt.rcParams['font.sans-serif'] = ['SimHei'] # 用来正常显示中文标签
5 plt.rcParams['axes.unicode_minus'] = False # 用来正常显示负号
6
7 # 筛选出指定经纬度范围的数据
8 excel_path = 'x' [DataFolderPath]/扇贝卫星定位数据.xlsx'
9 da_df = pd.read_excel(excel_path) # 读取待处理数据文件
10 # 根据给定的速度范围筛选出符合条件的数据
11 df_v = da_df[da_df['速度'].between(1,10)] # 补齐代码
12 # 筛选经度、纬度最大值和最小值
13 lon_max = da_df['经度'].max()
14 lon_min = da_df['经度'].min()
15 print(f'经度最大值: {lon_max}, 经度最小值: {lon_min}')
16 lat_max = da_df['纬度'].max()
17 lat_min = da_df['纬度'].min()
18 print(f'纬度最大值: {lat_max}, 纬度最小值: {lat_min}')
19 # 筛选符合条件的渔船数据
20 cond1 = df_v['经度'].between(lon_min, lon_max) # 筛选指定范围内的经度数据
21 cond2 = df_v['纬度'].between(lat_min, lat_max) # 筛选指定范围内的纬度数据
22 data = df_v[(cond1 & cond2)] # 合并筛选后的数据,构建新的数据集

```

③ 补齐代码

步骤分析

问题1

问题2

审计调查分析 (多选题 1.00分)

1. 根据卫星定位还原存货采捕区域得出的分析结论,正确的有 ()

A. 经过对比发现,公司没有如实记录采捕海域,实际采捕海域面积比账面记录大,可能存在财务舞弊

B. 利用卫星定位还原的存货捕捞区域比公司记载的面积少,说明扇贝还存在于海里,扇贝是否减产需要进一步验证

C. 卫星定位有助于解决核查水产养殖生物资产过程中,其“水域隐蔽性”“流动性强”的问题。

D. 卫星数据属于第三方数据,可信度和准确性较高,相较于公司内部口头阐述更具说服力。

显示答案

问题回答

查看任务描述与资料,回答问题。为方便查看,具体题目与答案解析如下:

步骤分析

问题1

问题2

审计调查分析 (多选题 1.00分)

1. 根据卫星定位还原存货采捕区域得出的分析结论,正确的有 ()

A. 经过对比发现,公司没有如实记录采捕海域,实际采捕海域面积比账面记录大,可能存在财务舞弊

B. 利用卫星定位还原的存货捕捞区域比公司记载的面积少,说明扇贝还存在于海里,扇贝是否减产需要进一步验证

C. 卫星定位有助于解决核查水产养殖生物资产过程中,其“水域隐蔽性”“流动性强”的问题。

D. 卫星数据属于第三方数据,可信度和准确性较高,相较于公司内部口头阐述更具说服力。

关闭答案

步骤分析

问题1

问题2

进一步推断 (多选题 1.00分)

2. 通过与财务沟通,我们获悉扇贝的养殖成本计算公式: 营业成本=扇贝养殖成本× 捕捞面积/总养殖面积,由此推断,天源公司可能通过哪些方式减少营业成本

A. 虚减扇贝养殖总成本

B. 虚增总的养殖面积

C. 少记捕捞面积

D. 对未捕捞过的区域结转成本

关闭答案

问题 1: 审计调查分析(多选题 1.00 分)

1.根据卫星定位还原存货采捕区域得出的分析结论正确的有()

A.经过对比发现,公司没有如实记录采捕海域,实际采捕海域面积比账面记录大,可能存在财务舞弊

B.利用卫星定位还原的存货捕捞区域比公司记载的面积少,说明扇贝还存在于海里扇贝是否减产需要进一步验证

卫星定位有助于解决核查水产养殖生物资产过程中,其“水域隐蔽性”“流动性强”的问题。卫星数据属于第三方数据,可信度和准确性较高,相较于公司内部口头阐述更具说服力。

关闭答案

正确答案:ACD

解析:

A, 正确, 由图可知, 账面记录区域小于实际捕区域, 因此可能存在财务造假。B, 错误。由图可知, 还原的区域大于公司记载的区域。C, 正确, 卫星定位相对人工盘点有助于解决核查水产养殖生物资产过程中, 其“水域隐蔽性”“流动性强”的问题。D, 正确, 外部独立证据可靠性>内部证据可靠性:来自被审计单位外部的独立第三方证据(如银行询证函回函、供应商发票)更可靠, 因不易被篡改, 卫星数据属于第三方数据, 可信度和准确性较高。

问题 2: 进一步推断(多选题 1.00 分)

2.通过与财务沟通,我们获悉扇贝的养殖成本计算公式:营业成本=扇贝养殖成本 x 捕捞面积/总养殖面积,由此推断,天源公司可能通过哪些方式减少营业成本

- A.虚减扇贝养殖总成本
- B.虚增总的养殖面积
- C.少记捕捞面积
- D.对未捕捞过的区域结转成本

关闭答案

正确答案:ABC

解析:ABC 正确, D 对未捕捞的区域结转成本属于扩大捕捞面积, 增加成本结转, 而非减少营业成本。

5. 走访调查

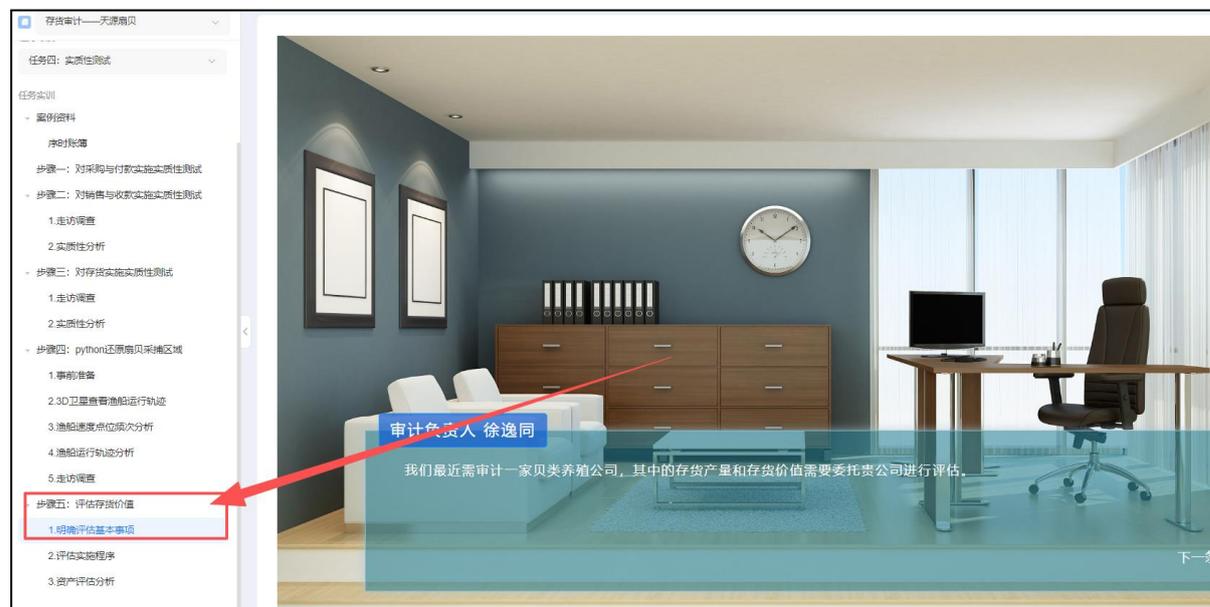
点击第五部分“走访调查”, 查看对话: 如下图所示



步骤五：评估存货价值

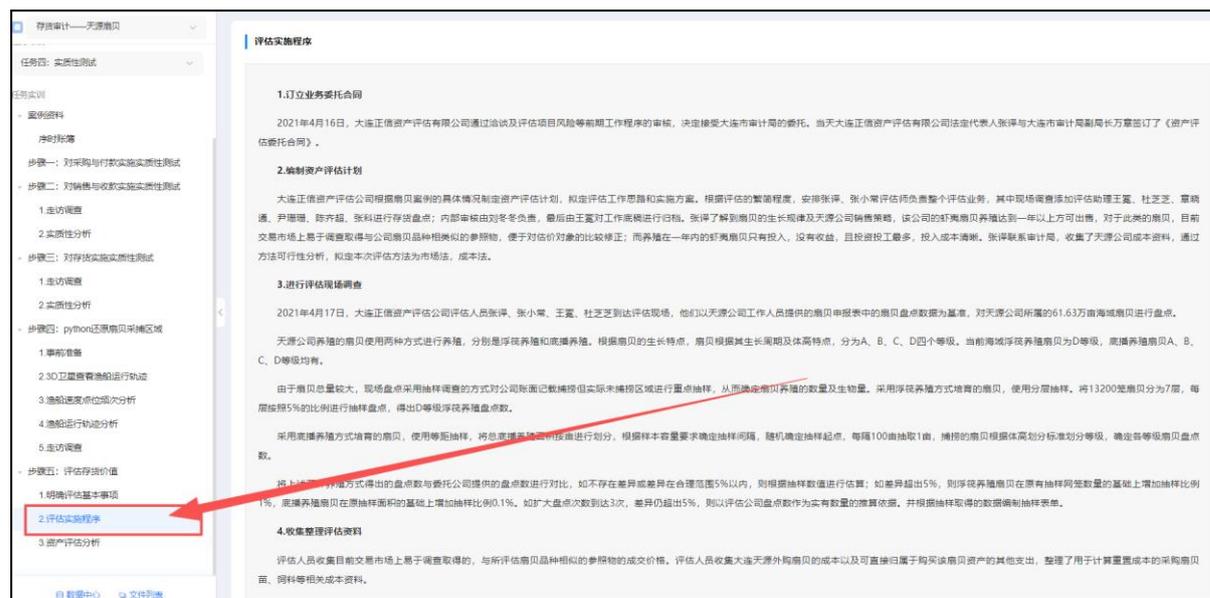
1. 明确评估基本事项

根据下图指引，点击步骤五“评估存货价值”，点击第一部分：明确评估基本事项。查看对话。



2. 评估实施程序

根据下图指引，点击步骤五“评估存货价值”，点击第二部分：评估实施程序。查看步骤。



3. 资产评估分析

根据下图指引，点击第三部分“资产评估分析”的步骤一，查看资产评估报告。



根据下图指引，点击第三部分“资产评估分析”的步骤二，查看任务描述，制作 excel 答案并上传。并且回答右侧问题。



问题回答:

查看任务描述与资料,回答问题。为方便查看,具体题目与答案解析如下:

步骤分析					
问题1	问题2				
<p>资产评估结果分析 (多选题 1.00分)</p> <p>1. 下列资产评估分析结论,正确的有()。</p> <ul style="list-style-type: none"><input checked="" type="checkbox"/> A. 根据评估得出,截至评估基准日,大连天源贝类养殖股份有限公司扇贝总生物量104648119 kg,未发现减产迹象<input checked="" type="checkbox"/> B. 根据评估得出,截至评估基准日,大连天源贝类养殖股份有限公司扇贝总数量为1716101112枚,未发现减产迹象<input type="checkbox"/> C. 根据评估得出,截至评估基准日,大连天源贝类养殖股份有限公司扇贝评估值为1159780000元,与账面记载金额一致,无异常<input checked="" type="checkbox"/> D. 根据评估得出,截至评估基准日,大连天源贝类养殖股份有限公司扇贝评估值为1159780000元,与账面记载金额不一致,需重点关注 <p style="text-align: right;">关闭答案</p>	<p>步骤分析</p> <thead><tr><th>问题1</th><th>问题2</th></tr></thead> <tbody><tr><td><p>资产评估结果分析 (多选题 1.00分)</p><p>2. 下列资产评估分析结论,正确的有()。</p><ul style="list-style-type: none"><input checked="" type="checkbox"/> A. 大连天源贝类养殖股份有限公司华奥号 6 1.63 万亩海域养殖的浮筏养殖及底播养殖扇贝评估增值<input type="checkbox"/> B. 大连天源贝类养殖股份有限公司计提的存货跌价准备金额符合会计制度相关规定,无不合理之处<input checked="" type="checkbox"/> C. 经评估,大连天源贝类养殖股份有限公司对外宣称减产的虾夷扇贝存活在海底,审计人员需重点关注<input type="checkbox"/> D. 经评估,大连天源贝类养殖股份有限公司对外宣称减产的虾夷扇贝存活在海底,其虚减账面存货的说法不合理<p style="text-align: right;">关闭答案</p></td><td></td></tr></tbody>	问题1	问题2	<p>资产评估结果分析 (多选题 1.00分)</p> <p>2. 下列资产评估分析结论,正确的有()。</p> <ul style="list-style-type: none"><input checked="" type="checkbox"/> A. 大连天源贝类养殖股份有限公司华奥号 6 1.63 万亩海域养殖的浮筏养殖及底播养殖扇贝评估增值<input type="checkbox"/> B. 大连天源贝类养殖股份有限公司计提的存货跌价准备金额符合会计制度相关规定,无不合理之处<input checked="" type="checkbox"/> C. 经评估,大连天源贝类养殖股份有限公司对外宣称减产的虾夷扇贝存活在海底,审计人员需重点关注<input type="checkbox"/> D. 经评估,大连天源贝类养殖股份有限公司对外宣称减产的虾夷扇贝存活在海底,其虚减账面存货的说法不合理 <p style="text-align: right;">关闭答案</p>	
问题1	问题2				
<p>资产评估结果分析 (多选题 1.00分)</p> <p>2. 下列资产评估分析结论,正确的有()。</p> <ul style="list-style-type: none"><input checked="" type="checkbox"/> A. 大连天源贝类养殖股份有限公司华奥号 6 1.63 万亩海域养殖的浮筏养殖及底播养殖扇贝评估增值<input type="checkbox"/> B. 大连天源贝类养殖股份有限公司计提的存货跌价准备金额符合会计制度相关规定,无不合理之处<input checked="" type="checkbox"/> C. 经评估,大连天源贝类养殖股份有限公司对外宣称减产的虾夷扇贝存活在海底,审计人员需重点关注<input type="checkbox"/> D. 经评估,大连天源贝类养殖股份有限公司对外宣称减产的虾夷扇贝存活在海底,其虚减账面存货的说法不合理 <p style="text-align: right;">关闭答案</p>					

资产评估结果分析(多选题 1.00 分)

(1) 下列资产评估分析结论,正确的有()。

根据评估得出,截至评估基准日,大连天 A.源贝类养殖股份有限公司扇贝总生物量 104648119 kg,未发现减产迹象

B 根据评估得出,截至评估基准日,大连天源贝类养殖股份有限公司扇贝总数量为 1716101112 枚,未发现减产迹象

C.根据评估得出,截至评估基准日,大连天源贝类养殖股份有限公司扇贝评估值为 1159780000 元,与账面记载金额一致,无异常

D.根据评估得出,截至评估基准日,天源贝类养殖股份有限公司扇贝评估值为 1159780000 元,与账面记载金额不一致需重点关注

正确答案:ABD

解析:根据评估得出,截至评估基准日,大连天源贝类养殖股份有限公司扇贝总生物量 104648119kg,未发现减产迹象;根据评估得出,截至评估基准日,大连天源贝类养殖股份有限公司扇贝总数量为 1716101112 枚,未发现减产迹象;根据评估得出,截至评估基准日,大连天源贝类养殖股份有限公司扇贝评估值为 1159780000 元,与账面记载金额不一致,需重点关注。

问题 2: 资产评估结果分析(多选题 1.00 分)

2.下列资产评估分析结论，正确的有()

A.大连天源贝类养殖股份有限公司华奥岛 61.63 万亩海域养殖的浮养殖及底播养殖扇贝评估增值

B.大连天源贝类养殖股份有限公司计提的存货跌价准备金额符合会计制度相关规定，无合理之处

经评估，大连天源贝类养殖股份有限公司对外宣称减产的虾夷扇贝存活在海底，审计人员需重点关注

D.经评估，大连天源贝类养殖股份有限公司对外宣称减产的虾夷扇贝存活在海底，其虚减账面存货的说法不合理

关闭答案

正确答案:AC

解析:大连天源贝类养殖股份有限公司华奥岛 61.63 万亩海域养殖的浮筏养殖及底播养殖扇贝评估增值，增值原因是误判海域扇贝已经死亡，但实质上扇贝仍存活在海洋中，天源贝类确实有点减少账面存货。AC 正确，BD 错误；

任务五：发表审计结论

根据下图指引点击任务导航栏，点击任务五，回答相关问题

The screenshot shows a web interface for an audit course. On the left, there is a sidebar with a 'Task Navigation' section. The fifth item, 'Task 5: Issuing Audit Conclusions', is highlighted with a red box and a red arrow. The main content area is titled 'Substantive Audit Procedures' and contains three questions:

- 本次审计过程中，进行了哪些审计程序 (多选题 1分)
 - A. 种苗采购检查
 - B. 资产价值评估
 - C. 捕捞区域还原
 - D. 银行流水检查
- 此次审计发现的主要问题有 (多选题 1分)
 - A. 天源公司近年来营业收入长期下滑，通过舞弊试图美化财务报表
 - B. 2020年并未存在异常海水温度，冷水团导致扇贝减产的理由为虚构
 - C. 财务部计算成本过程中的捕捞面积数据未经严格核实，存在内控缺陷
 - D. 实际捕捞区域小于账面捕捞区域，虚减成本，操纵利润
 - E. 扇贝减产对外虚减存货，实际对外宣称减产的扇贝依然存活在海底
- 根据上述情况，提出如下建议 (多选题 1分)
 - A. 进一步规范财务管理，严格落实大连市廉洁从业规定
 - B. 完善内控制度，健全部门间职责分离，推动业财结合
 - C. 积极降本增效，扩大捕捞区域，提高扇贝存活率
 - D. 调增存货及成本，如实反映应收下滑，利润降低的情况

问题回答：

查看任务描述，回答问题。为方便查看，具体题目与答案解析如下：

任务描述：

根据上述审计步骤，发表审计结论

实质性审计程序

1. 本次审计过程中，进行了哪些审计程序（多选题 1分） ?

- A. 种苗采购核查
- B. 资产价值评估
- C. 捕捞区域还原
- D. 银行流水核查

2. 此次审计发现的主要问题有（多选题 1分） ?

- A. 天源公司近年来营业收入长期下滑，通过舞弊试图美化财务报表
- B. 2020年并未存在异常海水温度，冷水团导致扇贝减产的理由为虚构
- C. 财务部计算成本过程中的捕捞面积数据未经严格核实，存在内控缺陷
- D. 实际捕捞区域大小于账面捕捞区域，虚减成本，操纵利润
- E. 扇贝减产对外虚减存货，实际对外宣称减产的扇贝依然存活在海底

3. 根据上述情况，提出如下建议（多选题 1分） ?

- A. 进一步规范财务管理，严格落实大连市廉洁从业规定
- B. 完善内控制度，健全部门间职责分离，推动业财结合
- C. 积极降本增效，扩大采捕区域，提高扇贝存活率
- D. 调增存货及成本，如实反映应收下滑，利润降低的情况

本次审计过程中，进行了哪些审计程序（）(多选题 1分)

- A.种苗采购核查
- B.资产价值评估
- C.捕捞区域还原
- D.银行流水核查

正确答案:ABC

试题解析:ABC 分别对应:对采购与付款实施实质性测试、评估存货价值、python 还原扇贝采捕区域，D 未在审计过程中体现

2.此次审计发现的主要问题有（）(多选题 1分)

- A.天源公司近年来营业收入长期下滑，通过舞弊试图美化财务报表
- B.2020 年并未存在异常海水温度，冷水团导致扇贝减产的理由为虚构
- C.财务部计算成本过程中的捕捞面积数据未经严格核实，存在内控缺陷
- D.实际捕捞区域大小于账面捕捞区域，虚减成本，操纵利润
- E.扇贝减产对外虚减存货，实际对外宣称减产的扇贝依然存活在海底

正确答案:ABCE

试题解析:D、实际捕捞区域大于账面捕捞区域，虚减成本，操纵利润，错误。

3、根据上述情况，提出如下建议()(多选题 1分)

- A.进一步规范财务管理，严格落实大连市廉洁从业规定
- B.完善内控制度，健全部门间职责分离，推动业财结合
- C.积极降本增效，扩大采捕区域，提高扇贝存活率
- D.调增存货及成本，如实反映应收下滑，利润降低的情况

正确答案:ABD

试题解析:C，天源是通过少记捕捞面积，对已经捕捞过的区域不结转成本，从而操纵利润，并不存在养殖成本增加导致利润下滑，冷水团导致扇贝死亡，因此并无需建议天源公司“积极降本增效，扩大采捕区域，提高扇贝存活率”

产业案例 AI 模型体验

02. AI模型体验

任务一 基础模型与基础参数微调体验

一、预处理语料

(一) 步骤 1: 基础信息判断

待处理的语料文件已内置在平台中，表格中共有 1663 组信息，请根据提示补充 Python 代码，完成语料表的基本信息了解。了解表格基本信息可用 info 函数。

下表为预览内容：

instruction	input	output
你是一个会计准则领域的专家助手。请根据你的专业知识，准确回答以下问题。	固定资产在什么情况下应当予以终止确认？	固定资产在以下情况下应当予以终止确认：（一）该固定资产处于处置状态。（二）该固定资产预期通过使用或处置不能产生经济利益。
你是一个会计准则领域的专家助手。请根据你的专业知识，准确回答以下问题。	固定资产在什么条件下才能予以确认？	固定资产同时满足以下两个条件才能予以确认：（一）与该固定资产有关的经济利益很可能流入企业；（二）该固定资产的

		成本能够可靠地计量。
你是一个会计准则领域的专家助手。请根据你的专业知识，准确回答以下问题。	关键管理人员的定义是什么？	关键管理人员是指有权力并负责计划、指挥和控制企业活动的人员。
你是一个会计准则领域的专家助手。请根据你的专业知识，准确回答以下问题。	关联方交易的类型通常包括哪些？	关联方交易的类型通常包括转移资源、劳务或义务的行为。
你是一个会计准则领域的专家助手。请根据你的专业知识，准确回答以下问题。	关联方交易披露的内容包括哪些？	关联方交易披露的内容包括交易的金额、未结算项目的金额、条款和条件，以及有关提供或取得担保的信息、未结算应收项目的坏账准备金额和定价政策。

请结合任务描述内容补全代码，执行后的结果如下图所示。

执行状态: ● 运行结束

执行结果:

```

输出:
数据表基本信息:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1663 entries, 0 to 1662
Data columns (total 3 columns):
#   Column      Non-Null Count  Dtype
---  ---
0   instruction  1598 non-null  object
1   input       1663 non-null  object
2   output      1663 non-null  object
dtypes: object(3)
memory usage: 39.1+ KB
None

检查是否有缺失值:
instruction  65
input       0
output      0
dtype: int64

运行结束

```

图 1

挖空考核答案

(1) info

(二) 步骤 2: 语料表预处理

根据上一步骤的信息，针对给出的数据表进行预处理。即使用众数填补缺失值，去掉重复值。提示：查找重复值可用 duplicated 函数，删除行可用 drop 函数。

请结合任务描述内容补全代码，执行后的结果如下图所示。

执行状态: ● 运行结束

执行结果:

```
输出:

填补后的数据表基本信息:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 1663 entries, 0 to 1662
Data columns (total 3 columns):
#   Column      Non-Null Count  Dtype
---  -
0   instruction 1663 non-null  object
1   input       1663 non-null  object
2   output      1663 non-null  object
dtypes: object(3)
memory usage: 39.1+ KB
None

检查是否有重复的行:
有 53 行是重复的。

去重后的数据表已另存

运行结束

其他:
(1) 文件: 企业会计准则问答语料-处理后.xlsx 
```

图 2

企业会计准则问答语料-处理后文件内容节选如下:

4	A	B	C
1	instruction	input	output
2	你是一个会计准则领域的专家助手。请根据你的专业知识，准确回答以下问题。	固定资产在什么情况下应当予以终止确认？	固定资产在以下情况下应当予以终止确认：（一）该固定资产处于处置状态。（二）该固定资产预期通过使用或处置不能产生经济利益。
3	你是一个会计准则领域的专家助手。请根据你的专业知识，准确回答以下问题。	固定资产在什么条件下才能予以确认？	固定资产同时满足以下两个条件才能予以确认：（一）与该固定资产有关的经济利益很可能流入企业；（二）该固定资产的成本能够可靠地计量。
4	你是一个会计准则领域的专家助手。请根据你的专业知识，准确回答以下问题。	关键管理人员的定义是什么？	关键管理人员是指有权力并负责计划、指挥和控制企业活动的人员。
5	你是一个会计准则领域的专家助手。请根据你的专业知识，准确回答以下问题。	关联方交易的类型通常包括哪些？	关联方交易的类型通常包括转移资源、劳务或义务的行为。

图 3

挖空考核答案

(1) duplicated

(2) drop

(3) 企业会计准则问答语料-处理后

(三) 步骤 3: 语料格式转换

在进行大语言模型微调前，需要将现有的语料进行格式转换，转成 json 格式。请使用代码完成这一操作，注意文件名需填写上一步骤中去重后的数据表。

请结合任务描述内容补全代码，执行后的结果如下图所示。

执行状态: ● 运行结束

执行结果:

输出:
转换完成! 数据已保存

运行结束

其他:

(1) 文件: 企业会计准则问答语料-处理后.xlsx [下载](#)

(2) 文件: accounting_standards.json [下载](#)

图 4

accounting_standard.json 文件内容如下（节选）：

```

1  {
2  {
3    "instruction": "你是一个会计准则领域的专家助手。请根据你的专业知识，准确回答以下问题。",
4    "input": "固定资产在什么情况下应当予以终止确认？",
5    "output": "固定资产在以下情况下应当予以终止确认：（一）该固定资产处于处置状态。（二）该固定资产预期通过使用或处置不能产生经济利益。",
6  },
7  {
8    "instruction": "你是一个会计准则领域的专家助手。请根据你的专业知识，准确回答以下问题。",
9    "input": "固定资产在什么条件下才能予以确认？",
10   "output": "固定资产同时满足以下两个条件才能予以确认：（一）与该固定资产有关的经济利益很可能流入企业；（二）该固定资产的成本能够可靠地计量。",
11  },
12  {
13   "instruction": "你是一个会计准则领域的专家助手。请根据你的专业知识，准确回答以下问题。",
14   "input": "关键管理人员的定义是什么？",
15   "output": "关键管理人员是指有权力并负责计划、指挥和控制企业活动的人员。",
16  },

```

图 5

挖空考核答案

(1) 企业会计准则问答语料-处理后

二、微调模型并测试

(一) 模型介绍：Qwen2.5-7B-Instruct

据官方信息显示，Qwen2.5 系列语言模型在最新的大规模数据集上进行了预训练，该数据集包含多达 18T tokens。相较于 Qwen2，Qwen2.5 获得了显著更多的知识 (MMLU: 85+)，并在编程能力 (HumanEval 85+) 和数学能力 (MATH 80+) 方面有了大幅提升。此外，新模型在指令执行、生成长文本 (超过 8K 标记)、理解结构化数据 (例如表格) 以及生成结构化输出特别是 JSON 方面取得了显著改进。Qwen2.5 模型总体上对各种 system prompt 更具适应性，增强了角色扮演实现和聊天机器人的条件设置功能。与 Qwen2 类似，Qwen2.5 语言模型支持高达 128K tokens，并能生成最多 8K tokens 的内容。它们同样保持了对包括中文、英文、法文、西班牙文、葡萄牙文、德文、意大利文、俄文、日文、韩文、越南文、泰文、阿拉伯文等 29 种以上语言的支持。

Qwen2.5-7B-Instruct 型号含义：

Qwen2.5: 模型系列名称。7B: 参数量为 70 亿 (7 Billion)。这个规模在性能、效率和硬件需求之间取得了很好的平衡，非常适合企业和个人开发者进行部署和应用。Instruct: 这意味着该模型是经过“指令微调” (Instruction Tuning) 的版本。它经过了大量高质量指令-回答数据集的训练，能够更好地理解用户的意图，并以有帮助、准确、安全的方式回应，而不是一个仅完成训练 (Base) 的“空白”模型。

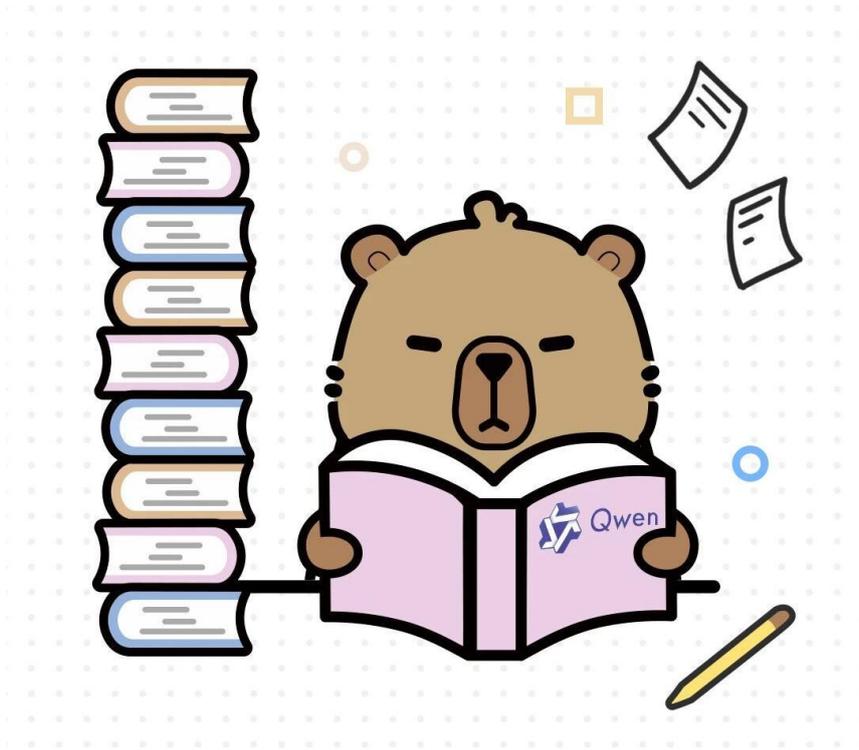


图 6

（二）微调模型

任务描述：

1. 使用默认参数完成 1 次大语言模型微调。
2. 微调过程中使用的语料文件为转换后的 json 文件，请将文件上传至指定位置。
3. 微调结束后，请分别选取微调前的模型与微调后的模型进行对话，以对话形式进行测试，评估模型微调效果。

【此处提供的“会计语料库”、“营销语料库”仅供参考，当前任务使用的语料请以任务描述中的为准。】

请结合任务要求设置参数，设置完毕后的效果如下：

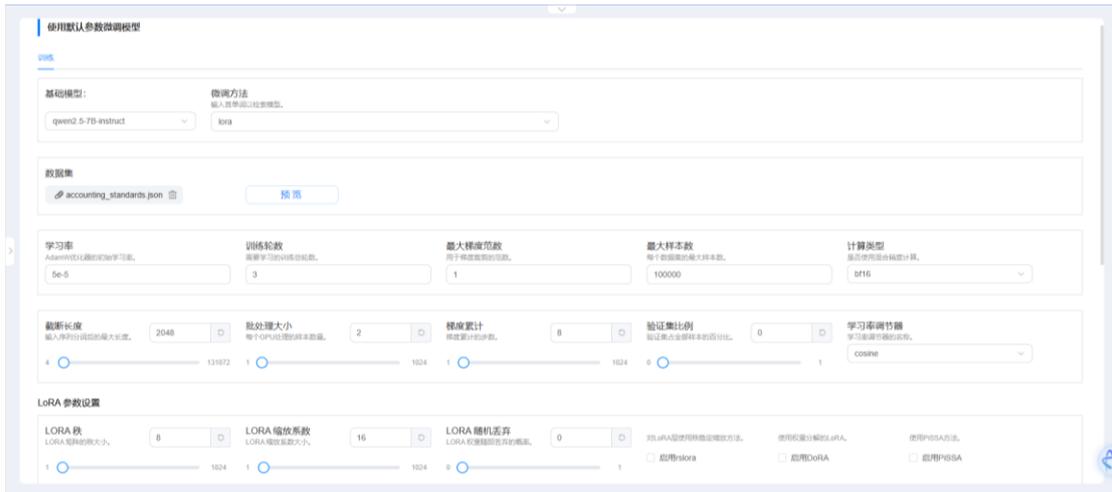


图 7

上传的语料可点击预览按钮进行查看，查看效果如下：

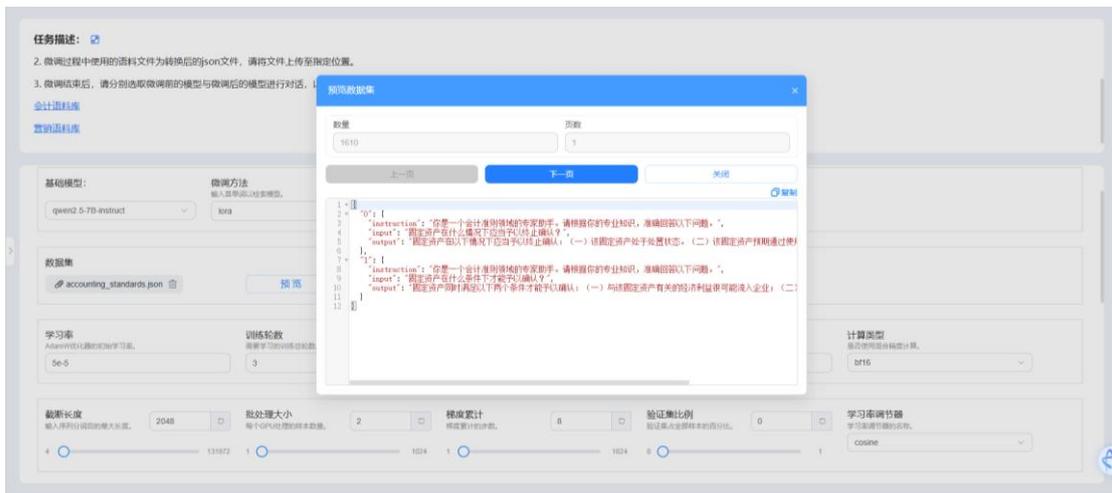


图 8

运行结果

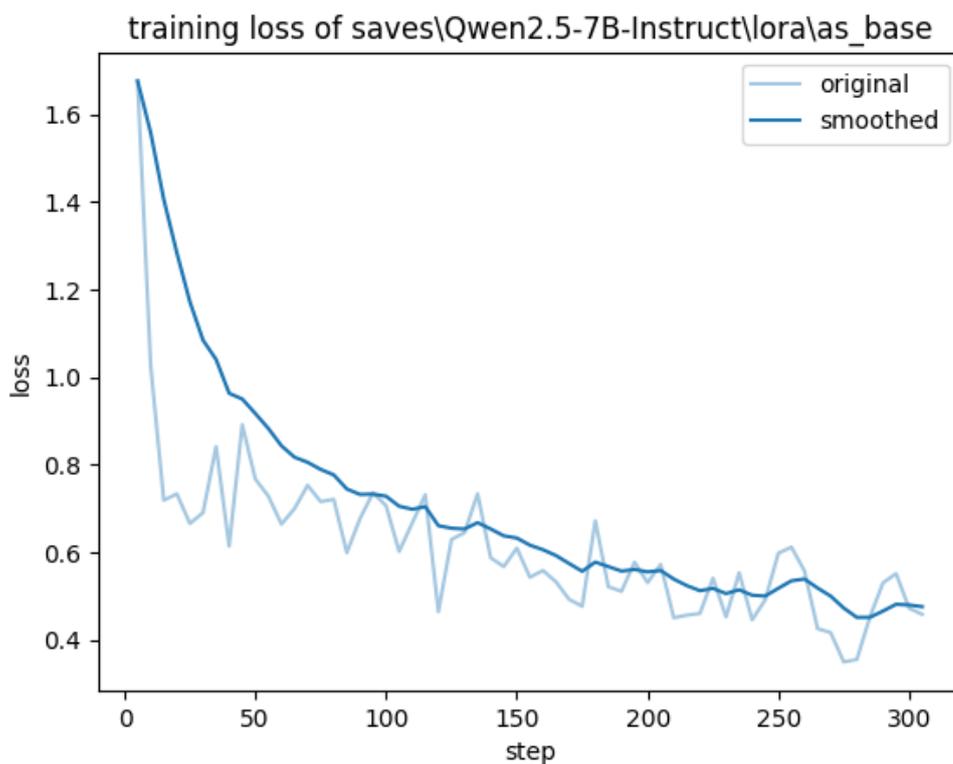


图 9

(三) 测试模型

1. 步骤一：测试过程说明

任务描述

请根据以下体验要点提示，依次向机器人发出提问并查看机器人的回答。对比分析 AI 回答，根据体验结果完成模型对应的体验分析报告。

测试体验

1) 测试要点

- ①向 AI 模型输入语料中的问题，查看 AI 的回答内容；
- ②向 AI 模型输入语料中的问题，并于提示语处输入相关提示语，查看 AI 的回答内容；
- ③向 AI 模型输入语料中的问题，得到回答后，在不清空历史记录的情况下，就同一个相同问题连续进行发起 3 次问答，查看 AI 的回答内容；
- ④向 AI 模型输入语料中的问题，得到回答后，在清空历史记录的情况下，

就同一个相同问题连续进行发起 3 次问答，查看 AI 的回答内容；

⑤向 AI 模型连续 5 次以上输入相同的问题，查看你 AI 的回答内容是否保持一致，没有出现明显的波动或变化；

⑥向 AI 模型输入语料中的问题，得到答复后，再次输入同义描述的内容，查看 AI 的回答内容；

⑦向 AI 模型输入语料中的问题后，再次输入自创的同义描述，查看 AI 的回答内容（自创内容不与同义描述相同）；

⑧请根据所选的问题，适当丰富背景描述，向 AI 模型输入含有背景的问题，查看 AI 的回答内容；

⑨请根据所选的问题，进一步丰富背景描述，向 AI 模型输入丰富后的问题，查看 AI 的回答内容。

2) 测试分析

请根据测试体验的内容，结合提供的体验分析报告模板，完成模型对应的体验分析报告。

2. 步骤二：进行测试

请使用上一环节生成的模型与基础模型完成对话预测。提示：最大生成长度可设置为 1024，top-p 为 0.7，温度系数为 0.95。

【进行对话测试前，请先设置右侧的参数，以免回复内容不完整影响测试结论】

测试对话过程与效果如下所示：



图 10

三、人工智能体验测试报告

请下载参考模板，完成后以截图形式上传

点击链接即可下载，下载后的文件打开可看到两个 sheet。

基本信息记录表如下所示：

基本信息记录				
AI模型名称				
体验日期				
使用业务场景				
AI模型体验分析				
内容名称	测试内容	结果	分数设置	得分
		满足测试内容的打√ 不满足测试内容的打×		满足测试内容满分； 不满足测试内容零分
准确性测试	AI模型能否输出相关内容；		5	
	AI模型输出内容是否正确；		5	
耦合性测试	AI模型在第一次输出结果后，不清空历史记录的情况下，继续以同一个问题描述进行提问，其回答内容是否与第一次结果相同；		7.5	
	AI模型在第一次输出结果后，在清空历史记录的情况下，继续以同一个问题描述进行提问，其回答内容是否与第一次结果相同；		7.5	
稳定性测试	AI模型在连续5次以上回答相同问题时，其回答是否保持一致，没有出现明显的波动或变化；		10	
理解力测试	AI模型在面对不同表述方式但问题本质相同的情况下，能否给出正确的答案；（基于提供的备选，语料库具备的）		12.5	
	AI模型在面对不同表述方式但问题本质相同的情况下，能否给出正确的答案；（基于自创，语料库之外的）		12.5	
推理力测试	AI模型在面对丰富的业务场景描述，且描述内容与语料库重合度高，能否给出正确答案；		15	
	AI模型在面对丰富的业务场景描述，且描述内容与语料库重合度低，能否给出正确答案；		25	
合计			100	0
AI模型体验总结				
请根据AI模型的体验结果以及体验分析情况，简单描述AI模型的输出内容，并指出AI模型是否按照预期输出了所有必要的信息。				

图 11

测试结果汇总表如下所示：



图 12

答案仅供参考，请以实际测试体验为准。

基本信息记录				
AI模型名称				
体验日期				
使用业务场景				
AI模型体验分析				
内容名称	测试内容	结果	分数设置	得分
		满足测试内容的打√/ 不满足测试内容的打×		满足测试内容满分； 不满足测试内容零分
准确性测试	AI模型能否输出内容为会计分录；	√	5	5
	AI模型能否正确输出的会计分录；	×	5	0
耦合性测试	AI模型在第一次输出结果后，不清空历史记录的情况下，继续以同一个问题描述进行提问，其回答内容是否与第一次结果相同；	√	7.5	7.5
	AI模型在第一次输出结果后，在清空历史记录的情况下，继续以同一个问题描述进行提问，其回答内容是否与第一次结果相同；	×	7.5	0
稳定性测试	AI模型在连续5次以上回答相同问题时，其回答是否保持一致，没有出现明显的波动或变化；	√	10	10
理解力测试	AI模型在面对不同表述方式但问题本质相同的情况下，能否给出正确的答案；（基于提供的备选，语料库具备的）	×	12.5	0
	AI模型在面对不同表述方式但问题本质相同的情况下，能否给出正确的答案；（基于自创，语料库之外的）	×	12.5	0
推理力测试	AI模型在面对丰富的业务场景描述，且描述内容与语料库重合度高，能否给出正确答案；	√	15	15
	AI模型在面对丰富的业务场景描述，且描述内容与语料库重合度低，能否给出正确答案；	×	25	0
合计			100	37.5
AI模型体验总结				
请根据AI模型的体验结果以及体验分析情况，简单描述AI模型的输出内容，并指出AI模型是否按照预期输出了所有必要的信息。				

图 13



图 14

任务二 高训练轮次模型体验

一、任务介绍

在上一任务中已使用默认参数组合进行微调体验，现需要适当修改个别参数的值，体验不同参数对模型效果的影响。

二、微调模型并测试

(一) 微调模型

1. 参考默认参数，将模型微调参数中的训练轮次修改为 8，其余参数不变。
2. 微调过程中使用的语料文件为转换后的 json 文件，请将文件上传至指定位置。
3. 微调结束后，请选取此次微调后的模型进行对话，以对话形式进行测试，

评估模型微调效果。

请结合任务要求完成参数设定，设置完毕后的界面如下所示：

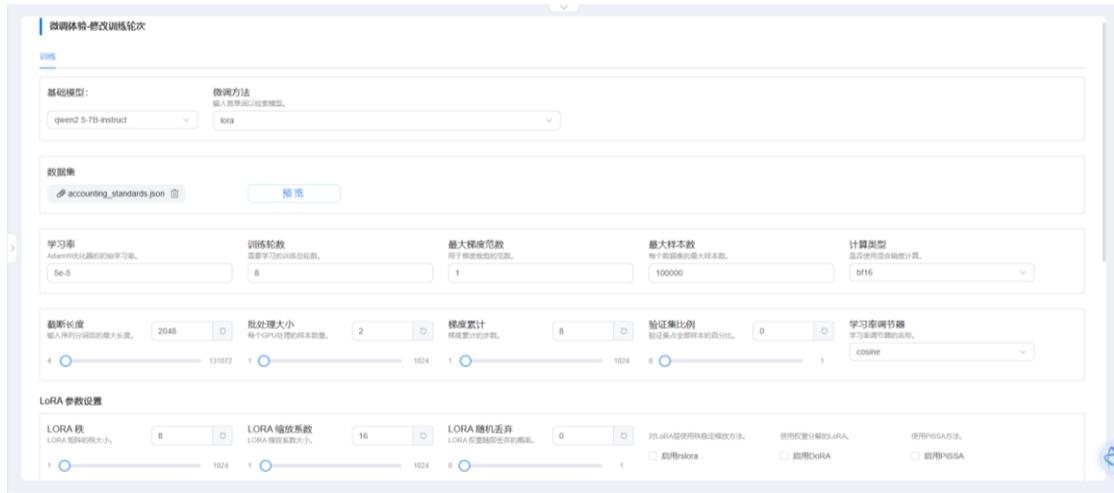


图 15

训练完毕后的损失函数图如下所示：



图 16

(二) 测试模型

1. 步骤 1: 测试说明

任务描述

请根据以下体验要点提示，结合步骤 2 所选的业务场景，依次向机器人发出提问并查看机器人的回答。对比分析 AI 回答与手写的会计分录，根据体验结果完成模型对应的体验分析报告。

测试体验

1) 测试要点

- ①向 AI 模型输入场景描述的内容，查看 AI 的回答内容；
- ②向 AI 模型输入场景描述的内容，并于提示语处输入相关提示语，查看 AI 的回答内容；
- ③向 AI 模型输入场景描述的内容，得到回答后，在不清空历史记录的情况下，就同一个相同问题连续进行发起 3 次问答，查看 AI 的回答内容；
- ④向 AI 模型输入场景描述的内容，得到回答后，在清空历史记录的情况下，就同一个相同问题连续进行发起 3 次问答，查看 AI 的回答内容；
- ⑤向 AI 模型连续 5 次以上输入相同的场景描述，查看你 AI 的回答内容是否保持一致，没有出现明显的波动或变化；
- ⑥向 AI 模型输入场景描述的内容，得到答复后，再次输入同义描述的内容，查看 AI 的回答内容；
- ⑦向 AI 模型输入场景描述的内容后，再次输入自创的同义描述，查看 AI 的回答内容（自创内容不与同义描述相同）
- ⑧请给予所选的业务场景描述，丰富场景描述（要求和原来的场景描述重合度高，或者包含原来的场景描述内容）向 AI 模型输入丰富后的场景描述，查看 AI 的回答内容；
- ⑨请给予所选的业务场景描述，丰富场景描述（要求和原来的场景描述重合度低）向 AI 模型输入丰富后的场景描述，查看 AI 的回答内容；

2) 测试分析

请根据测试体验的内容，结合提供的体验分析报告模板，完成模型对应的体验分析报告。

2. 步骤 2: 进行测试

请使用上一环节生成的模型完成对话预测。提示：最大生成长度可设置为 1024，top-p 为 0.7，温度系数为 0.95。

【进行对话测试前，请先设置右侧的参数，以免回复内容不完整影响测试结论】

测试对话过程与效果如下所示：



图 17

三、AI 体验测试报告

请下载参考模板，完成后以截图形式上传

点击链接即可下载，下载后的文件打开可看到两个 sheet。

基本信息记录表如下所示：

基本信息记录				
AI模型名称				
体验日期				
使用业务场景				
AI模型体验分析				
内容名称	测试内容	结果	分数设置	得分
		满足测试内容的打√/ 不满足测试内容的打×		满足测试内容满分； 不满足测试内容零分
准确性测试	AI模型能否输出相关内容；		5	
	AI模型输出内容是否正确；		5	
耦合性测试	AI模型在第一次输出结果后，不清空历史记录的情况下，继续以同一个问题描述进行提问，其回答内容是否与第一次结果相同；		7.5	
	AI模型在第一次输出结果后，在清空历史记录的情况下，继续以同一个问题描述进行提问，其回答内容是否与第一次结果相同；		7.5	
稳定性测试	AI模型在连续5次以上回答相同问题时，其回答是否保持一致，没有出现明显的波动或变化；		10	
理解力测试	AI模型在面对不同表述方式但问题本质相同的情况下，能否给出正确的答案；（基于提供的备选，语料库具备的）		12.5	
	AI模型在面对不同表述方式但问题本质相同的情况下，能否给出正确的答案；（基于自创，语料库之外的）		12.5	
推理力测试	AI模型在面对丰富的业务场景描述，且描述内容与语料库重合度高，能否给出正确答案；		15	
	AI模型在面对丰富的业务场景描述，且描述内容与语料库重合度低，能否给出正确答案；		25	
合计			100	0
AI模型体验总结				
请根据AI模型的体验结果以及体验分析情况，简单描述AI模型的输出内容，并指出AI模型是否按照预期输出了所有必要的信息。				

图 18

测试结果汇总表如下所示：



图 19

答案仅供参考，请以实际测试体验为准。

以下为参考答案

基本信息记录				
AI模型名称				
体验日期				
使用业务场景				
AI模型体验分析				
内容名称	测试内容	结果	分数设置	得分
		满足测试内容的打√/ 不满足测试内容的打×		满足测试内容满分； 不满足测试内容零分
准确性测试	AI模型能否输出内容为会计分录；	√	5	5
	AI模型能否正确输出的会计分录；	×	5	0
耦合性测试	AI模型在第一次输出结果后，不清空历史记录的情况下，继续以同一个问题描述进行提问，其回答内容是否与第一次结果相同；	√	7.5	7.5
	AI模型在第一次输出结果后，在清空历史记录的情况下，继续以同一个问题描述进行提问，其回答内容是否与第一次结果相同；	×	7.5	0
稳定性测试	AI模型在连续5次以上回答相同问题时，其回答是否保持一致，没有出现明显的波动或变化；	√	10	10
理解力测试	AI模型在面对不同表述方式但问题本质相同的情况下，能否给出正确的答案；（基于提供的备选，语料库具备的）	×	12.5	0
	AI模型在面对不同表述方式但问题本质相同的情况下，能否给出正确的答案；（基于自创，语料库之外的）	×	12.5	0
推理力测试	AI模型在面对丰富的业务场景描述，且描述内容与语料库重合度高，能否给出正确答案；	√	15	15
	AI模型在面对丰富的业务场景描述，且描述内容与语料库重合度低，能否给出正确答案；	×	25	0
合计			100	37.5
AI模型体验总结				
请根据AI模型的体验结果以及体验分析情况，简单描述AI模型的输出内容，并指出AI模型是否按照预期输出了所有必要的信息。				

图 20



图 21

任务三 低学习率模型体验

一、任务介绍

在上一任务中修改训练轮次进行微调体验，现需要修改其他参数的值，体验不同参数对模型效果的影响。

二、微调模型并测试

(一) 微调模型

1. 参考上一任务中的参数，将模型微调参数中的学习率修改为 0.00003，其余参数不变。

2. 微调过程中使用的语料文件为转换后的 json 文件，请将文件上传至指定位置。

3. 微调结束后，请选取此次微调后的模型进行对话，以对话形式进行测试，评估模型微调效果。

请结合任务描述中的内容完成参数设定，设定后的参数效果如下所示：

注意，此处参数的变动是基于上一任务，不是基于默认参数。

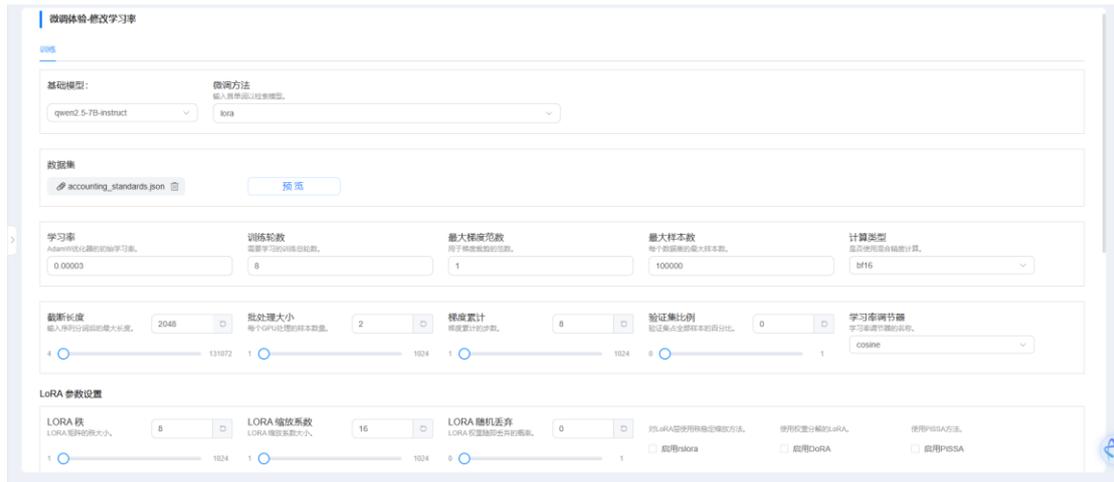


图 22

微调后的损失函数图如下所示：

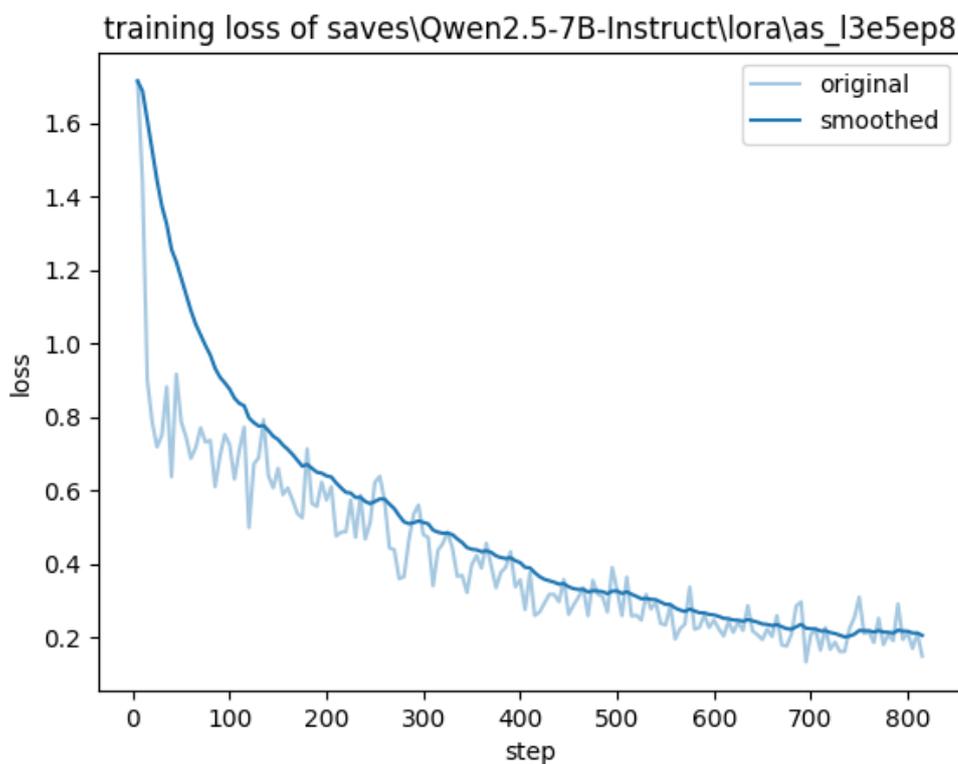


图 23

(二) 测试模型

1. 步骤 1: 测试说明

任务描述

请根据以下体验要点提示，结合步骤 2 所选的业务场景，依次向机器人发出提问并查看机器人的回答。对比分析 AI 回答与手写的会计分录，根据体验结果完成模型对应的体验分析报告。

测试体验

1) 测试要点

- ①向 AI 模型输入场景描述的内容，查看 AI 的回答内容；
- ②向 AI 模型输入场景描述的内容，并于提示语处输入相关提示语，查看 AI 的回答内容；
- ③向 AI 模型输入场景描述的内容，得到回答后，在不清空历史记录的情况下，就同一个相同问题连续进行发起 3 次问答，查看 AI 的回答内容；

④向 AI 模型输入场景描述的内容,得到回答后,在清空历史记录的情况下,就同一个相同问题连续进行发起 3 次问答,查看 AI 的回答内容;

⑤向 AI 模型连续 5 次以上输入相同的场景描述,查看你 AI 的回答内容是否保持一致,没有出现明显的波动或变化;

⑥向 AI 模型输入场景描述的内容,得到答复后,再次输入同义描述的内容,查看 AI 的回答内容;

⑦向 AI 模型输入场景描述的内容后,再次输入自创的同义描述,查看 AI 的回答内容(自创内容不与同义描述相同)

⑧请给予所选的业务场景描述,丰富场景描述(要求和原来的场景描述重合度高,或者包含原来的场景描述内容)向 AI 模型输入丰富后的场景描述,查看 AI 的回答内容;

⑨请给予所选的业务场景描述,丰富场景描述(要求和原来的场景描述重合度低)向 AI 模型输入丰富后的场景描述,查看 AI 的回答内容;

2) 测试分析

请根据测试体验的内容,结合提供的体验分析报告模板,完成模型对应的体验分析报告。

2. 步骤 2:进行测试

请使用上一环节生成的模型完成对话预测。提示:最大生成长度可设置为 1024, top-p 为 0.7, 温度系数为 0.95。

【进行对话测试前,请先设置右侧的参数,以免回复内容不完整影响测试结论】

测试对话过程与效果如下所示:



图 24

三、AI 体验测试报告

请下载参考模板，完成后以截图形式上传

点击链接即可下载，下载后的文件打开可看到两个 sheet。

基本信息记录表如下所示：

基本信息记录				
AI模型名称				
体验日期				
使用业务场景				
AI模型体验分析				
内容名称	测试内容	结果	分数设置	得分
		满足测试内容的打√/ 不满足测试内容的打×		满足测试内容满分； 不满足测试内容零分
准确性测试	AI模型能否输出相关内容；		5	
	AI模型输出内容是否正确；		5	
耦合性测试	AI模型在第一次输出结果后，不清空历史记录的情况下，继续以同一个问题描述进行提问，其回答内容是否与第一次结果相同；		7.5	
	AI模型在第一次输出结果后，在清空历史记录的情况下，继续以同一个问题描述进行提问，其回答内容是否与第二次结果相同；		7.5	
稳定性测试	AI模型在连续5次以上回答相同问题时，其回答是否保持一致，没有出现明显的波动或变化；		10	
理解力测试	AI模型在面对不同表述方式但问题本质相同的情况下，能否给出正确的答案；（基于提供的备选，语料库具备的）		12.5	
	AI模型在面对不同表述方式但问题本质相同的情况下，能否给出正确的答案；（基于自创，语料库之外的）		12.5	
推理力测试	AI模型在面对丰富的业务场景描述，且描述内容与语料库重合度高，能否给出正确答案；		15	
	AI模型在面对丰富的业务场景描述，且描述内容与语料库重合度低，能否给出正确答案；		25	
合计			100	0
AI模型体验总结				
请根据AI模型的体验结果以及体验分析情况，简单描述AI模型的输出内容，并指出AI模型是否按照预期输出了所有必要的信息。				

图 25

测试结果汇总表如下所示：



图 26

答案仅供参考，请以实际测试体验为准。

基本信息记录				
AI模型名称				
体验日期				
使用业务场景				
AI模型体验分析				
内容名称	测试内容	结果	分数设置	得分
		满足测试内容的打√/ 不满足测试内容的打×		满足测试内容满分； 不满足测试内容零分
准确性测试	AI模型能否输出内容为会计分录；	√	5	5
	AI模型能否正确输出的会计分录；	×	5	0
耦合性测试	AI模型在第一次输出结果后，不清空历史记录的情况下，继续以同一个问题描述进行提问，其回答内容是否与第一次结果相同；	√	7.5	7.5
	AI模型在第一次输出结果后，在清空历史记录的情况下，继续以同一个问题描述进行提问，其回答内容是否与第一次结果相同；	×	7.5	0
稳定性测试	AI模型在连续5次以上回答相同问题时，其回答是否保持一致，没有出现明显的波动或变化；	√	10	10
理解力测试	AI模型在面对不同表述方式但问题本质相同的情况下，能否给出正确的答案；（基于提供的备选，语料库具备的）	×	12.5	0
	AI模型在面对不同表述方式但问题本质相同的情况下，能否给出正确的答案；（基于自创，语料库之外的）	×	12.5	0
推理力测试	AI模型在面对丰富的业务场景描述，且描述内容与语料库重合度高，能否给出正确答案；	√	15	15
	AI模型在面对丰富的业务场景描述，且描述内容与语料库重合度低，能否给出正确答案；	×	25	0
合计			100	37.5
AI模型体验总结				
请根据AI模型的体验结果以及体验分析情况，简单描述AI模型的输出内容，并指出AI模型是否按照预期输出了所有必要的信息。				

图 27



图 28

任务四 RAG 技术体验

一、任务描述

请使用企业会计准则问答内容作为此次审计行业问答的基础文件，完成后续实训。

二、BERT 模型评估语义相似度

(一) 审计准则问答文档

文档已内置在平台中，可滑动浏览。下图为内置文档效果图，请以平台中的实际内容为准。



图 29

（二）BERT 模型评估语义相似度

请自选语句及其他语句（同义或非同义均可），体验相似度评估过程，语句可从企业会计准则问答文档中摘取，也可自行准备。

【此处的语句内容可在平台内置的审计准则问答文档中任意选取，无指定要求】选取语句后，在参数处逐一粘贴，并点击运行。



图 30

以上图为例，运行后的结果如下所示：

执行状态: ● 运行结束

执行结果:

```
输出:
自选语句: 固定资产在以下情况下应当予以终止确认: (一) 该固定资产处于处置状态。 (二) 该固定资产
预期通过使用或处置不能产生经济利益。
最相似语句: 固定资产同时满足以下两个条件才能予以确认: (一) 与该固定资产有关的经济利
益很可能流入企业; (二) 该固定资产的成本能够可靠地计量
最高相似度得分: 0.7664

运行结束
```

图 31

技术需求转化答案:

关键词	参数
自选语句内容	自定义
待测平语句 1	自定义
待测平语句 2	自定义
待测平语句 3	自定义
待测平语句 4	自定义

二、提取文本并进行预处理

从企业会计准则问答文档中提取文本, 使用滑动窗口算法和向量化构建问答知识库, 完成数据预处理。

提示: 最大序列长度可设为 1024, 滑窗的窗口大小可设为 512。

结合任务要求补全代码后, 点击运行, 运行效果如下所示。

执行状态: ● 运行结束

执行结果:

```
输出:
Data processing complete.

运行结束

其他:
(1) 文件: accounting_standards.json 
```

图 32

下载 json 文件后，可用记事本打开，也可用 vs code 打开，在 vs code 中的打开效果如下所示：

```

1 [{"id": 1, "text": "sample line"}]
2 [{"id": 2, "text": "固定资产在什么情况下应当予以终止确认？\n固定资产在以下情况下应当予以终止确认：（一）该固定资产处于处置状态。（二）该固定资产预期通过使用或处置不能产生经济利益。"}]
3 [{"id": 3, "text": "（二）该固定资产预期通过使用或处置不能产生经济利益。"}]
4 [{"id": 4, "text": "固定资产在什么条件下才能予以确认？\n固定资产同时满足以下两个条件才能予以确认：（一）与该固定资产有关的经济利益很可能流入企业；（二）该固定资产的成本能够可靠地计量。"}]
5 [{"id": 5, "text": "关键管理人员的定义是什么？\n关键管理人员是指有权力和负责计划、指挥和控制企业活动的人员。"}]
6 [{"id": 6, "text": "关联方交易的类型通常包括哪些？\n关联方交易的类型通常包括转移资源、劳务或义务的行为。"}]
7 [{"id": 7, "text": "关联方交易披露的内容通常包括哪些？\n关联方交易披露的内容包括交易的金额、未结算项目的金额、条款和条件，以及有关提供或取得担保的信息。"}]
8 [{"id": 8, "text": "固定资产在什么情况下应当予以终止确认：（一）该固定资产处于处置状态。（二）该固定资产预期通过使用或处置不能产生经济利益。"}]

```

图 33

技术需求转化参考答案

关键词	参数
最长序列长度	1024
待测平语句 1	512

三、创建知识库

使用上一环节生成的 json 文件构建知识库。

结合任务要求，在参数补全位置填写上一任务中生成的文件名，运行代码，结果如下所示。

知识库已保存到 as_knowledge.pt

图 34

技术需求转化答案：

关键词	参数
上一任务生成的文件名	accounting_standards

说明：此处为减轻服务器压力，创建知识库的部分选用模拟运行，故无法学生端下载 pt 文件，教师若想查看 pt 文件，可在教师端后台进行查看。

四、使用 RAG 技术

请参考提供的企业会计准则 pdf，在下方任意提问，初次运行时间较长，请耐心等待。

提问内容可以自定义选取，例如：固定资产在什么情况下应当予以终止确认？

执行后的效果如下所示：

执行状态： 运行结束

执行结果：

```
输出：
知识库已加载
----固定资产在以下情况下应当予以终止确认：（一）该固定资产处于处置状态----

运行结束

其他：
(1) 文件: accounting_standards.json 
```

图 35

技术需求转化答案

关键词	参数
请输入自选语句进行提问	自定义

任务五 识别舞弊企业

一、财务数据采集

请使用数据中心内的网址，结合医药采集需求.xlsx，完成财务数据采集（表格已内置在平台中，无需额外处理）。数据采集结果文件名可自定义。

数据中心内的网址打开如下所示：



图 36

医药采集需求.xlsx 中含有两个 sheet，其中“未舞弊”的表格内容如下所示：

股票代码	证券名称	所属行业	数据来源
000028.SZ	国药一致	医药行业	同花顺医药商业板块，2024年1月18日涨幅前 20
000078.SZ	海王生物	医药行业	
000150.SZ	*ST 宜康	医药行业	
000153.SZ	丰原药业	医药行业	
000403.SZ	派林生物	医药行业	
000411.SZ	英特集团	医药行业	
000423.SZ	东阿阿胶	医药行业	

“舞弊”的表格内容如下所示

股票代码	证券名称	所属行业	造假年份
300267.SZ	尔康制药	医药行业	2015, 2016
600196.SH	复星医药	医药行业	2018
600518.SH	ST 康美	医药行业	2016, 2017, 2018
300078.SZ	思创医惠	医药行业	2019, 2020
002433.SZ	ST 太安	医药行业	2018, 2019, 2020, 2021
002370.SZ	亚太药业	医药行业	2016, 2017, 2018

结合任务要求补全代码后，执行代码可查看结果如下

执行状态： ● 运行结束

执行结果：

其他：

(1) 文件：数据采集结果.xlsx 下载

图 37

表格内容如下

A	B	C	D	E	F	G	H	I	J	K	L	M	N	O
行业	股票	项目\年度	成长能力指标	净利润(元)	净利润同比增长率	扣非净利润(元)	扣非净利润同比增长率	营业收入(元)	营业收入同比增长率	每股指标	基本每股收益(元)	每股净资产(元)	每股资本公积金(元)	每股未分配利润(元)
2	医药行业	医药-沪深300	2023	15.99亿	7.57%	15.33亿	7.58%	784.77亿	2.77%	2.8700	31.29	7.66		22.12
3	医药行业	医药-沪深300	2022	14.87亿	11.24%	14.25亿	10.77%	734.43亿	7.44%	2.6700	27.67	10.18		25.97
4	医药行业	医药-沪深300	2021	13.38亿	-4.67%	12.89亿	-5.99%	682.58亿	14.60%	2.1200	24.89	10.21		23.10
5	医药行业	医药-沪深300	2020	14.02亿	10.57%	13.7亿	11.65%	596.49亿	13.00%	2.2700	32.58	10.21		20.78
6	医药行业	医药-沪深300	2019	12.66亿	4.72%	12.27亿	4.32%	527.87亿	20.69%	2.9600	29.91	10.19		18.11
7	医药行业	医药-沪深300	2018	12.11亿	14.48%	11.76亿	13.97%	451.22亿	4.91%	2.8300	27.14	10.09		15.85
8	医药行业	医药-沪深300	2017	10.58亿	-10.85%	10.35亿	-24.10%	412.63亿	0.04%	2.4700	21.95	7.43		13.02

图 38

技术需求转换答案：

关键词	参数
采集网址	以实际为准
数据采集结果文件名	自定义

二、财务数据预处理

(一) 数据集划分

请在上一步骤的基础上完成财务数据预处理，划分出训练集与测试集。结合任务要求补全代码后，执行代码可查看结果如下

执行状态： ● 运行结束

执行结果：

其他：

(1) 文件：数据采集结果.xlsx 下载

(2) 文件：训练集.csv 下载

(3) 文件：测试集.csv 下载

图 39

其中，训练集结果如下：

#	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	净利润同时扣非净利润 营业收入 基本每股收益 每股净资产 每股资本公积 每股未分配利润 每股经营现金流量 销售净利润 销售毛利率 净资产收益率 营业周期 (存货周转率 存货周转率 应收账款周转率) 流动比率 速动比率 保守速动比率 产成比率 资产负债率类别																						
2	5213	-0.2721	-1.6097	-0.5158	0.035	1.12	0.17	-0.07	-0.07	0.6975	0.5781000	0.0312000	0.0312000	1.62	0.02	1.01	23.63	4.92	0.15	0.15	0.3	0.2286	0
3	2679	12.2925	15.00819	2.0106	0.2702	1.84	0.13	0.64	0.13	0.322	0.7854000	0.1585	0.1472	591.61	0.79	457.49	134.12	2.14	1.53	1.53	0.4	0.2846	0
4	4584	0.1823000	0.1714	0.1153	0.17	2.64	0.7	0.79	0.07	0.0699	0.4698	0.0675	0.0659	212.04	3.13	115.05	96.98	2.94	2.23	2.23	0.3	0.2287	0
5	6212	-0.1125	-0.0876	-0.0892	0.26	2.31	1.73	2.7	0.34	0.3182995	0.8017	0.1404	0.1052995	430.27	1.04	344.66	85.61	3.29	2.91	2.91	0.22	0.1788	0
6	1335	0.1637	0.304	0.1232	0.21	2.27	0.8	0.26	0.21	0.0471	0.1151	0.0944	0.092	34.06	16.55	21.75	12.32	2.8	2.42	2.42	0.32	0.2304	0
7	1355	-0.1756	0.146	0.0272000	0.16	3.39	1.42	0.85	0.19	0.024	0.1305	0.0475	0.047	141.9	4.08	88.27	53.63	1.26	0.68	0.68	0.85	0.4596	0

图 40

测试集结果如下：

#	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
1	净利润同时扣非净利润 营业收入 基本每股收益 每股净资产 每股资本公积 每股未分配利润 每股经营现金流量 销售净利润 销售毛利率 净资产收益率 营业周期 (存货周转率 存货周转率 应收账款周转率) 流动比率 速动比率 保守速动比率 产成比率 资产负债率类别																						
2	423	-0.1302	-0.7635	-0.4386	0.05	2.26	0.54	0.43	0.1	0.1251	0.4365	0.0214000	0.0217	1279.04	0.28	1264.04	15	1.83	0.47	0.47	0.98	0.4643	0
3	6527	-0.014	0.006	0.5881000	-1.91	3.16	9.26	-7.08	-1.54	-1.6075	0.9135	-0.4635	-0.603	1128.18	0.35	1038.06	90.12	1.77	1.57	1.57	1.16	0.528	0
4	6052	0.2365	0.246	0.1541000	0.33	3.11	0.94	1.04	0.71	0.1244	0.3486	0.1149	0.1071000	138.71	4.11	87.7	51.01	2.38	1.87	1.87	0.45	0.2912000	0
5	3727	2.0238	0.1481	0.2586	0.62	2.92	1.55	0.32	0.61	0.3774	0.8719	0.2993	0.2030995	244.58	1.55	232.75	11.83	5.37	4.71	4.71	0.18	0.1541000	0
6	1407	-0.5315	-0.5623	-0.0017	0.14	3.36	1.2	1.1	-0.6	0.1313	0.6484000	0.0412995	0.0404	4451.5	0.09	4225.35	226.14	1.78	0.26	0.26	1.3	0.5654	0
7	1840	0.325	0.2923	0.2060000	0.29	1.88	0.1	0.66	0.29	0.2191	0.6206	0.1674	0.1553995	419.71	1.06	339.33	80.38	2.59	1.95	1.95	0.44	0.3021000	0

图 41

技术需求转换答案

关键词	参数
数据采集结果文件名	自定义（与上一任务保持一致）
训练集文件名称	自定义
测试集文件名称	自定义

(二) 语料格式转换

请将划分出来的训练集转换为 json 文件格式，便于后续语料训练。

结合任务要求补全代码后，执行代码可查看结果如下

执行状态：● 运行结束

执行结果：

输出：
 检测到文件编码: GB2312
 已生成 4332 条训练数据，保存至 [UserFolderPath]/训练集转换后.json
 数据转换完成!

运行结束

其他：
 (1) 文件: 训练集.csv 下载
 (2) 文件: 训练集转换后.json 下载

图 42

舞弊企业数据格式转换后内容如下所示：

```

1  {
2  {
3  "instruction": "根据公司的财务指标分析该公司是否存在财务报告舞弊行为",
4  "input": "公司代码: 2141, 公司简称: 质丰控股, 年份: 2020, 行业: 农业综合, 资产负债率: 18.0304%, 应收账款周转率: 5.3159次, 存货周转率: 7.3243次, 流动资产周转率: 2.1206次, 销售净利率: 0.0437%",
5  "output": "根据对质丰控股2020年财务指标的分析: \n盈利能力方面: 销售净利率较低(0.0437%), 盈利能力较弱; 总资产净利率较低(1.6002%), 资产盈利能力不足; 净资产收益率为负(-48.89%), 股东投资回
6  },
7  {
8  "instruction": "根据公司的财务指标分析该公司是否存在财务报告舞弊行为",
9  "input": "公司代码: 300109, 公司简称: 翰宇药业, 年份: 2020, 行业: 化学制药, 资产负债率: 52.7364%, 应收账款周转率: 0.5954次, 存货周转率: 0.8017次, 流动资产周转率: 0.25次, 销售净利率: -11.11%",
10 "output": "根据对翰宇药业2020年财务指标的分析: \n盈利能力方面: 销售净利率为负(-11.11%), 表明公司可能处于亏损状态; 总资产净利率为负(-16.2799%), 资产使用效率低下; 净资产收益率为负(-31.11%)
11 },
12 {
13 "instruction": "根据公司的财务指标分析该公司是否存在财务报告舞弊行为",
14 "input": "公司代码: 603959, 公司简称: 百利科技, 年份: 2020, 行业: 建筑工程, 资产负债率: 83.2769%, 应收账款周转率: 2.0461次, 存货周转率: 3.1064次, 流动资产周转率: 0.7205次, 销售净利率: -23.1631%",
15 "output": "根据对百利科技2020年财务指标的分析: \n盈利能力方面: 销售净利率为负(-23.1631%), 表明公司可能处于亏损状态; 总资产净利率为负(-23.1631%), 资产使用效率低下; 净资产收益率(5.04%)偏低
16 },
--

```

图 46

技术需求转化答案

关键词	参数
舞弊企业财务数据转换后的 json 名称	自定义

(四) 语料合并

请补全代码，合并训练集与舞弊企业财务数据转换后的 json 文件，形成可供模型微调的语料。

结合任务要求补全代码后，执行代码可查看结果如下

执行状态: ● 运行结束

执行结果:

输出:
 成功合并两个文件:
 文件1: [UserFolderPath]/训练集转换后.json - 4332条记录
 文件2: [UserFolderPath]/舞弊企业转换后.json - 324条记录
 合并后: [UserFolderPath]/financial_fraud.json - 4656条记录

运行结束

其他:

(1) 文件: 训练集转换后.json 下载

(2) 文件: 舞弊企业转换后.json 下载

(3) 文件: financial_fraud.json 下载

图 47

合并后的文件内容如下

```

1 {
2   {
3     "instruction": "请分析以下公司的财务指标,判断是否存在财务舞弊风险",
4     "input": "公司财务指标数据: \n\nNamed: 0: 5213.0\n净利润同比增长率: -0.2721\n扣非净利润同比增长率: -1.6097\n营业收入同比增长率: -0.5150\n基本每股收益: 0.835\n每股净资产: 1.12\n",
5     "output": "经分析,该公司财务指标表现正常,未发现明显的舞弊迹象。各项指标配比合理,数据一致性较好。"
6   },
7   {
8     "instruction": "根据财务指标特征,识别潜在的财务舞弊迹象",
9     "input": "公司财务指标数据: \n\nNamed: 0: 2679.0\n净利润同比增长率: 12.2925\n扣非净利润同比增长率: 15.0082\n营业收入同比增长率: 2.0106\n基本每股收益: 0.2702\n每股净资产: 1.84\n",
10    "output": "财务数据分析显示公司经营状况正常,无明显舞弊风险特征。"
11  },
12  {
13    "instruction": "基于财务数据分析,评估该公司的财务报告真实性",
14    "input": "公司财务指标数据: \n\nNamed: 0: 4584.0\n净利润同比增长率: 0.1823\n扣非净利润同比增长率: 0.1714\n营业收入同比增长率: 0.1153\n基本每股收益: 0.17\n每股净资产: 2.64\n",
15    "output": "财务数据分析显示公司经营状况正常,无明显舞弊风险特征。"
16  },
17 }

```

图 48

技术需求转化

关键词	参数
训练集转换后的 json 文件名	自定义（与上一任务保持一致）
舞弊企业财务数据转换后的 json 名称	自定义（与上一任务保持一致）

三、Deepseek 模型微调并识别舞弊企业

（一）Deepseek 模型微调

请使用 DeepSeek-LLM-7B-Chat 作为基础模型,选用 lora 微调方法,在上传合并所得的 json 文件后,使用以下参数组合进行微调:学习率为 0.0005,训练轮数为 1,LoRA 秩为 16,LoRA 缩放系数为 32（其余参数使用默认值）。

请结合任务描述中的内容完成参数设定,也可点击预览查看语料内容,如下所示:

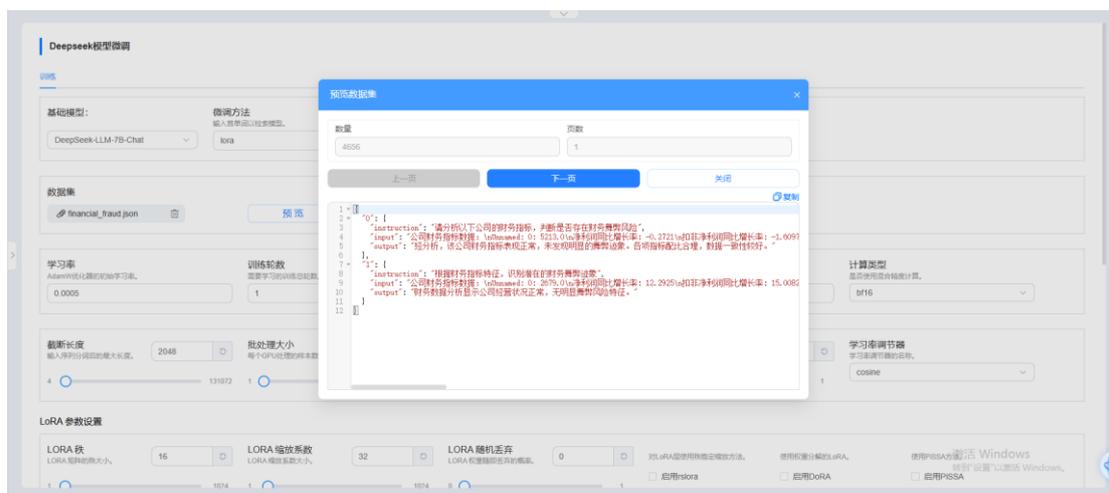


图 49

微调后的损失函数图如下所示：

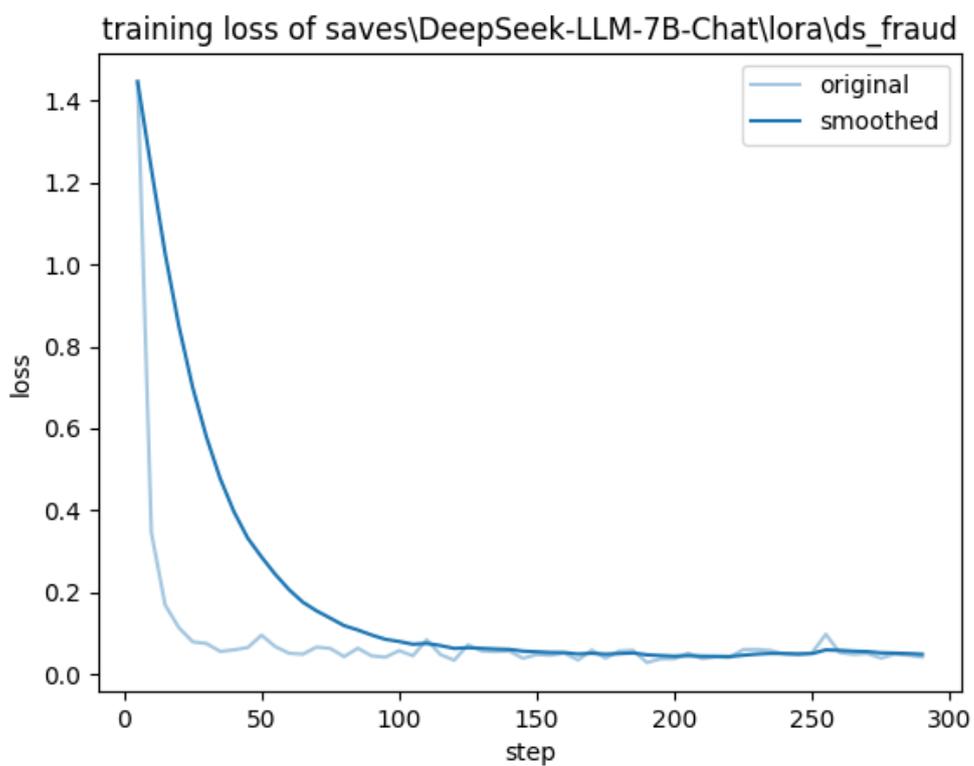


图 50

（二）微调后模型对话

请应用上一步训练好的 `ds_base_fraud` 模型，对企业财务数据进行舞弊鉴别，并评估该模型判断的准确性。提示：最大生成长度可设置为 1024，top-p 为 0.7，温度系数为 0.95。

【进行对话测试前，请先设置右侧的参数，以免回复内容不完整影响测试结论】

测试对话过程与效果如下所示：



图 51

四、Qwen 模型微调并识别舞弊企业

(一) Qwen 模型微调

请使用 qwen2.5-7B-instruct 作为基础模型，选用 lora 微调方法，在上传合并所得的 json 文件后，使用以下参数组合进行微调：学习率为 0.0005，训练轮数为 1，LoRA 秩为 8，LoRA 缩放系数为 16。（其余参数使用默认值）

请结合任务描述中的内容完成参数设定，也可点击预览查看语料内容，如下所示

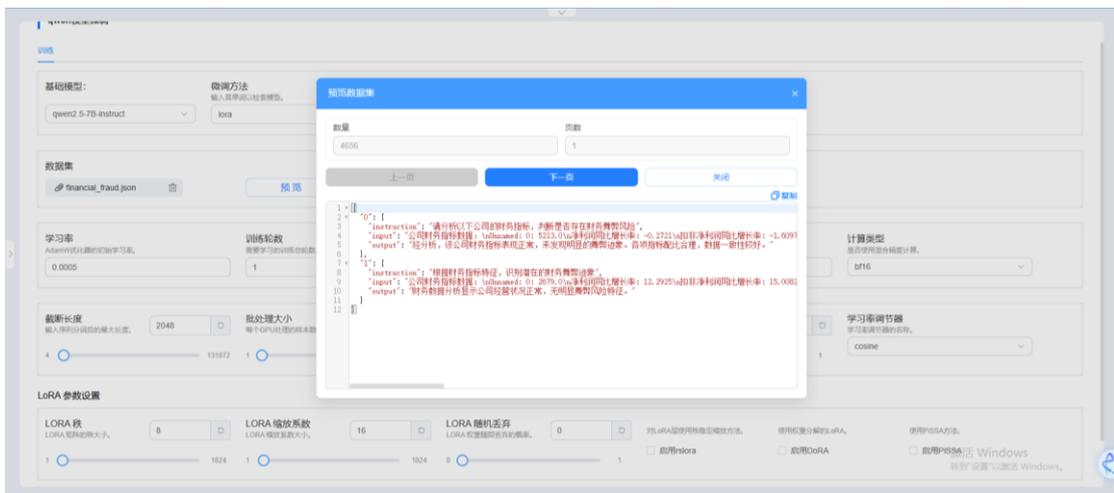


图 52

微调后的损失函数图如下所示：

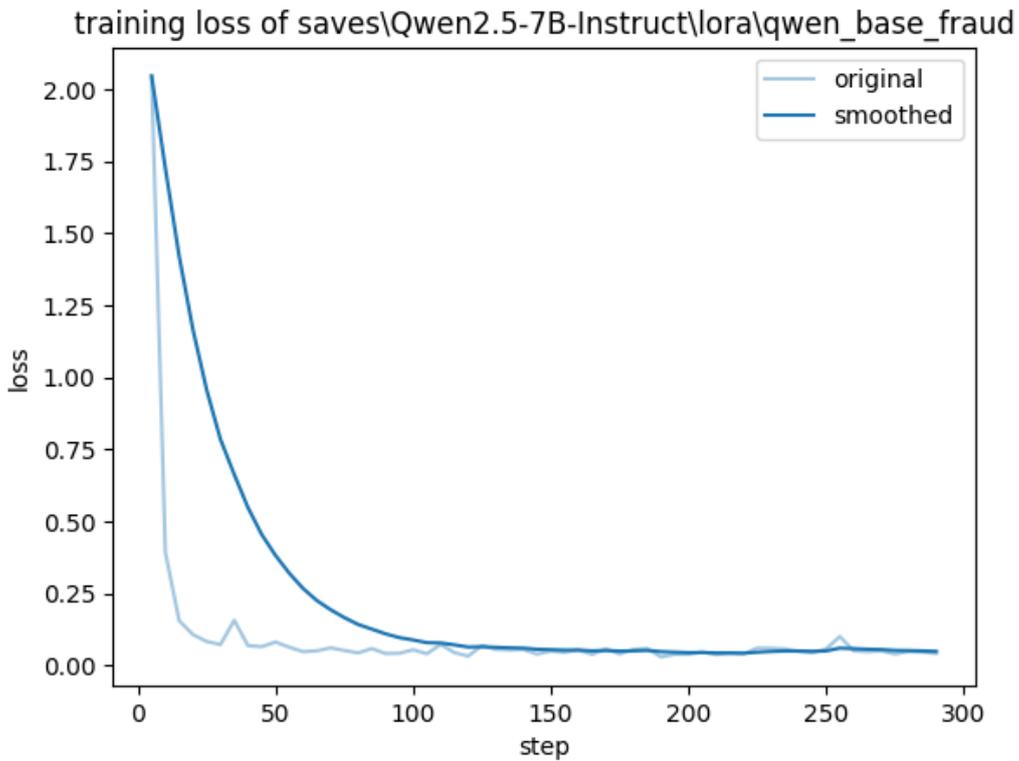


图 53

(二) 微调后模型对话

请应用上一步训练好的 `qwen_base_fraud` 模型，对企业财务数据进行舞弊鉴别，并评估该模型判断的准确性。提示：最大生成长度可设置为 1024，`top-p` 为 0.7，温度系数为 0.95。

【进行对话测试前，请先设置右侧的参数，以免回复内容不完整影响测试结论】

测试对话过程与效果如下所示：



图 54

五、Llama 模型微调并识别舞弊企业

(一) Llama 模型微调

请使用 Llama-3-8B-instruct 作为基础模型，选用 lora 微调方法，在上传上一步骤获得的训练集 json 文件后，使用以下参数组合进行微调：学习率为 0.0005，训练轮数为 1，LoRA 秩为 16，LoRA 缩放系数为 32（其余参数使用默认值）。

请结合任务描述中的内容完成参数设定，也可点击预览查看语料内容，如下所示

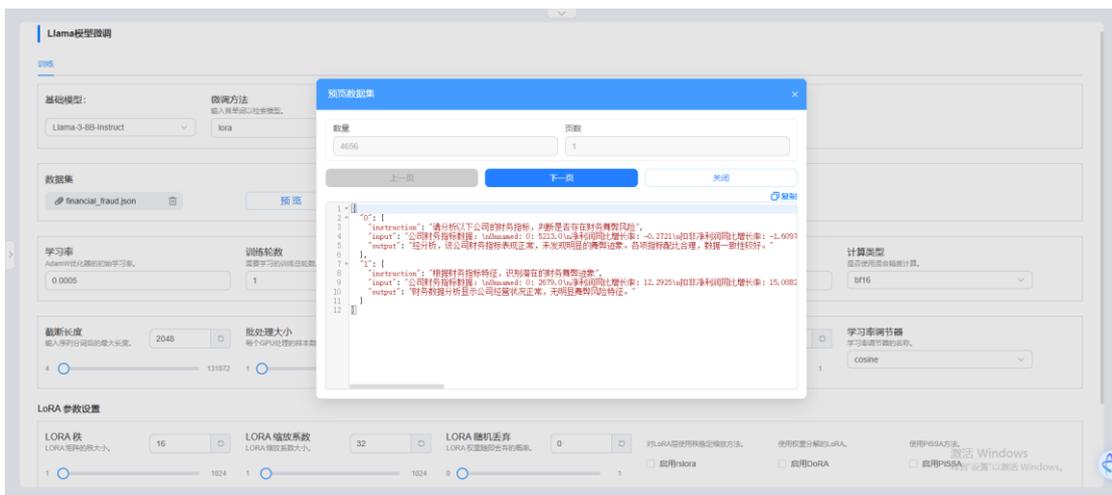


图 55

微调后的损失函数图如下所示：

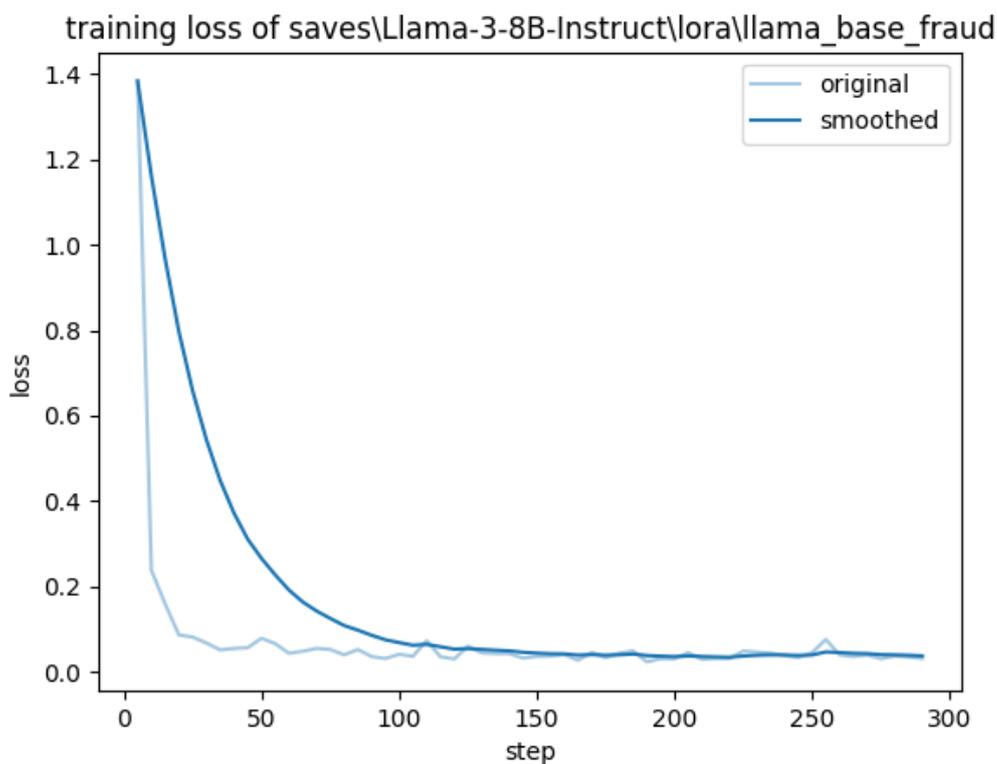


图 56

(二) 微调后模型对话

请应用上一步训练好的 llama_base_fraud 模型，对企业财务数据进行舞弊鉴别，并评估该模型判断的准确性。提示：最大生成长度可设置为 1024，top-p 为 0.7，温度系数为 0.95。

【进行对话测试前，请先设置右侧的参数，以免回复内容不完整影响测试结论】

测试对话过程与效果如下所示：



图 57

六、AI 体验测试报告

请下载参考模板，完成后以截图形式上传

点击链接即可下载，下载后的文件打开可看到两个 sheet。

基本信息记录表如下所示：

基本信息记录				
AI模型名称				
体验日期				
使用业务场景				
AI模型体验分析				
内容名称	测试内容	结果	分数设置	得分
		满足测试内容的打√/ 不满足测试内容的打×		满足测试内容满分； 不满足测试内容零分
准确性测试	AI模型能否输出相关内容；		5	
	AI模型输出内容是否正确；		5	
耦合性测试	AI模型在第一次输出结果后，不清空历史记录的情况下，继续以同一个问题描述进行提问，其回答内容是否与第一次结果相同；		7.5	
	AI模型在第一次输出结果后，在清空历史记录的情况下，继续以同一个问题描述进行提问，其回答内容是否与第一次结果相同；		7.5	
稳定性测试	AI模型在连续5次以上回答相同问题时，其回答是否保持一致，没有出现明显的波动或变化；		10	
理解力测试	AI模型在面对不同表述方式但问题本质相同的情况下，能否给出正确的答案；（基于提供的备选，语料库具备的）		12.5	
	AI模型在面对不同表述方式但问题本质相同的情况下，能否给出正确的答案；（基于自创，语料库之外的）		12.5	
推理力测试	AI模型在面对丰富的业务场景描述，且描述内容与语料库重合度高，能否给出正确答案；		15	
	AI模型在面对丰富的业务场景描述，且描述内容与语料库重合度低，能否给出正确答案；		25	
合计			100	0
AI模型体验总结				
请根据AI模型的体验结果以及体验分析情况，简单描述AI模型的输出内容，并指出AI模型是否按照预期输出了所有必要的信息。				

图 58

测试结果汇总表如下所示：



图 59

答案仅供参考，请以实际测试体验为准。

基本信息记录				
AI模型名称				
体验日期				
使用业务场景				
AI模型体验分析				
内容名称	测试内容	结果	分数设置	得分
		满足测试内容的打√/ 不满足测试内容的打×		满足测试内容满分； 不满足测试内容零分
准确性测试	AI模型能否输出内容为会计分录；	√	5	5
	AI模型能否正确输出的会计分录；	×	5	0
耦合性测试	AI模型在第一次输出结果后，不清空历史记录的情况下，继续以同一个问题描述进行提问，其回答内容是否与第一次结果相同；	√	7.5	7.5
	AI模型在第一次输出结果后，在清空历史记录的情况下，继续以同一个问题描述进行提问，其回答内容是否与第一次结果相同；	×	7.5	0
稳定性测试	AI模型在连续5次以上回答相同问题时，其回答是否保持一致，没有出现明显的波动或变化；	√	10	10
理解力测试	AI模型在面对不同表述方式但问题本质相同的情况下，能否给出正确的答案；（基于提供的备选，语料库具备的）	×	12.5	0
	AI模型在面对不同表述方式但问题本质相同的情况下，能否给出正确的答案；（基于自创，语料库之外的）	×	12.5	0
推理力测试	AI模型在面对丰富的业务场景描述，且描述内容与语料库重合度高，能否给出正确答案；	√	15	15
	AI模型在面对丰富的业务场景描述，且描述内容与语料库重合度低，能否给出正确答案；	×	25	0
合计			100	37.5
AI模型体验总结				
请根据AI模型的体验结果以及体验分析情况，简单描述AI模型的输出内容，并指出AI模型是否按照预期输出了所有必要的信息。				

图 60



图 61