用户数据采集作业

陈绵山

信管2301

202321054008

第一讲 课程导言与分词

- 1.学习使用在线NLPIR分词系统或微词云分词或清华大学分词演示系统（案例演示截图）；

第一讲 课程导言与分词

- 2安装python（anaconda）（编写输出“Hello World. Hello ‘你的姓名’”）；

```
(base) C:\Users\陈绵山>python
Python 3.13.5 | packaged by Anaconda, Inc. | (main, Jun 12 2025, 16:37:03) [MSC v.1929 64 bit (AMD64)] on win32
Type "help", "copyright", "credits" or "license" for more information.
>>> print("Hello World. Hello '陈绵山'")
Hello World. Hello '陈绵山'
>>>
```

第一讲 课程导言与分词

- 001

```
[73]: text = "钱学森是两弹一星功勋科学家，屠呦呦发现青蒿素，黄旭华研制核潜艇"

[74]: seg_list = [w for w in jieba.lcut(text) if w not in {'的','是',', '} and len(w)>0]

[75]: print("分词结果：", seg_list)
分词结果： ['钱学森', '两弹一星', '功勋', '科学家', '屠', '呦', '呦', '发现', '青蒿素', '黄旭华', '研制', '核潜艇']

[76]: print("拼接查看：", '/'.join(seg_list))
拼接查看： 钱学森/两弹一星/功勋/科学家/屠/呦/呦/发现/青蒿素/黄旭华/研制/核潜艇

好，本次练习结束了，恭喜你!
```

第一讲 课程导言与分词

- 002

第一讲 课程导言与分词

- 003
- 1. 词频统计能反映什么?
- 高频词可能指示核心主题：例如“管理”“企业”“体系”“智能化”“供应链”等高频词，可推测文本主题涉及企业管理、智能化转型与供应链升级。
- 揭示文本重点：重复出现的词汇通常是作者强调的概念。
- 为深入分析提供基础：词频是关键词提取、主题建模等后续分析的数据基础。
- 2.词频统计本身存在什么问题呢？去停用词没有完成?
- 统计结果中包含大量无实际意义的标点符号（如逗号、引号、换行符）和虚词（如“与”），这些停用词占据了前20位的大部分位置。
- 影响：停用词会掩盖真正有意义的关键词，降低分析效果，停用词处理不彻底

第一讲 课程导言与分词

- 003
- 3仅仅统计词频有什么不足呢？能够确定“数字化”与“安全”之间的关系吗？那如何改进呢？

不足： 缺乏语境信息：无法得知词语出现的上下文，例如“数字化”是否与“安全”在相同语境中出现。忽略词语关系：无法识别共现、因果、并列等语义关系。 无法区分多义词：例如“管理”可能指行政管理或供应链管理，词频无法区分。 重要性衡量片面：低频但关键的词（如专业术语）可能被忽略。

仅凭词频无法判断两者关系。

改进：使用TF-IDF代替纯词频，降低常见词权重，突出重要词汇。

第一讲 课程导言与分词

- 004

1，使用deepseek开展工作的感觉如何？

很快很智能

2，你觉得大语言模型的活干的怎么样？

速度快

3，还是那个问题，如果可以抽取实体，那么如何识别关系呢？你试试用大语言模型识别下关系？

提取到的实体和专业术语：

```json
{
  "理论": [
    "肿瘤免疫微环境",
    "T细胞耗竭",
    "免疫编辑理论"
  ],
  "方法": [
    "单细胞RNA测序",
    "细胞亚群聚类",
    "轨迹分析",
    "pseudotime推断",
    "细胞间通讯网络构建"
  ],
  "工具": [
    "Seurat",
    "Monocle3",
    "CellChat"
  ],
  "专业术语": [
    "TIME",
    "scRNA-seq",
    "非小细胞肺癌",
    "PD-1/PD-L1",
    "TGF-β路径",
    "免疫抑制信号通路",
    "个体化免疫治疗"
  ]
}
```

基于您提供的实体，我可以尝试识别它们之间的关系。以下是分析结果：

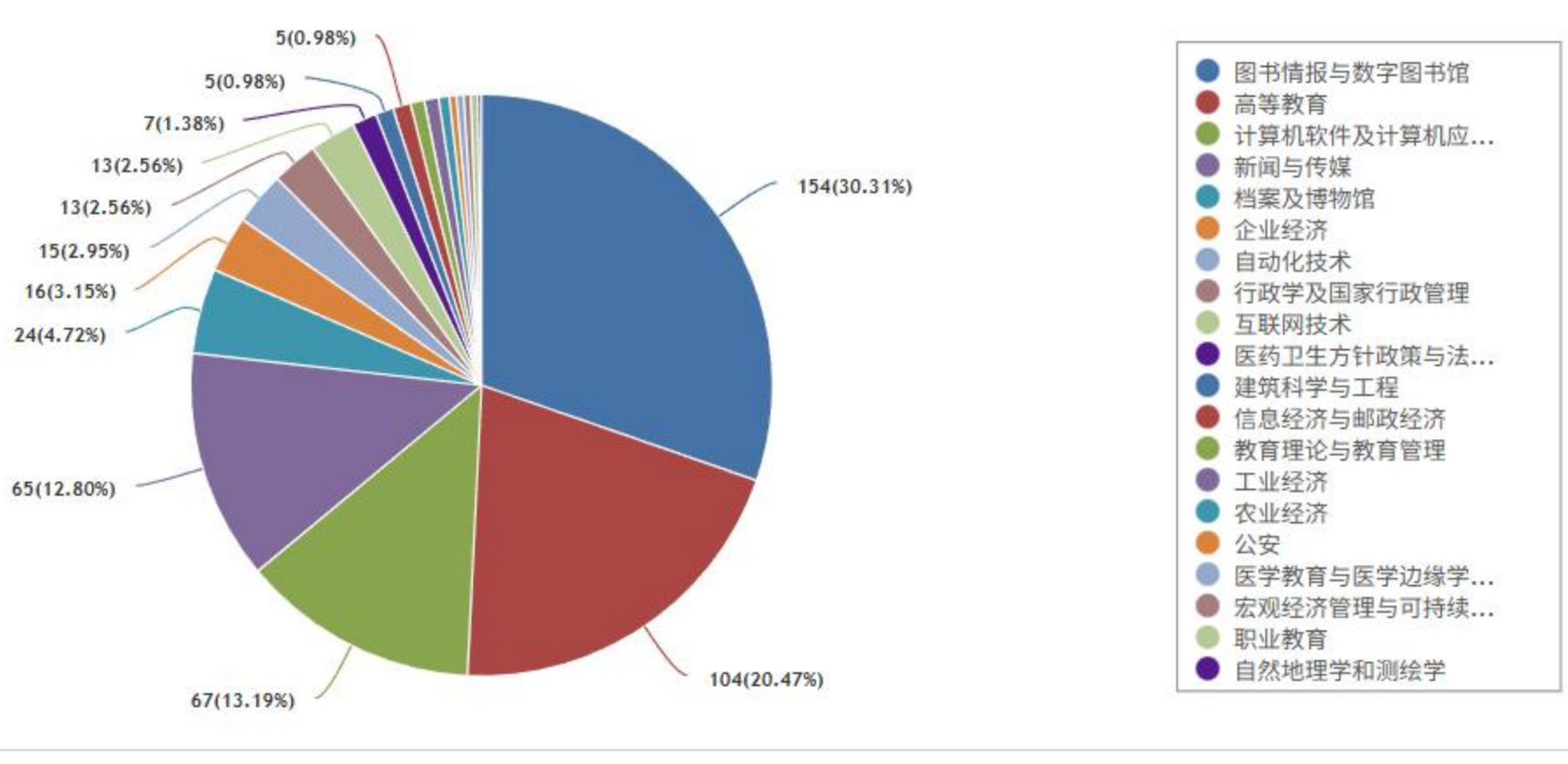
第一讲 课程导言与分词

- 阅读压缩文件中（“实体抽取论文-换成PDF”）中的其中一篇论文，并做阅读总结（1页PPT即可）。

通过阅读基于关键词的学术文本聚类集成研究，我学习到了文本聚类是一种无监督且高效的文本类别划分方法。在聚类集成方法的选择中，这篇文章选择两种聚类集成方法，第一种是以 K-means 为基础的，基于数据的聚类集成方式；第二种是以增量聚类为基础的，基于算法的聚类集成方式。在关键词抽取方法选择上，选择 TF-ISF、CSI、ECC、TextRank 这四种较经典的无监督单文本关键词抽取方法。聚类集成方法显著提升了学术文本聚类的性能；同时，当使用不同关键词抽取方法和不同关键词个数时，聚类集成方法更具稳定性。在学术文本类别划分中，需慎重选择关键词抽取方法，并在条件允许范围内抽取数量较多的关键词。另外，当关键词个数较少时，由于聚类集成方法在不同关键词个数下较稳定且在关键词个数较少时呈现较大的性能上的优势，因此应选择聚 类集成方法以减轻关键词个数不足而产生的影响。

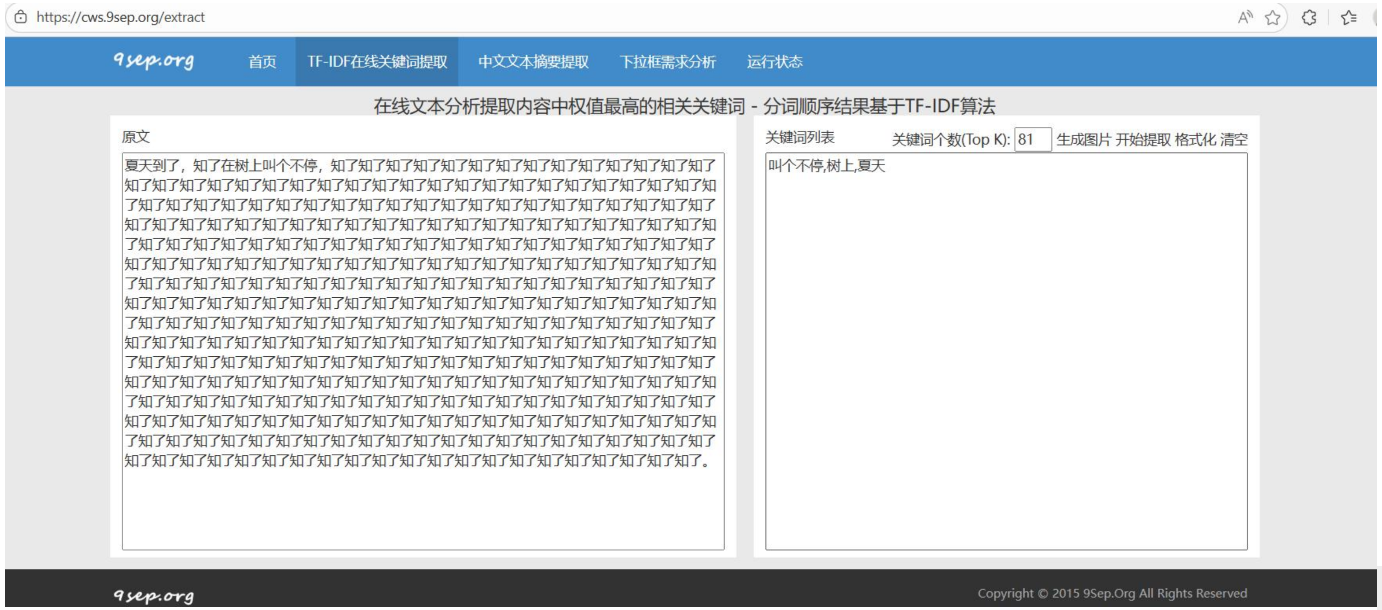
第二讲 词频统计

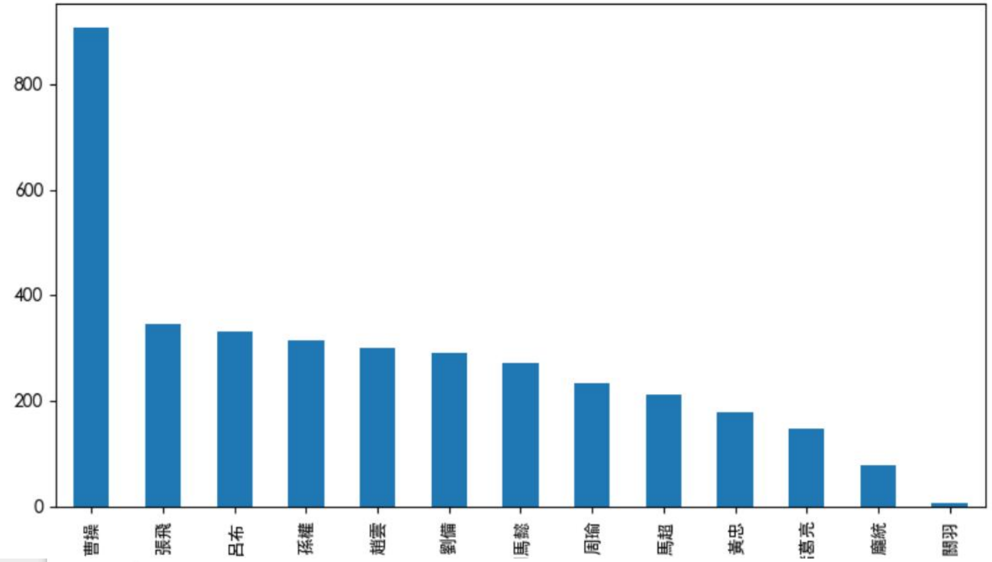
- 基于CNKI数据库统计分析2014-2024年（近10年），“信息资源管理”



第二讲 词频统计

- 2.完成ppt中的程序运行，包括全文词频统计，指定类型词频统计；



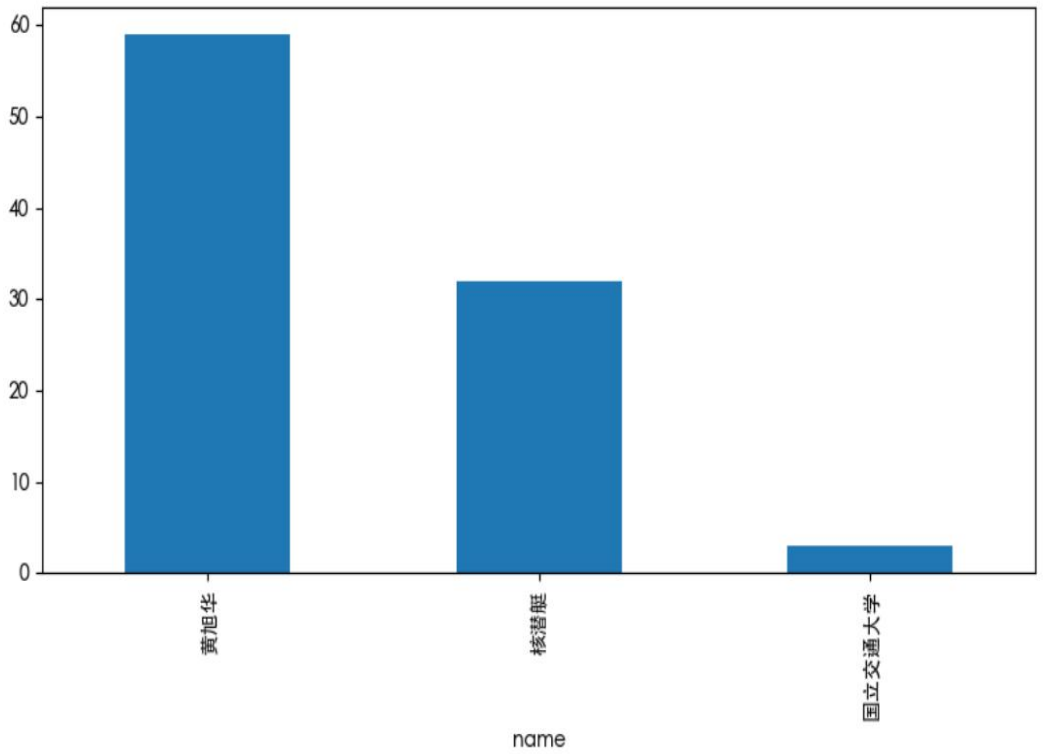


```
[17]: # 输出词频的前N个
for i in range(100):
    print(articlelist[i])

('黄旭华', 53)
('核潜艇', 32)
('采集', 29)
('学术', 22)
('资料', 21)
('工作', 17)
('成长', 15)
('小组', 14)
('院士', 13)
('进行', 13)
('专业', 13)
('技术', 12)
('研制', 12)
('我国', 12)
('工程', 11)
('访谈', 10)
```

第二讲 词频统计

- 链接功勋科学家：把ppt中的文本换成功勋科学家黄旭华院士的传记序言文本（文件夹中，科学家博物馆-黄旭华传记序言.txt），1）统计全文词频；2）统计指定词频，如“黄旭华”；

```
[18]: draw_dict(terms_dict)
```



第二讲 词频统计

- 阅读论文“2018-Wang 等 - Long live the scientists Tracking the scientific”，并做阅读总结（1页PPT即可）。

本研究利用Google Books和Google Scholar数据，通过词频分析追踪物理学家的持久科学声望。研究发现，牛顿与爱因斯坦等伟大科学家虽已离世，但其影响力持续数世纪。全球数据显示爱因斯坦的总体声望更高，但细分语言后发现“同群偏好”：牛顿在英式英语书籍中更受关注，而爱因斯坦在美式英语和德语书籍中提及更多。共现分析揭示爱因斯坦的声望主要关联相对论与量子理论，牛顿则关联万有引力与运动定律。研究提出，图书提及频率可作为传统引文指标的补充，用于衡量科学家超越学术界的长期社会影响，但需注意数据存在语言偏差、姓名歧义等局限性。

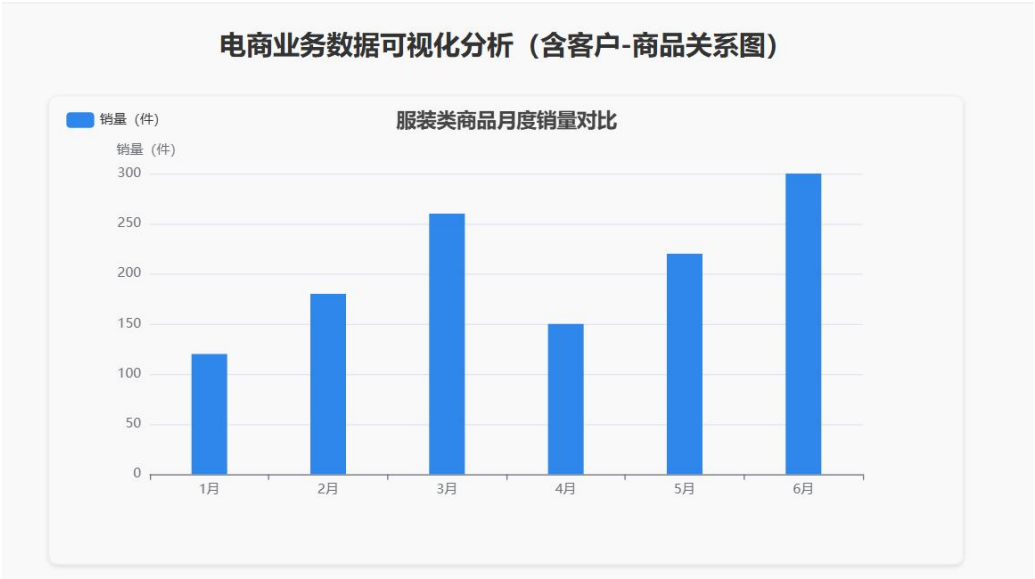
第三讲 词云与可视化

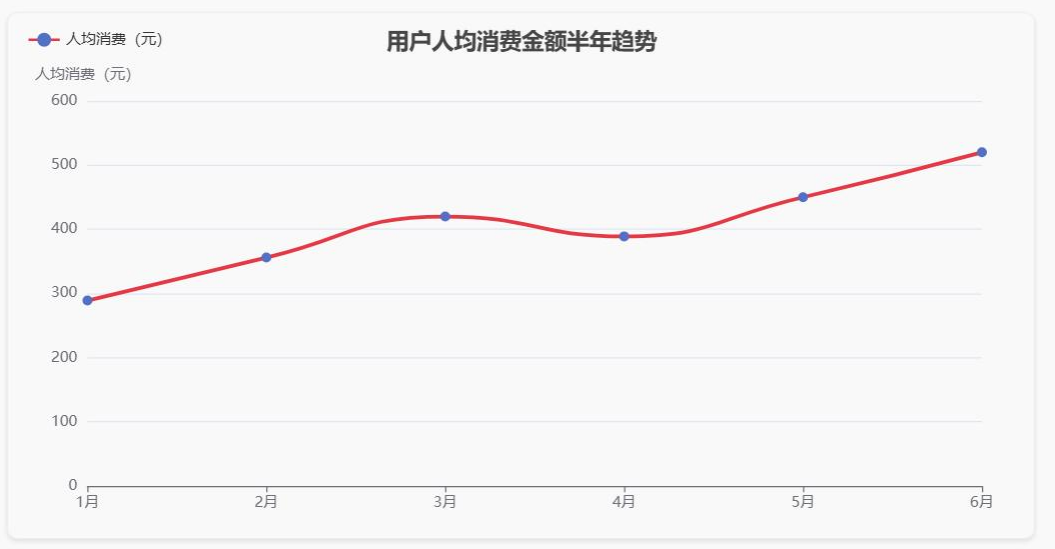
- 1.用任意一款词云工具，制作一个好看的词云（内容合理即可），并对词云图有一段话的解释。

从词云图可知，该商品这面评价多个头饱满，新鲜，总体评价高

第三讲 词云与可视化

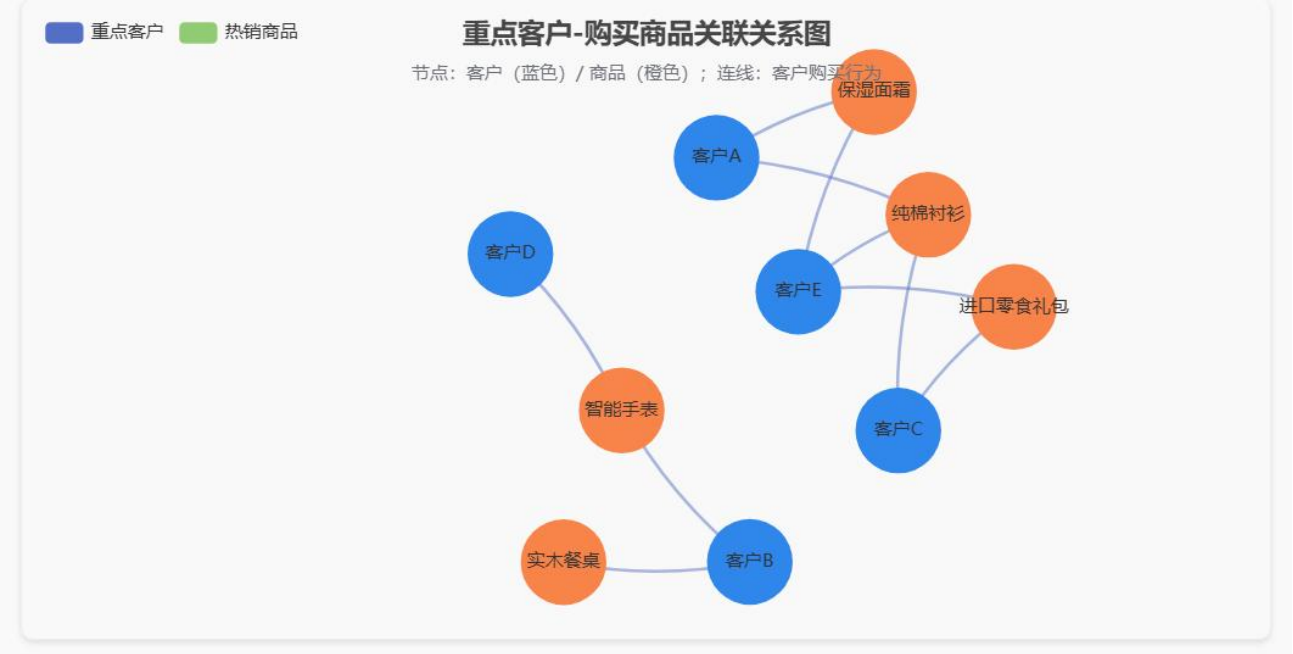
- 2使用Echarts，制作3个以上图，其中一个必须是“关系”，图的概念越明确（可解释，而不是自带的模板）越好。2. 其他图表概念明确性
- 柱状图：聚焦「服装类商品月度销量」，数据对应具体月份和销量，可用于电商月度销售复盘
- 折线图：聚焦「用户人均消费趋势」，平滑曲线清晰展示消费金额的上升 / 下降变化，可用于消费市场预判





第三讲 词云与可视化

- 3. 关系图（核心）业务解释
- 节点含义：蓝色节点是「重点客户」（A-E），橙色节点是「热销商品」（5 类核心商品），分类清晰，一眼可辨
- 连线含义：节点之间的连线代表「购买关系」，即左侧客户购买了右侧对应的商品（如客户 A 同时购买了纯棉衬衫和保湿面霜）
- 交互特性：支持鼠标缩放 / 平移（roam: true），鼠标悬浮可显示节点类型（客户 / 商品）和购买关系，力导向布局让节点自动分散，不重叠



第三讲 词云与可视化

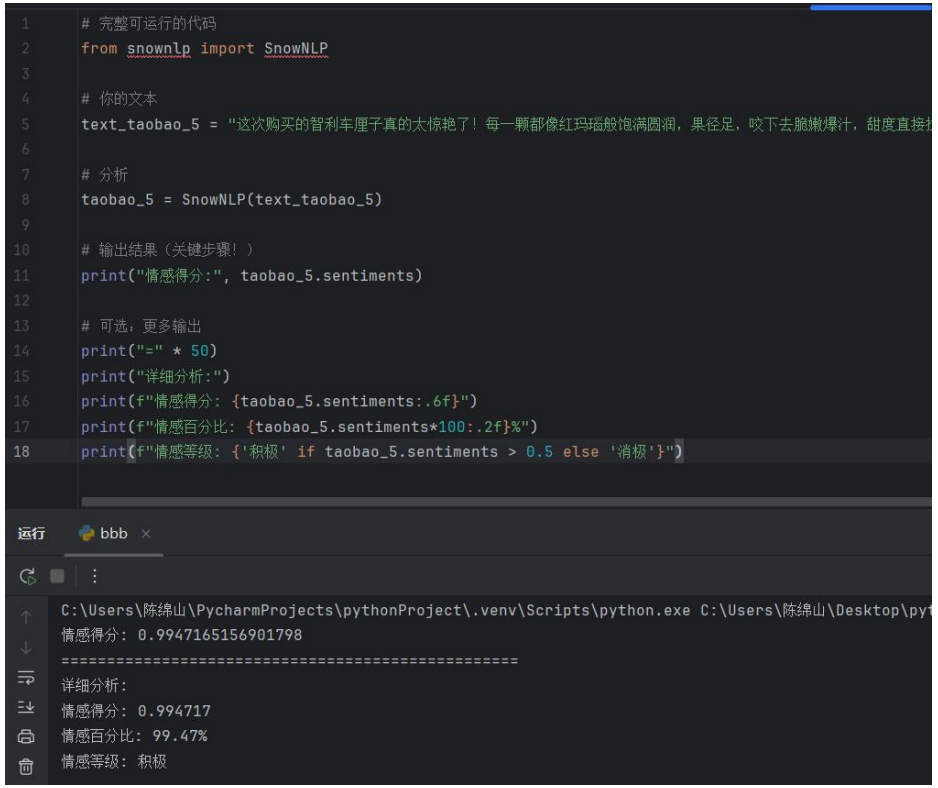
2. 4采用给的程序，实现一段科学家文本的词云图绘制，越清晰越好（生成的词云图要单独拿出来）。

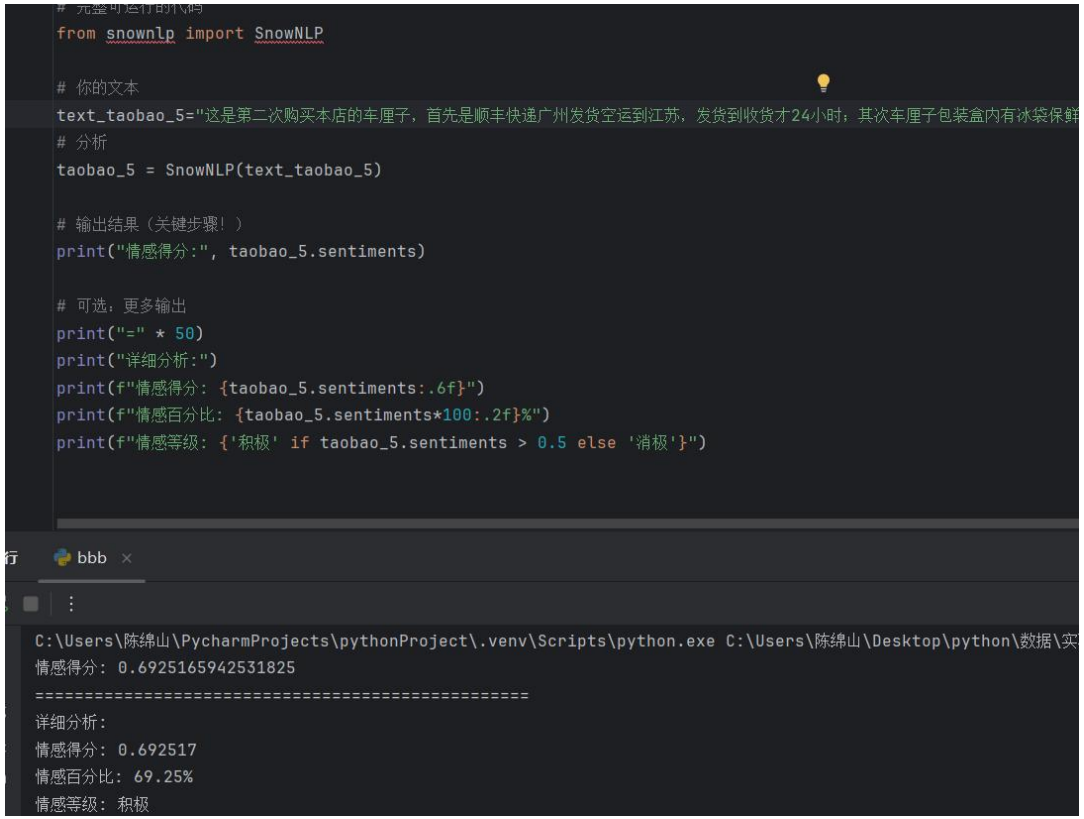
第四讲 情感分析

- 1使用PPT给的情感分析平台（或其它平台），对文本情感进行分析，并截图；

第四讲 情感分析

- 2完成sentiment_analysis_1-sentiment_analysis_4，4份代码。做截图，并简要做代码运行总结分析。





第四讲 情感分析

- .sentiment_analysis_3
- 1请问你觉得大语言模型在识别患者的身体、心理、情感的工作中，表现如何？

模型可作为辅助筛查工具，快速梳理患者主诉中的关键信息，帮助医护人员关注身心关联线索。但其结果需结合专业评估，尤其对隐性情感、复杂心理状态的识别仍需人类专业判断。

- 2除了健康领域的情感世界理解与识别，你还能够想到哪些其他重要的领域，可以开展类似的工作呢？
- 教育领域、消费者洞察、社会舆情分析、职场管理、文学/影视研究

第四讲 情感分析

sentiment_analysis_4

文学作品的情感轨迹

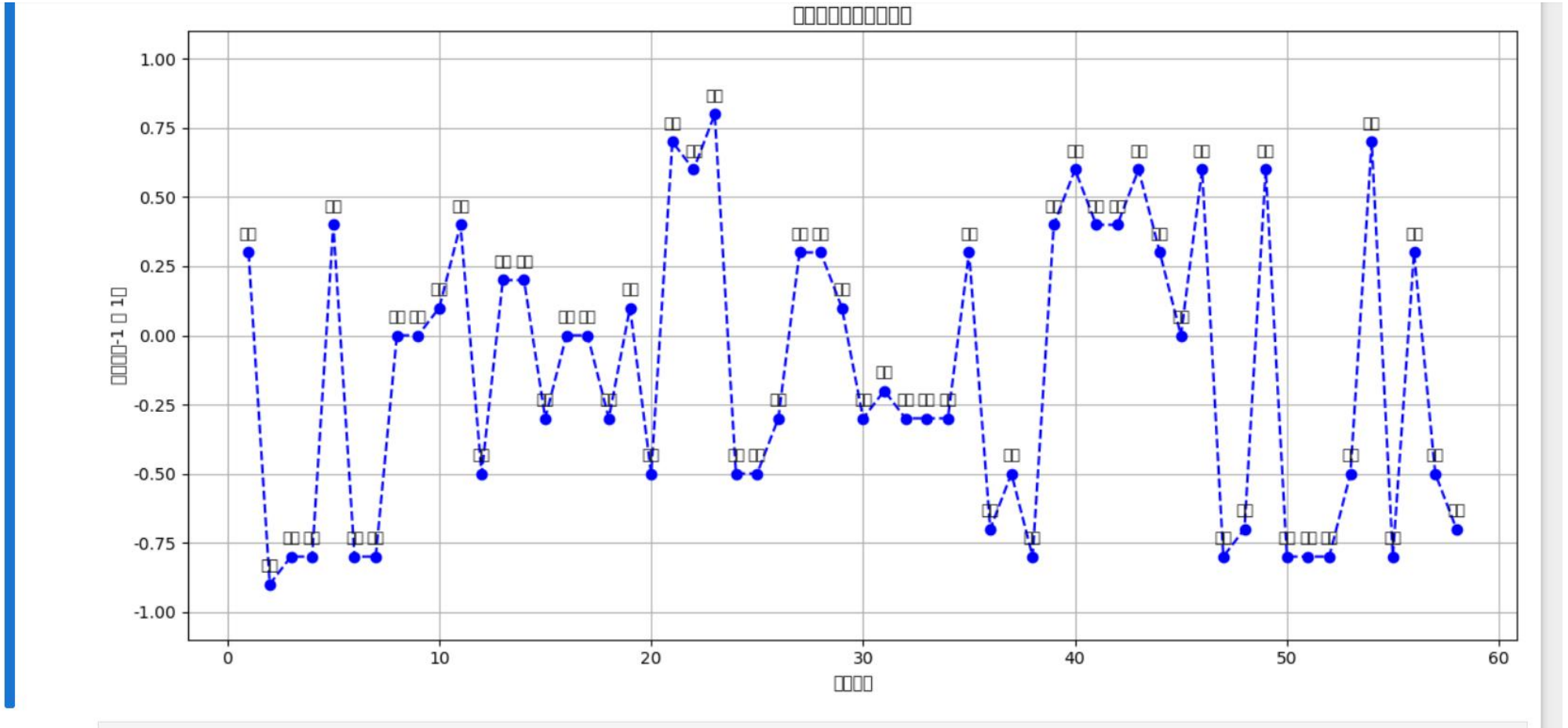
- 步骤是这样的:
- 1）读取txt文本，按照句号、感叹号，对文本进行分句，并且对句子进行编号;
- 2）使用大语言模型，提取以上编号句子，每句话中的情感词，或者隐含的情感实体;
- 3）用大语言模型计算出，每一个情感词或者情感词所在句子的情感得分;
- 4）然后绘制一个时间序列的可视化图，横坐标是句子的序号，纵坐标是情感得分，每一个散点是句子对应的情感词，情感词作为散点的标签，然后用虚线把这些情感词连起来。

[1]:
```
import re

with open("背影.txt", "r", encoding="utf-8") as f:
    text = f.read()

# 按照句号、感叹号进行分句（保留标点）
sentences = re.split(r'(?<=[。！])', text)
sentences = [s.strip() for s in sentences if s.strip()]
numbered_sentences = [(i + 1, s) for i, s in enumerate(sentences)]
#numbered_sentences = numbered_sentences[0:5] #大模型抽取很慢，提取文本的前n条句子

for num, sent in numbered_sentences:
    print(f"{num}: {sent}")
```

```
1: 我与父亲不相见已二年余了，我最不能忘记的是他的背影。
2: 那年冬天，祖母死了，父亲的差使也交卸了，正是祸不单行的日子。
3: 我从北京到徐州，打算跟着父亲奔丧回家。
4: 到徐州见着父亲，看见满院狼藉的东西，又想起祖母，不禁簌簌地流下眼泪。
5: 父亲说：“事已如此，不必难过，好在天无绝人之路！
6: ”
```

第四讲 情感分析

sentiment_analysis_4

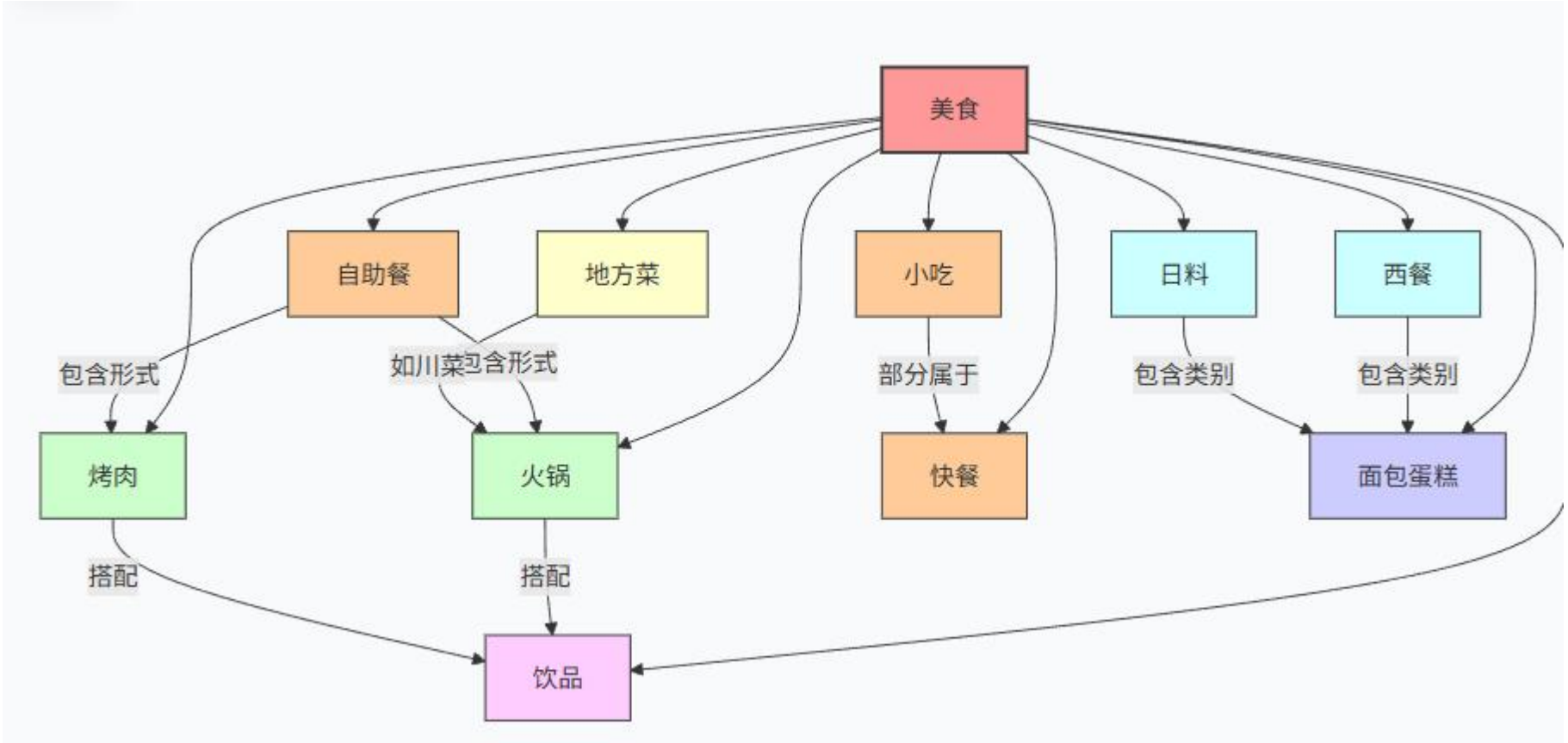


第六讲 知识图谱理念

- 美团大脑知识图谱

搜索案例：美食

美团美食知识图谱


美食
自助餐
地方菜
小吃
日料
西餐
包含形式
如川菜包含形式
部分属于
包含类别
包含类别
烤肉
火锅
快餐
面包蛋糕
搭配
搭配
饮品

美团美食知识图谱

- 小结：美团大脑知识图谱通过用户搜索内容，自动向用户推荐有关搜索内容的相关美食，做到精准推送。除此之外还对不同的美食进行智能分类，提供给用户更多样化的美食推荐。

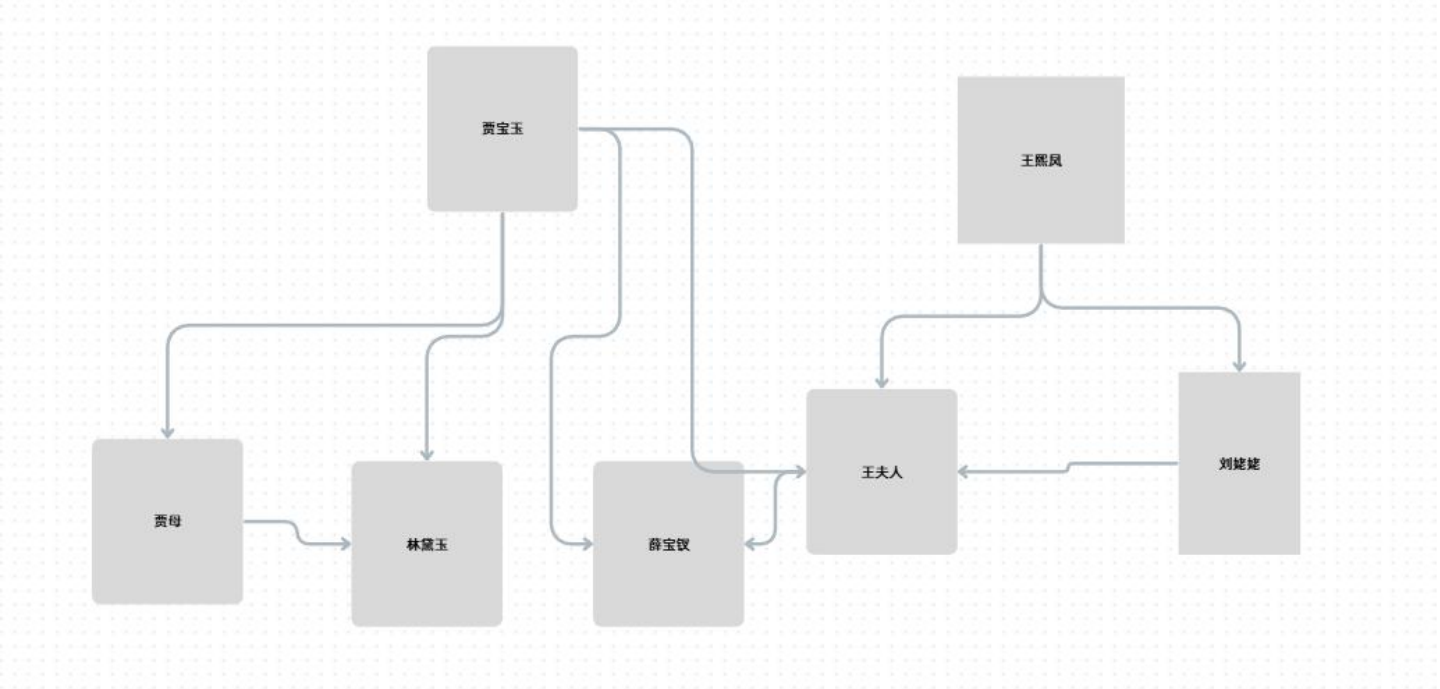
第六讲（2） 知识图谱工具

- 1使用PPT中知识图谱链接平台，检索、截图（大词林等，可用的）；

第六讲（2） 知识图谱工具

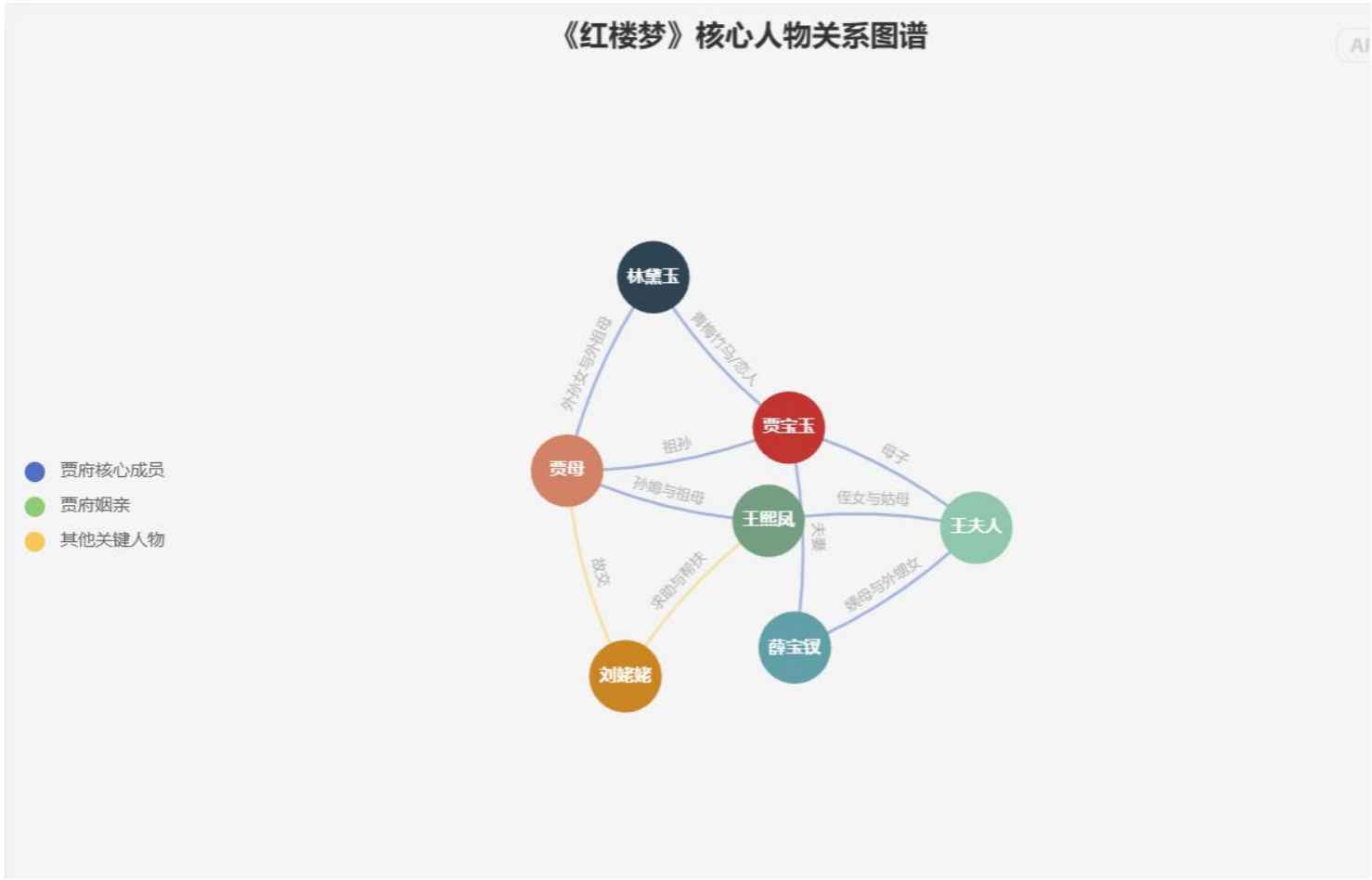
- 2.使用白板建模绘制一个你感兴趣的“知识图谱”，可以是人物关系，也可以是事物关系，或者概念之间的关系等等，并解释你绘制的图谱；

我绘制的是红楼梦中主要任务的关系知识图谱，具体的人物关系在echart制作的图谱中可以清晰的看出。



第六讲（2） 知识图谱工具

- 3使用echarts中的关系图，绘制作业2）中的“知识图谱”。



第六讲（2） 知识图谱工具

- 4使用Neo4j（可在线版本），编程绘制一款（简单）知识图谱（内容不限）
- 川菜口味知识图谱