南京工业大学
NANJING TECH UNIVERSITY

用户数据采集与关联分析

（结课作业）

班级：信管2301

学号：202321054022

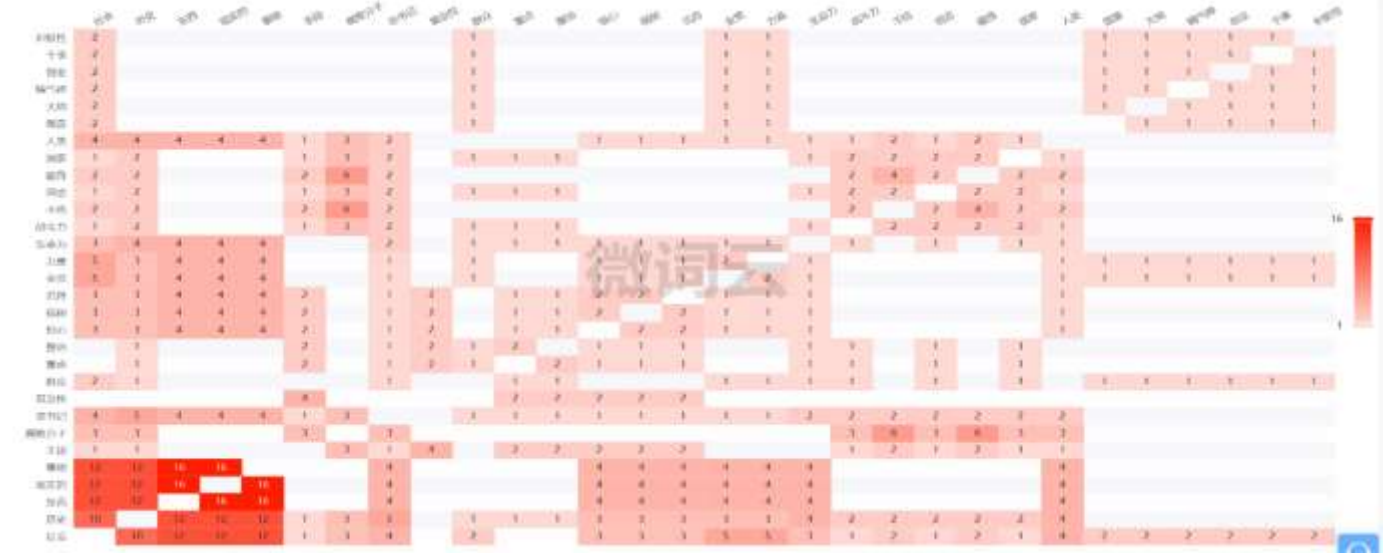
姓名：韦钟凯

南京工業大學
NANJING TECH UNIVERSITY

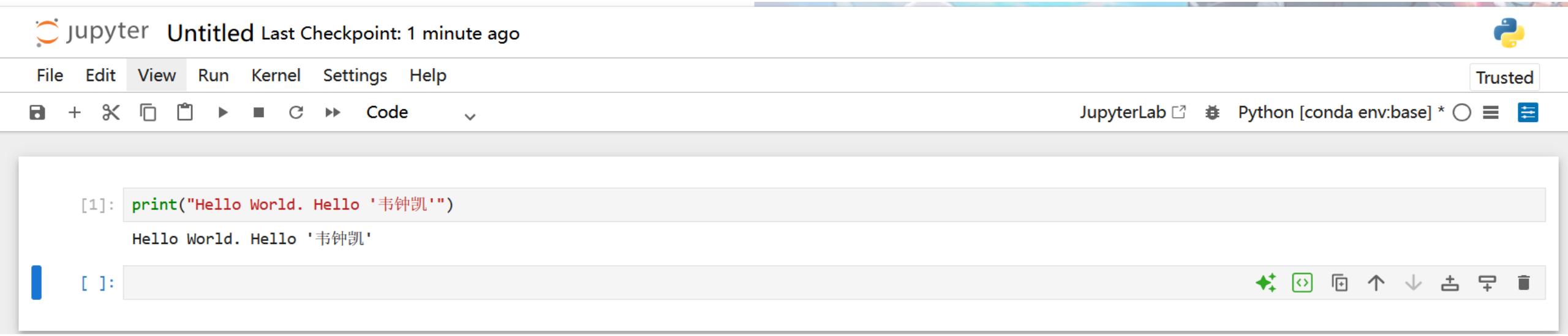
第一讲作业

第一讲1、清华大学分词演示系统（案例演示截图）；

第一讲1、微词云分词（案例演示截图）；



第一讲2.安装python（anaconda）（编写输出“Hello World. Hello ‘你的姓名’”）；



第一讲3、课后作业（001代码）

1.基本分词

[8]:
```
import jieba
```

[9]:
```
seg_list1 = jieba.cut("黄旭华，1926年3月12日出生于广东省汕尾市，原籍广东省揭阳市。1949年毕业于上海交通大学。历任北京海军核潜艇研究室副总工程师、中船重工集团公司核潜艇总
```

[10]:
```
print('$'.join(seg_list1))
```

```
黄旭华$，$1926$年$3$月$12$日出$生于$广东省$汕尾市$，$原籍$广东省$揭阳市$。$1949$年$毕业$于$上海交通大学$。$历任$北京$海军$核潜艇$研究室$副$总工程师$、$中$船$重工$集团公司$核潜艇$总体$研究$设计所$研究员$、$名誉$所长$。$1994$年$当选$为$中国工程院$院士$。
```

[11]:
```
seg_list2 = jieba.cut("LSTM（Long Short-Term Memory）是长短期记忆网络，是一种时间递归神经网络，适合于处理和预测时间序列中间隔和延迟相对较长的重要事件。")
```

[12]:
```
print('@'.join(seg_list2))
```

```
LSTM@（@Long@ @Short@-@Term@ @Memory@）@是@长短期@记忆@网络@，@是@一种@时间@递归@神经网络@，@适合@于@处理@和@预测@时间@序列@中@间隔@和@延迟@相对@较长@的@重要@事件@。
```

2.加入词典，是针对第二个片段的，希望是能够完整把"长短期记忆网络"这个术语整体分割出来 ¶

[37]:
```
jieba.load_userdict('dict.txt')
```

[38]:
```
seg_list_dict = jieba.cut("LSTM（Long Short-Term Memory）是长短期记忆网络，是一种时间递归神经网络，适合于处理和预测时间序列中间隔和延迟相对较长的重要事件。")
```

[39]:
```
print('/'.join(seg_list_dict))
```

```
LSTM/（/Long/ /Short/-/Term/ /Memory/）/是/长短期记忆网络/，/是/一种/时间递归神经网络/，/适合/于/处理/和/预测/时间/序列/中/间隔/和/延迟/相对/较长/的/重要/事件/。
```

课后作业（001代码）续

3.加入停用词，针对第一个片段，希望的结果是，结果中不会出现“的、是”等虚词

```
[13]: stopwords = [line.strip() for line in open('stop_words.txt', 'r', encoding='utf-8').readlines()]

[14]: seg_list_stopw = jieba.cut("黄旭华，1926年3月12日出生于广东省汕尾市，原籍广东省揭阳市。1949年毕业于上海交通大学。历任北京海军核潜艇研究室副总工程师、中船重工集团公司核

[15]: final = ''

[16]: #这是一行注释，进行分词结果的过滤
      for seg in seg_list_stopw:
          if seg not in stopwords:
              final += seg + '/' #叠加，累加

[17]: print(final)
```

```
黄旭华/1926/年/3/月/12/日出/生于/广东省/汕尾市/原籍/广东省/揭阳市/1949/年/毕业/于/上海交通大学/历任/北京/海军/核潜艇/研究室/副/总工程师/中/船/重工/集团公司/核潜艇/总体/研究/设计所/研究员/名誉/所长/1994/年/当选/为/中国工程院/院士/
```

分得好的地方

1.人名、地名、机构名等专有名词：“黄旭华”“广东省”“汕尾市”“揭阳市”“上海交通大学”“北京”“海军”“核潜艇研究室”“中船重工集团公司”“核潜艇总体研究设计所”“中国工程院”等专有名词都分得很准确

2.日期和时间表达：“1926年3月12日”“1949年”“1994年”等日期表达被正确分词，年份、月份和日期都被合理地分隔开来，符合常见的日期表达习惯，有助于理解时间信息。

3.标点符号的处理：标点符号（如逗号、句号）被正确地分隔出来，没有与文本内容混淆，有助于清晰地划分句子结构。

分得不好的地方

1.职务名称的拆分：“副$总工程师”被拆分成了“副”和“总工程师”，这是不合理的。职务名称“副总工程师”应该作为一个整体来分词，因为“副”是修饰“总工程师”的，拆分后会让人误解为“副”和“总工程师”是两个不同的概念或职务。

2.部分词语的拆分：“出 所长”被拆分成了“名誉”和“所长”，虽然“名誉”和“所长”可以单独作为词语，但在这里它们共同组成一个短语“名誉所长”，应该作为一个整体来分词，这样更符合语言习惯和语义表达的准确性。

课后作业（002代码）

功勋科学家-黄旭华-传记文本分词

现在，可以开启你的小组项目的第一个小小任务啦！就是对一小段有关"功勋科学家"的文本进行分词处理。

[1]:
```
# 简单分词
```

[2]:
```
import jieba
```

[3]:
```
seg_list_huang = jieba.cut('黄旭华，1926年3月12日出生于广东省汕尾市，原籍广东省揭阳市，1949年毕业于上海交通大学，历任北京海军核潜艇研究室副总工程师、中船重工集团公司核
```

[4]:
```
print('/'.join(seg_list_huang))
```

```
Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\27160\AppData\Local\Temp\jieba.cache
Loading model cost 0.535 seconds.
Prefix dict has been built successfully.
```

黄旭华/，/1926/年/3/月/12/日出/生于/广东省/汕尾市/，/原籍/广东省/揭阳市/，/1949/年/毕业/于/上海交通大学/，/历任/北京/海军/核潜艇/研究室/副/总工程师/、/中/船/重工/集团公司/核潜艇/总体/研究/设计所/研究员/、/名誉/所长/，/1994/年/当选/为/中国工程院/院士/，

[5]:
```
# 加入用户词典
```

[6]:
```
jieba.load_userdict('dict.txt')
```

[7]:
```
seg_list_huang = jieba.cut('黄旭华，1926年3月12日出生于广东省汕尾市，原籍广东省揭阳市，1949年毕业于上海交通大学，历任北京海军核潜艇研究室副总工程师、中船重工集团公司核
```

[8]:
```
print('/'.join(seg_list_huang))
```

黄旭华/，/1926/年/3/月/12/日出/生于/广东省/汕尾市/，/原籍/广东省/揭阳市/，/1949/年/毕业/于/上海交通大学/，/历任/北京/海军/核潜艇/研究室/副/总工程师/、/中船重工集团公司/核潜艇/总体/研究/设计所/研究员/、/名誉/所长/，/1994/年/当选/为/中国工程院/院士/，

[9]:
```
# 加入词典之后，哪些词汇被分出来了呢？
```

[10]:
```
# 使用停用词表
```

[11]:
```
# stopwords = [line.strip() for line in open('stop_words.txt','r', encoding='utf-8').readlines()]
```

[12]:
```
stopwords = open('stop_words.txt','r', encoding='utf-8').read()
stopwords = stopwords.split('\n')
```

[13]:
```
stopwords
```

[13]:
```
['的', '了', '是', '啊', '。', '，', '、', '使用']
```

[14]:
```
seg_list_huang = jieba.cut('黄旭华，1926年3月12日出生于广东省汕尾市，原籍广东省揭阳市，1949年毕业于上海交通大学，历任北京海军核潜艇研究室副总工程师、中船重工集团公司核
```

[15]:
```
final = ''
```

[16]:
```
for seg in seg_list_huang:
    if seg not in stopwords:
        final+= seg+'/'
```

[17]:
```
print(final)
```

黄旭华/1926/年/3/月/12/日出/生于/广东省/汕尾市/原籍/广东省/揭阳市/1949/年/毕业/于/上海交通大学/历任/北京/海军/核潜艇/研究室/副/总工程师/中船重工集团公司/核潜艇/总体/研究/设计所/研究员/名誉/所长/1994/年/当选/为/中国工程院院士/

作业的意义：

- 你可以处理比较复杂的文本啦
- 你开始尝试接触和理解，一些具有文化内涵的科技文献资源

课后作业（003代码）

```
[16]: # 输出所有词汇的词频（按频率降序）
print("\n所有词汇词频统计（前20个）：")
for word, count in word_counts.most_common(20):
    print(f"'{word}': {count}次")
```

```
所有词汇词频统计（前20个）：
'，': 13次
'
': 9次
'管理': 5次
'、': 5次
'与': 4次
'"': 3次
'企业': 3次
'"': 3次
'体系': 3次
'智能化': 3次
'经营': 3次
'打造': 3次
'能力': 3次
'供应链': 3次
'建设': 2次
'通过': 2次
'自动化': 2次
'数字化': 2次
'升级': 2次
'生产': 2次
```

提问

- 1，以上仅仅是统计了文本中的词的频次，这能反映什么呢？
- 2，词频统计本身存在什么问题呢？去停用词没有完成？
- 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？

1、以上仅仅是统计了文本中的词的频次，这能反映什么呢？

答：仅通过词频统计，可以初步反映出以下信息：

1、文本主题倾向：高频率出现的词汇（如“管理”、“企业”、“体系”、“数字化”等）可能反映了文档的核心主题。例如，“数字化”“智能化”“自动化”等词频繁出现，说明该文本聚焦于企业数字化转型。

2、关键词识别：词频可以帮助快速识别出文本中重要的术语或概念，比如“数字化”出现了2次，虽然不高，但结合上下文可判断其重要性。

然而，这种简单的词频统计只能提供**表面的信息**，无法揭示词语之间的关系、语义层次或情感色彩。

2、词频统计本身存在什么问题呢？去停用词没有完成？

① 停用词未有效去除

图片中的词频结果显示了大量无实际意义的虚词和标点符号，如：

“，”：13次

“。”：9次

“与”：“’”、“”“、“…” 等

这些属于典型的**停用词（stop words）**，它们在自然语言处理中通常被过滤掉，因为它们不携带关键语义信息。

虽然代码中调用了 jieba.lcut() 分词，并尝试使用 jieba.analyse.extract_tags() 提取关键词，但在最终输出中仍保留了这些停用词，说明**去停用词步骤未完成或未生效**。

② 缺乏分词准确性

使用 jieba 分词时，若未加载自定义词典或未进行合理配置，可能导致短语被错误切分。例如，“数字化转型”可能被切分为“数字”+“化”+“转”+“型”，从而影响统计准确性。

③ 未做词形归一化

中文中虽无严格词形变化，但同义词或近义表达（如“智能” vs “智能化”）未能统一，导致语义重复分散

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？

答：**① 仅靠词频的局限性**

无法捕捉语义关系：词频只告诉我们某个词出现了多少次，但不能说明它与其他词的关系。例如，“数字化”出现了2次，“安全”未出现在前20个词频中，但这并不意味着两者无关——它们可能在同一个句子或段落中共同出现，但因频率低而被忽略。

忽略上下文语境：词频无法体现词语之间的搭配、共现关系或因果逻辑。例如，“数字化”是否提升了“安全”水平？这需要更深层次的分析。

忽略权重差异：所有词都按次数计数，没有考虑位置、句法角色或情感极性等因素。

② 无法确定文本关键词与其他关键词的关系

例如：不能确定“数字化”与“安全”的关系当前词频统计显示“数字化”出现2次，“安全”未出现在统计列表中（即使存在也可能频率太低），因此无法判断二者是否存在关联。

即使“安全”出现，仅凭频率也无法判断它是作为“数字化”的目标、手段还是障碍。

改进方法	具体内容
上下文分析	统计词汇之间的共现频率，即两个词汇在一定范围内（如同一句子或段落）同时出现的次数。例如，统计“数字化”和“安全”在同段落中出现的次数，可以初步判断它们之间的关联
主题建模（LDA）	对全文进行主题建模，自动发现潜在主题（如“数字化转型”、“信息安全”），并观察它们是否共享同一主题。
去除停用词	在分词后加入停用词表过滤，移除“的”“了”“是”“，”“。”等无意义词。

课后作业（004代码）

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",
    "免疫治疗",
    "免疫抑制信号通路",
    "个体化免疫治疗"
  ]
}
```

提问：

- 1，使用deepseek开展工作的感觉如何？
- 2，你觉得大语言模型的活干的怎么样？
- 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？

1. 使用DeepSeek开展工作的感觉如何？

答：使用DeepSeek进行科研辅助的感觉还可以！就像有个24小时的学术伙伴在身边。它能帮我快速理解复杂的专业概念，处理那些冗长的文献资料。最让我惊喜的是编程支持——写代码、调试、学新框架都变得容易多了。使用费用也比较低，这点对预算有限的学生比较友好，不用担心使用次数限制，可以大胆尝试各种想法。整体感觉很自然流畅，但是跟Deepseek交流还是存在一定困难的，有时候它呈现的效果没有达到我的预期，可能还是需要优化一下表达的语言，总之deepseek就是一个百科全书一样的伙伴。

2. 你觉得大语言模型的活干的怎么样？

答：从刚才的实体抽取结果看，大语言模型干得相当专业！它准确识别了：

理论层面：肿瘤免疫微环境、免疫编辑理论等核心概念

方法技术：单细胞RNA测序、轨迹分析等具体方法

工具软件：Seurat、Monocle3等专业工具

专业术语：包括缩略语和特定信号通路

分类清晰，没有遗漏关键术语，还做了合理的归类。对于科研文本的理解深度让我印象深刻，这已经达到了专业信息抽取的水平。

课后作业（004代码）续

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",
    "免疫治疗",
    "免疫抑制信号通路",
    "个体化免疫治疗"
  ]
}
```

提问：

- 1，使用deepseek开展工作的感觉如何？
- 2，你觉得大语言模型的活干的怎么样？
- 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？

1. 识别关系的能力

从提供的实体和术语中，AI可以识别出以下关系：

肿瘤免疫微环境（TIME）与非小细胞肺癌： TIME是非小细胞肺癌等肿瘤疾病的重要特征，它影响着肿瘤的进展和治疗反应。

T细胞耗竭与免疫抑制信号通路： T细胞耗竭是免疫抑制信号通路（如PD-1/PD-L1、TGF-β路径）过度激活的结果，这会导致T细胞功能下降，影响抗肿瘤免疫反应。

单细胞RNA测序（scRNA-seq）与细胞亚群聚类： scRNA-seq是一种强大的方法，用于研究细胞亚群的特征和功能，通过聚类分析可以发现不同细胞亚群在肿瘤免疫微环境中的作用。

细胞间通讯网络构建与免疫编辑理论： 细胞间通讯网络的构建有助于理解免疫细胞与肿瘤细胞之间的相互作用，这与免疫编辑理论密切相关，因为免疫编辑理论描述了肿瘤如何逃避免疫系统的监视。

Seurat、Monocle3、CellChat与单细胞RNA测序： 这些工具（Seurat用于细胞聚类和数据整合，Monocle3用于轨迹分析和pseudotime推断，CellChat用于细胞间通讯网络构建）是单细胞RNA测序数据分析的关键工具，它们帮助研究人员从不同角度分析和理解细胞行为和相互作用

第一讲4、关于《基于学术论文全文的研究方法实体自动识别研究_章成志》阅读总结

随着科技文献数量的快速增长，研究方法作为学术论文中的核心组成部分，其系统化识别与总结对于把握学科发展动态、提升科研效率具有重要意义。然而，现有研究多聚焦于理论探讨或仅从摘要中提取方法信息，未能充分利用全文中所包含的丰富方法论描述，也缺乏对“使用”与“引用”研究方法的明确区分。因此，本文提出从学术论文全文出发，开展研究方法实体的自动识别研究，并深入分析其在情报学领域的分布特征与应用趋势。

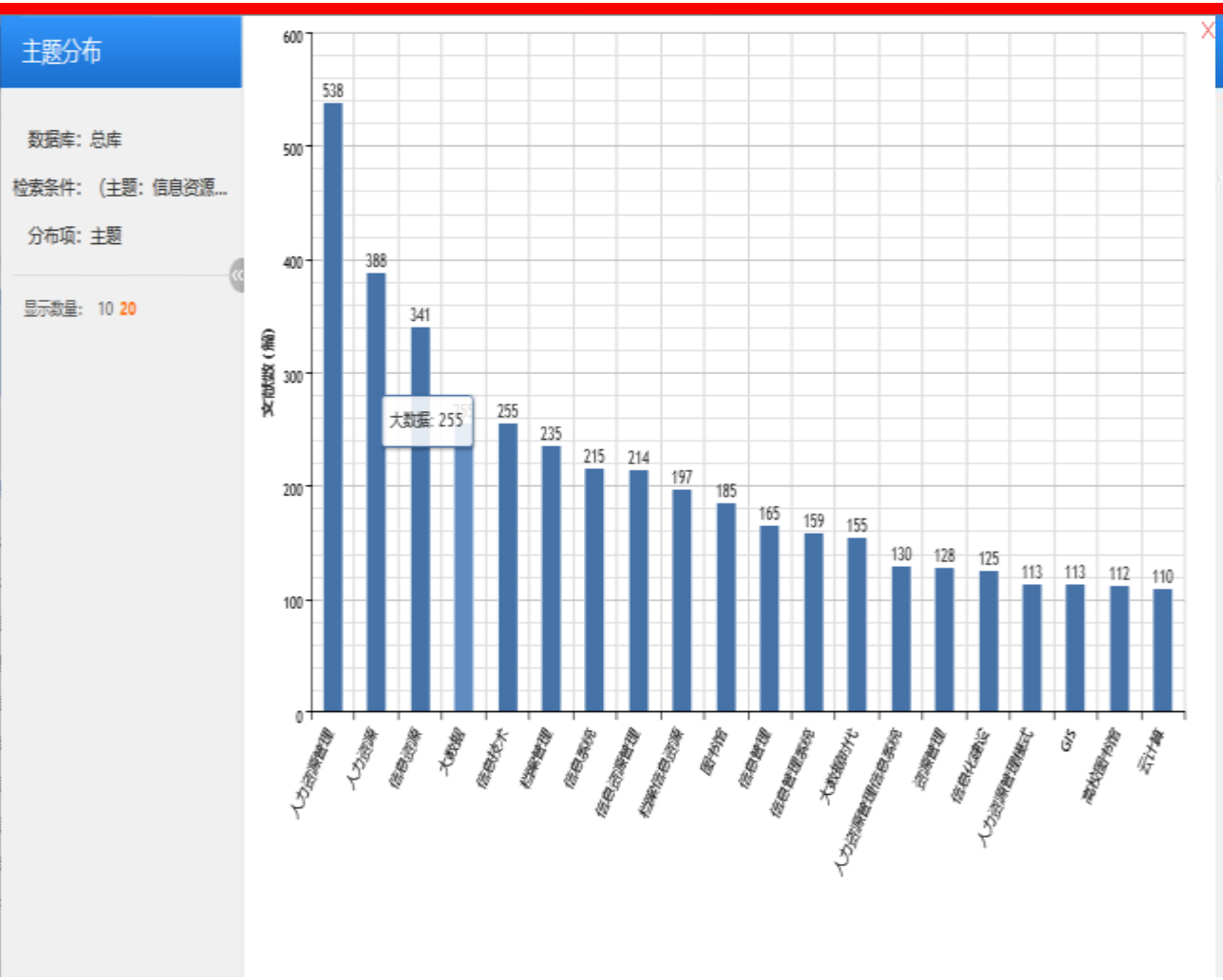
本文通过系统比较包括BiLSTM、BiLSTM-CRF在内的8种神经网络模型，结合字向量与领域词向量，构建了针对“论文使用研究方法”与“论文引用研究方法”的两类识别模型。实验表明，基于字向量与word2vec领域词向量的Char-CRF-BiLSTM模型在两类任务中均表现最优。基于《情报学报》近十年论文的实证分析发现，情报学领域内实验法相关研究方法的使用与引用频次均最高，反映出该学科对实证研究的重视。

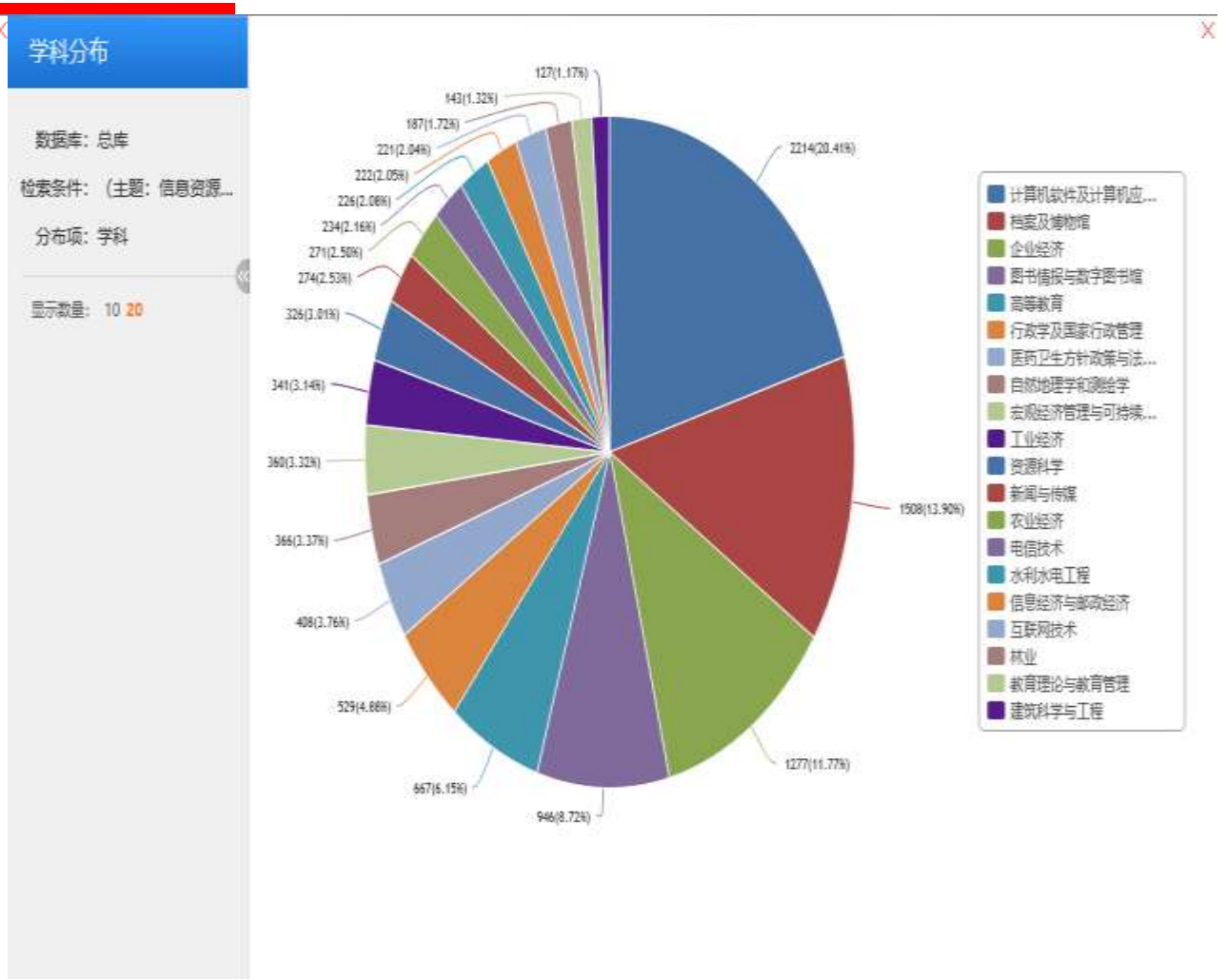
未来工作将进一步引入半监督学习方法以减少对标注数据的依赖，融合规则增强模型的准确性与可解释性，并扩展至更多情报学期刊文献，以构建更全面、动态的研究方法知识体系。

南京工業大學
NANJING TECH UNIVERSITY

第二讲作业

第二讲1、基于CNKI数据库统计分析2014-2024年（近10年），“信息资源管理” 主题变化趋势。





第二讲2、作业2和3:代码1

全文本词频统计的步骤 ¶

- 打开文本
- 分词
- 去除停用词 (集合的方式)
- 利用字典，进行词-词频的存储
- 排序
- 展示 (输出print)
-
- 这个是非常简单的python程序

```
[15]: import jieba

[16]: article = open('科学家博物馆-黄旭华传记序言.txt','r',encoding = 'utf-8').read() # 打开并读取三国前10回 #出现乱码提示，就把ANSI改成utf-8

[17]: dele = {'。','！','？','的','“','”','（','）',' ','》','《','、'} # 手动设计一些停用词和符号

[18]: jieba.add_word("国立交通大学") # 加入字典中没有的新词

[19]: words = list(jieba.cut(article)) #结巴分词出来的词汇

[20]: words

[20]: ['在',
       '核潜艇',
       '领域',
       '，',
       '我国',
       '已',
       '形成',
       '一套',
       '完整',
       '的',
       '研究',
       '、',
       '设计',
       '、',
       '试验',
       '、',
       '制造',
       '、'
```

集合 (数据结构的一种)

```
[22]: articleSet = set(words)-dele

[23]: articleSet

[23]: {'\n',
       '09',
       '1020',
       '1924',
       '1949',
       '1958',
       '1978',
       '1982',
       '1985',
       '1986',
       '1989',
       '1996',
       '2',
       '2013',
       '24',
       '29',
       '648',
       '7'

[24]: # 因为采用的是循环的方法，把所有的词都循环一遍，长度大于1，所以速度很慢！

[25]: for w in articleSet:
          if len(w)>1:
              articleDict[w] = words.count(w)  # 字典的key-value配对

[26]: articleDict

[26]: {'成熟期': 2,
       '这里': 1,
       '特点': 3,
       '李世英': 3,
       '华丽': 1,
       '战备': 1,
       '理想': 1,
       '资料': 21,
       '最终': 1,
       '采取': 2,
```

作业2和3:代码2

```
[1]: !pip install jieba
```

```
Requirement already satisfied: jieba in d:\anaconda3\lib\site-packages (0.42.1)
```

```
[2]: import jieba
```

```
[3]: article = open('sanguo_10.txt','r',encoding = 'utf-8').read() # ansi/打开并读取三国前10回 #出现乱码提示，就把ANSI改成utf-8
```

```
[4]: dele = ['。','！','？','的','“','”','（','）',' ','》','《','，'] # 手动设计一些停用词和符号
```

```
[5]: jieba.add_word('羅貫中') # 加入字典中没有的新词
```

```
Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\27160\AppData\Local\Temp\jieba.cache
Loading model cost 0.858 seconds.
Prefix dict has been built successfully.
```

```
[6]: words = list(jieba.cut(article)) #经已分词出来的词汇
```

```
[7]: words
```

```
[7]: ['《',
 '三国演义',
 '》',
 '（',
 '全',
 '）',
 '\n',
 '\n',
 '（',
 '明',
 '）',
 '羅貫中',
 '著',
 '\n',
 '\n',
 ' ',
 ' ',
 '\n'
```

字典

```
[8]: articleDict = {} # 这是一个字典，准备词-词频的保存
```

集合（数据结构的一种）

```
[9]: articleSet = set(words)-dele
```

```
[10]: articleSet
```

```
 '屢',
 '城',
 '此卑官',
 '提',
 '權軍馬',
 '袁隗',
 '諸侯立',
 '境界',
 '郭汜撥',
 '聽',
 '撫安',
 '依言',
 '荒地',
 '離十',
 '擊地',
 '此書',
 '再議',
 '先者',
```

```
[11]: # 因为采用的是循环的方法，把所有的词都循环一遍，长度大于1，所以速度很慢！
```

```
[12]: for w in articleSet:
          if len(w)>1:
              articleDict[w] = words.count(w)  # 字典的key-value配对
```

```
[13]: articleDict
```

```
[13]: {'布票': 1,
 '無門路': 1,
 '使議鞭': 1,
 '家家': 1,
 '此卑官': 1,
 '權軍馬': 1,
```

作业2和3:代码3

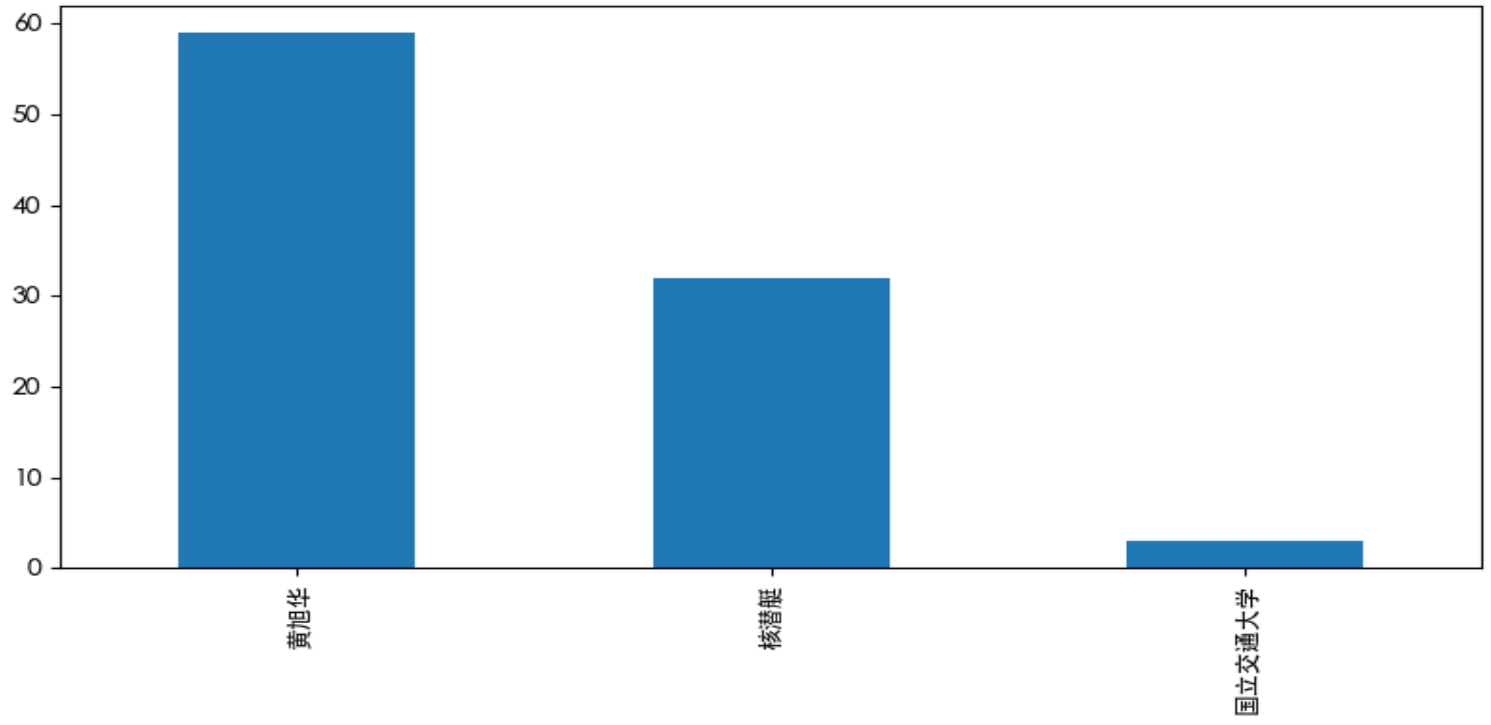
[14]:
```
# 定义 画图 函数
```

[15]:
```
def draw_dict(mydict, figsize=(8, 5)):
    import pandas as pd
    import matplotlib.pyplot as plt
    make_chinese_plot_ready()
    df = pd.DataFrame(list(mydict.items()), columns=['name', 'times'])
    df.set_index('name')['times'].sort_values(ascending=False).plot(kind='bar', figsize=figsize)  # 做好排序
    plt.tight_layout()
```

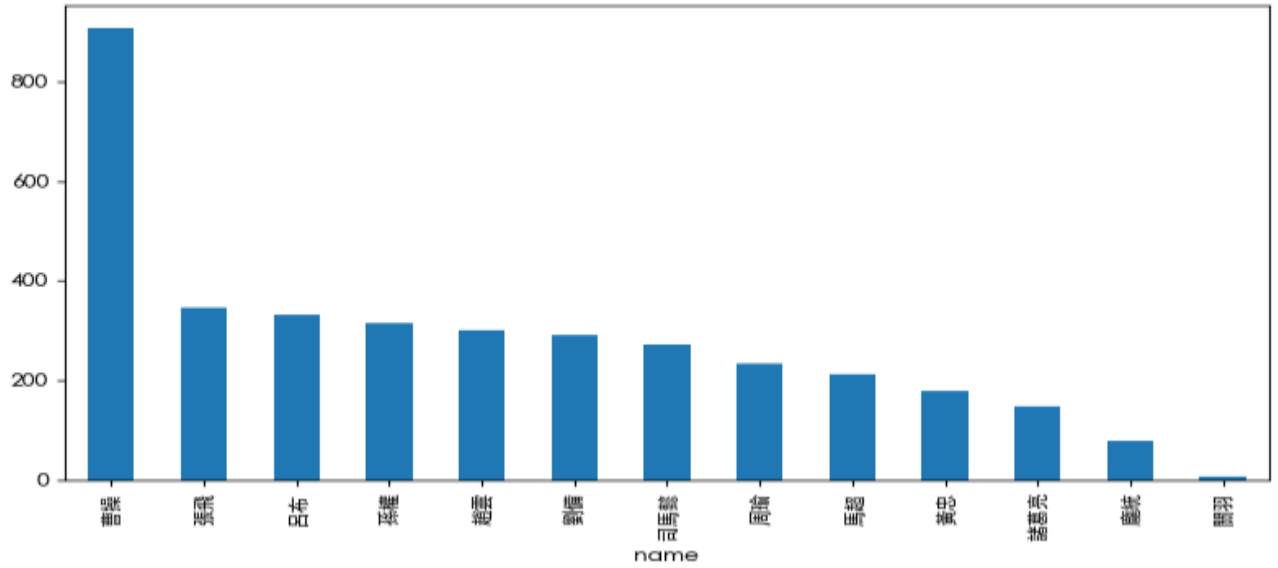
[16]:
```
 # %pylab inline
```

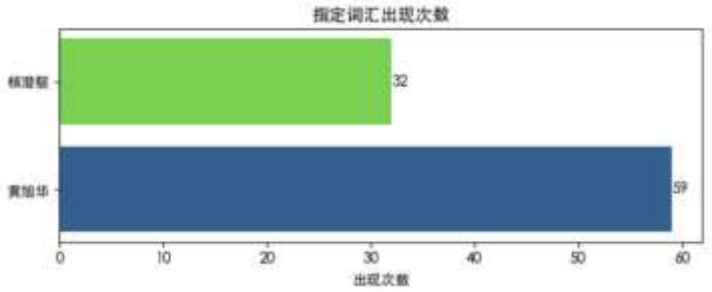
[17]:
```
%matplotlib inline
```

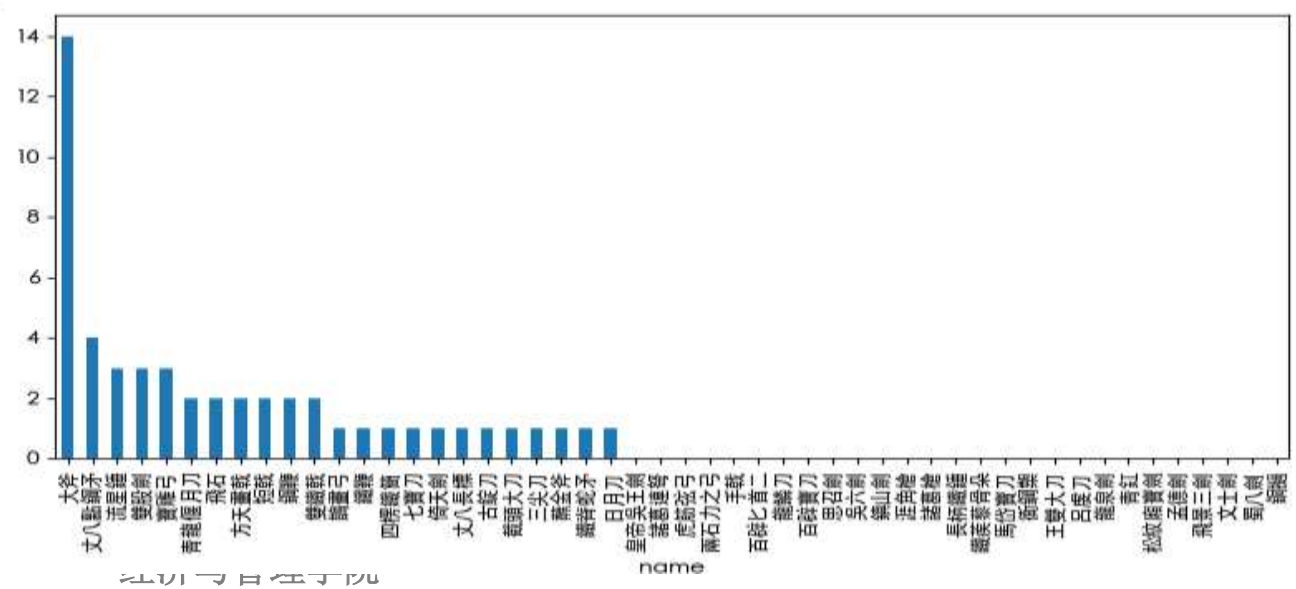
[18]:
```
draw_dict(terms_dict)
```



作业2和3:代码4


800
600
400
200
0
name


指定词汇出现次数
32
59
0
10
20
30
40
50
60
出现次数


14
12
10
8
6
4
2
0
name

第二讲4、关于《 2018-Wang 等 - Long live the scientists Tracking the scientific” 》阅读总结

科学家声誉的长期追踪研究

这篇研究探讨如何量化科学家的长期影响力与声誉。研究团队创新性地采用Google数字图书与学术文献数据库，通过分析科学家姓名在书籍中的提及频率，来追踪其声誉的历时演变。

研究以物理学领域的牛顿与爱因斯坦为典型案例。核心发现表明，伟大科学家的影响力具有惊人的持久性，跨越数百年。同时，研究揭示了显著的“群体偏好”现象：科学家在本国或使用其母语的群体中享有更高声誉。例如，在英文书籍中，牛顿的提及率更高；而在全球范围内，爱因斯坦的学术声誉自20世纪中期后已实现超越。

此外，通过共现分析发现，科学家的声誉与其核心成就紧密关联，但个人轶事与哲学思想等也构成其公众形象的一部分。

本研究的启示在于，它提供了一种超越短期学术引文、衡量科学家更广泛社会与文化影响的创新方法。该方法尤其适用于评估顶尖科学家的历史地位，并为科学计量学开辟了新的分析维度。研究也指出，该方法需注意数据在语言覆盖和姓名消歧等方面的局限性。

南京工業大學
NANJING TECH UNIVERSITY

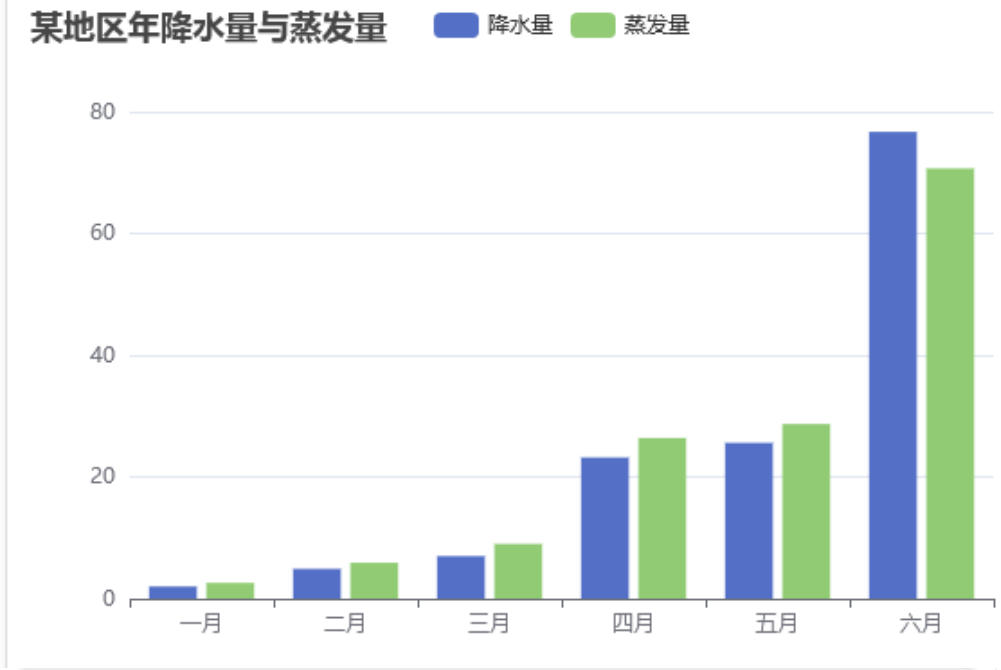
第三讲作业

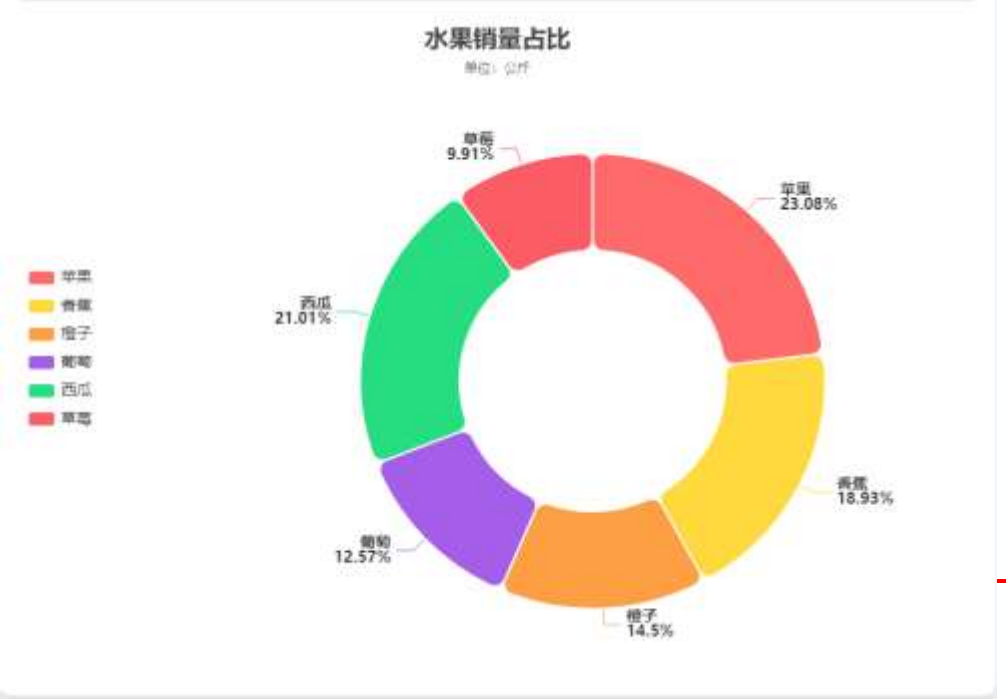
第三讲1、用微词云工具，制作一个好看的词云，并对词云图有一段话的解释

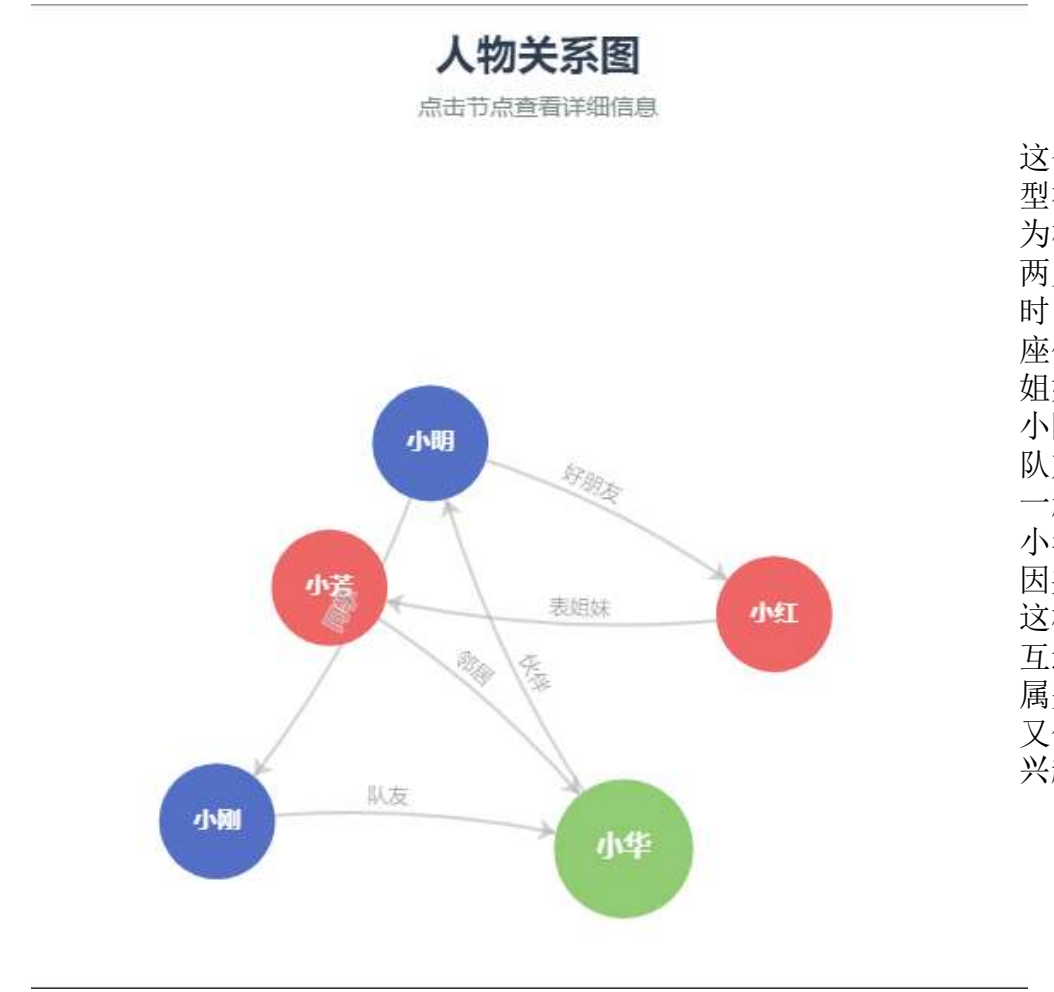
一段话解释：

这张词云图以“反腐败”为核心主题，突出呈现了“腐败”“斗争”“社会”“革命”“党中央”“全面从严治党”等高频关键词，直观反映了文本中关于党风廉政建设、政治生态净化和坚决反对腐败的鲜明立场，彰显了新时代推进自我革命、维护社会公平正义与巩固党的执政根基的强烈政治意志与时代要求。

第三讲2、使用Echarts，制作3个图

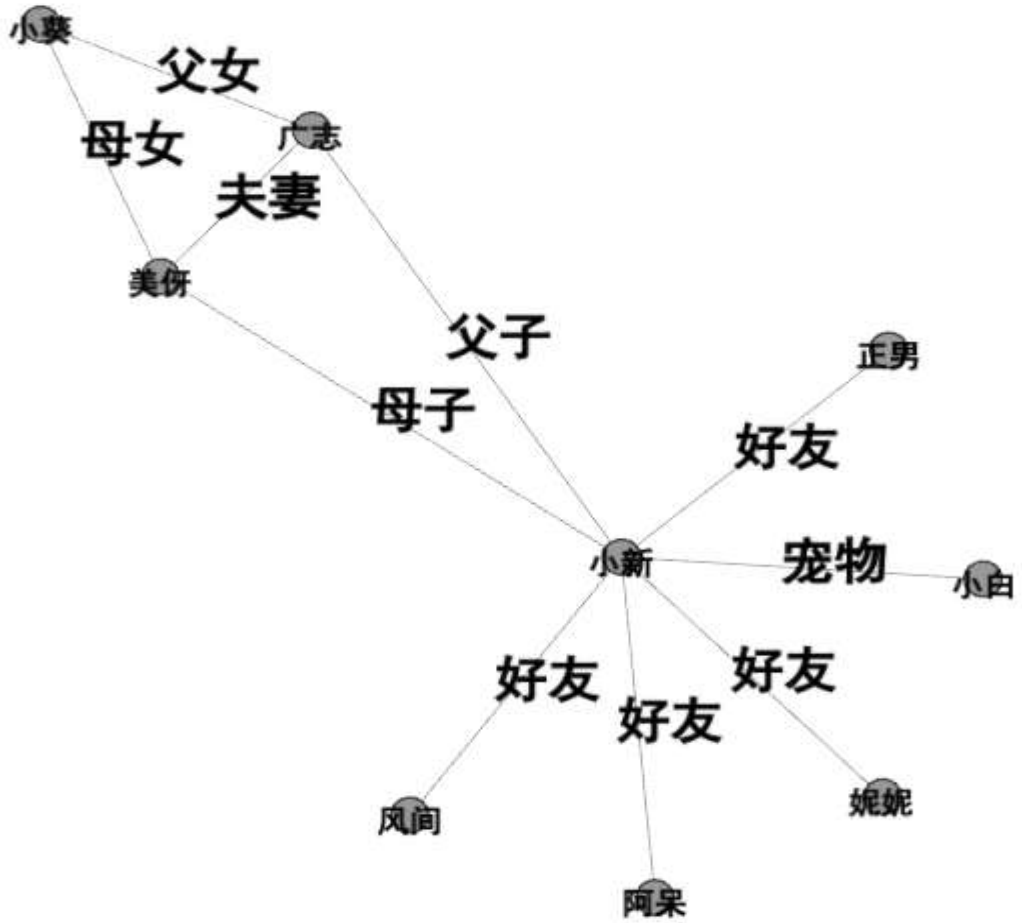




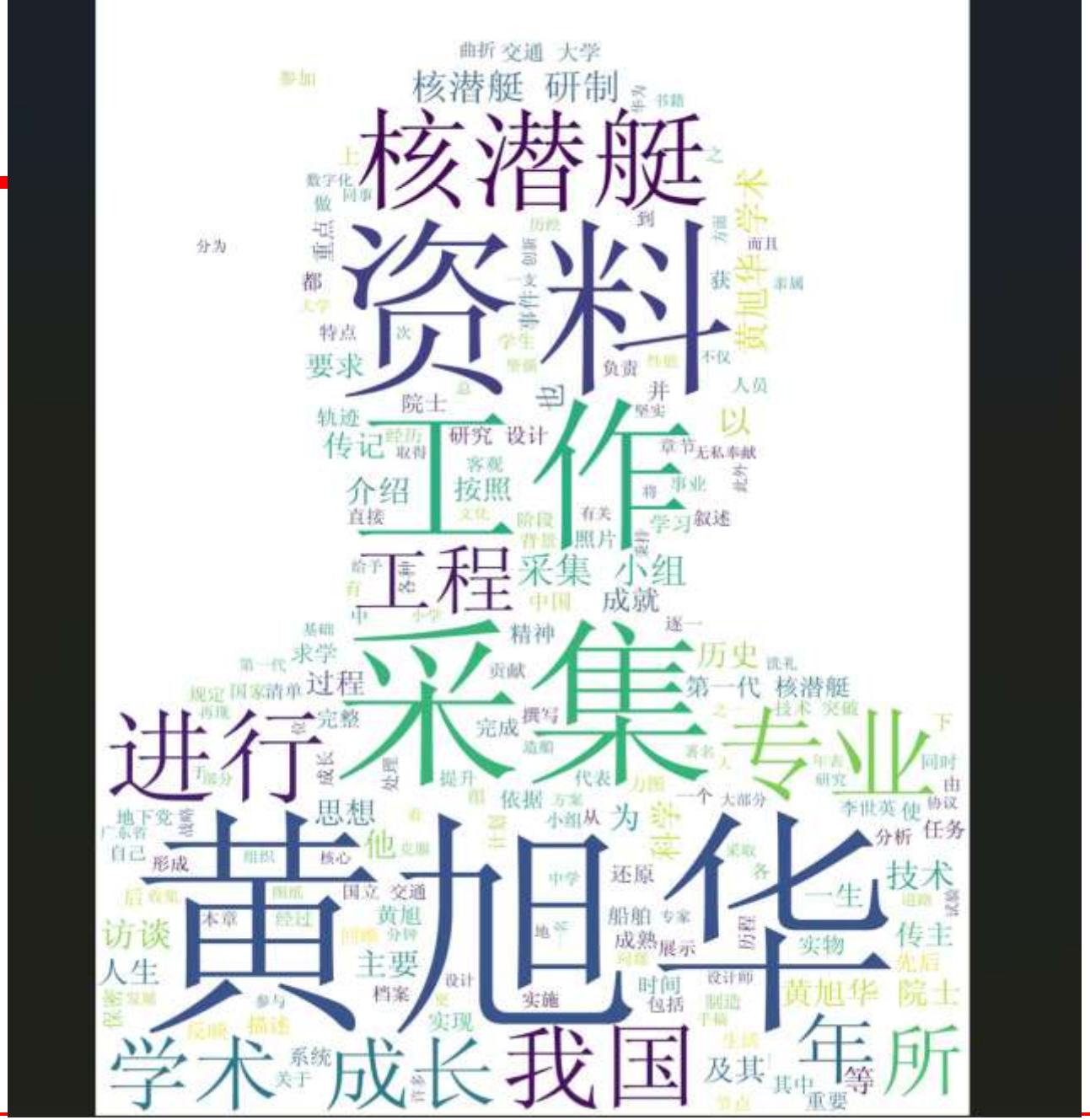


这个人物关系图展示了一个小型社交圈的人际互动：小明作为核心人物，与小红是好朋友，两人经常一起学习和玩耍；同时，小明与小刚是同班同学，座位相邻。小红则与小芳是表姐妹关系，周末常一起逛街。小刚在篮球队中与小华是默契队友，而小芳与小华又是从小一起长大的邻居。有趣的是，小华和小明在课外兴趣小组中因共同喜欢编程而成为伙伴。这样，五个人通过多重关系相互连接，形成了一个既包含亲属关系（表姐妹）、同学关系，又包含友谊、邻里关系和共同兴趣的紧密社交网络

第三讲3、使用Gehpi制作《蜡笔小新》人物关系图谱



第三讲4、采用给的程序，实现一段科学家文本的词云图绘制



南京工業大學
NANJING TECH UNIVERSITY

第四讲作业

第四讲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（文本来源：https://www.sx-dj.gov.cn/dylt/jwsp/2004358713224482818.html）

第四讲2、代码运行1

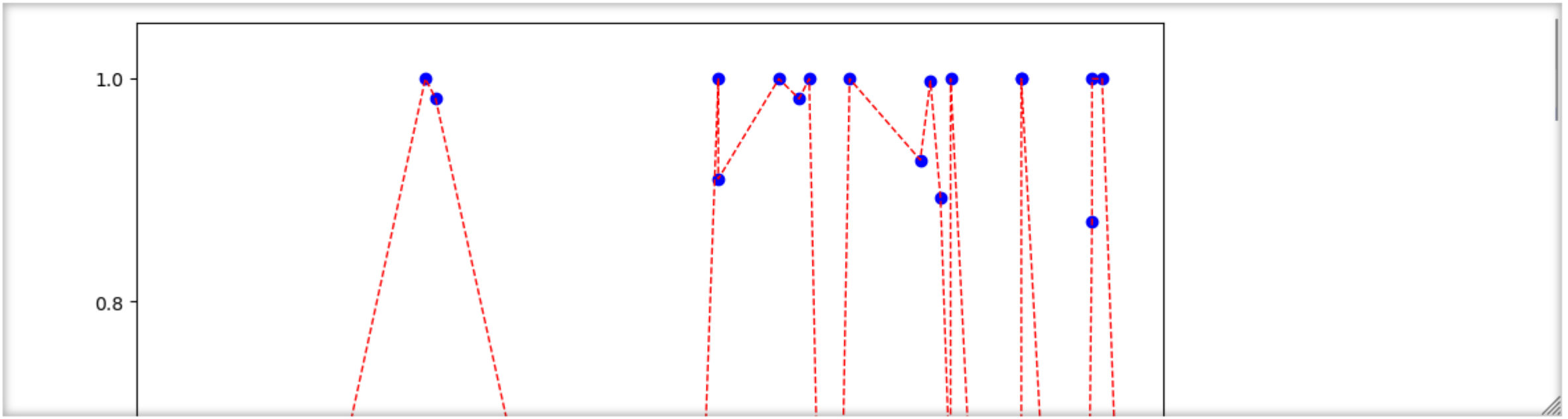
```
[24]: # 例如
text_taobao_1 = "显示效果：挺好的 运行速度：目前来说很流畅 拍照效果：拍照效果挺好的 电池续航：一天一冲 总结：目前没啥毛病，用了一天的体验"

[25]: taobao_1 = SnowNLP(text_taobao_1)

[26]: taobao_1.sentiments

[26]: 0.999947261146611

[27]: text_taobao_2 = "总结：这是我买过最不满意的一款手机！两千多元的手机这样，真的很不值！"

[28]: taobao_2 = SnowNLP(text_taobao_2)

[29]: for sentence in taobao_2.sentences:
          print(sentence)

      总结：这是我买过最不满意的一款手机
      两千多元的手机这样
      真的很不值

[30]: taobao_2.sentiments

[30]: 0.889005139666256

[31]: # 以上的结果看上去是有问题的。分析的不准确。

[32]: text_taobao_3 = "显示效果：像素不行 运行速度：微信有时发给不了语音，得重新开机后才能发，才买半个月的手机就这样，客服态度也很差！ 拍照效果：拍照不清晰！ 电池续

[33]: taobao_3 = SnowNLP(text_taobao_3)

[34]: taobao_3.sentiments

[34]: 5.7076094222563434e-05
```

简要做代码运行总结分析

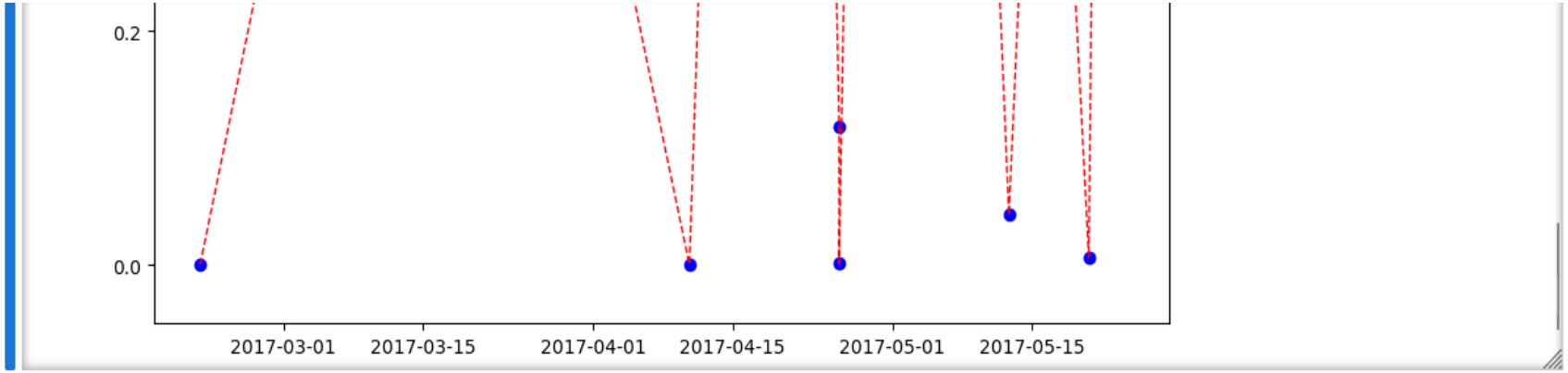
答：正面评论的代码运行得到了正确的情绪得分，而负面的评论代码运行却得到了接近1 的得分，得到了错误的结果，说明情感判断可能存在偏差或训练数据不匹配的问题。

代码运行2

[16]:
```
import matplotlib.pyplot as plt
#plt.scatter(df["date"],df["sentiments"],color='blue',figsize=(10,12))
plt.figure(figsize=(10, 12))
plt.scatter(df["date"],df["sentiments"],color='blue')
df = df.sort_values("date")
plt.plot(df["date"],df["sentiments"],color='red',linestyle='--',linewidth =1)
plt.show()
```



代码运行2续



[17]:
```
plt.savefig('timeline.png') # 看不到？改一改？
```

在图中，我们发现许多正面评价情感分析数值极端的高。同时，我们也清晰地发现了那几个数值极低的点。对应评论的情感分析数值接近于0。这几条评论，被Python判定为基本上没有正面情感了。

从时间上看，最近一段时间，几乎每隔几天就会出现一次比较严重的负面评价。

作为经理，你可能如坐针毡。希望尽快了解发生了什么事儿。你不用在数据框或者Excel文件里面一条条翻找情感数值最低的评论。Python数据框Pandas为你提供了非常好的排序功能。假设你希望找到所有评论里情感分析数值最低的那条，可以这样执行：

[18]:
```
df.sort_values(['sentiments'])[:1]
```

[18]:

	comments	date	sentiments
24	这次是在情人节当天过去的，以前从来没在情人节正日子出来过，不是因为没有男朋友，而是感觉哪哪人...	2017-02-20 16:00:00	6.334066e-08

情感分析结果数值几乎就是0啊！不过这里数据框显示评论信息不完全。我们需要将评论整体打印出来。

[19]:
```
print(df.sort_values(['sentiments']).iloc[0].comments)
```

这次是在情人节当天过去的，以前从来没在情人节正日子出来过，不是因为没有男朋友，而是感觉哪哪人都多，所以特意错开，这次实在是馋A餐厅了，所以赶在正日子也出来了，从下午四点多的时候我看排号就排到一百多了，我从家开车过去得堵的话一个小时，我一看提前两个小时就在网上先排着号了，差不多我们是六点半到的，到那的时候我看号码前面还有才三十多号，我想着肯定没问题了，等一会就能吃上的，没想到悲剧了，就从我们到那坐到等位区开始，大约是十分二十分一叫号，中途多次我都想走了，哈哈，哎，等到最后早上九点才吃上的，服务员感觉也没以前清闲时周到了，不过这肯定的，一人负责好几桌，今天节日这么多人，肯定是很累的，所以太多也都是我自己跑腿，没让服务员给弄太多，就虾滑让服务员下的，然后环境来说感觉卫生方面是不错，就是有些太吵了，味道还是一如既往的那个味道，不过A餐厅最人性化的就是看我们等了两个多小时，上来送了我们一张打折卡，而且当次就可以使用，这点感觉还是挺好的，不愧是A餐厅，就是比一般的要人性化，不过这次就是选错日子了，以后还是得提前预约，要不就别赶节日去，太火爆了！

简要做代码运行总结分析

答：这段代码通过Python实现了对特定时间段内用户评论的情感分析，并利用matplotlib绘制了情感趋势图。图表展示了2017年3月至5月期间，每天评论的情感得分（范围从0到1），其中蓝色点代表每条评论的情感值，红色虚线表示时间趋势。观察发现，尽管有多个接近1.0的正面评价，但也存在几处极低点，特别是2017-04-08左右的一条负面评价，涉及情人节排队、服务慢等问题。代码还找出了情感最低的评论，帮助识别服务中的关键问题。此分析揭示了顾客情绪的高峰与低谷，指出节假日或高峰期的服务压力会导致极端负面反馈，为改进服务提供了依据。总体上，这是一种数据驱动决策的有效方式，有助于快速定位问题并采取措施提升客户满意度。

代码运行3

细粒度情感实体抽取结果：

```json
{
  "实体": [
    {
      "部位": "头部",
      "症状": "头痛",
      "情感": "无具体描述"
    },
    {
      "部位": "全身",
      "症状": "疲乏无力",
      "情感": "无具体描述"
    },
    {
      "部位": "皮肤",
      "症状": "异常敏感，触碰疼痛",
      "情感": "无具体描述"
    },
    {
      "部位": "心脏",
      "症状": "心慌",
      "情感": "无具体描述"
    },
    {
      "部位": "胸部",
      "症状": "胸闷",
      "情感": "无具体描述"
    },
    {
      "部位": "背部",
      "症状": "沉重感",
      "情感": "无具体描述"
    },
    {
      "部位": "感官",
      "症状": "对光线和声音极度敏感",
      "情感": "惊恐"
    },
    {
      "部位": "心理",
      "症状": "不愿出门，不想与人交流",
      "情感": "生活毫无意义"
    }
  ]
}
```

提问

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？

答：优点：一、情感识别精准度高，当将专家知识直接整合到LLMs中后，其情感分析准确性与人类专家相当。二、具有个性化服务潜力，大语言模型能够为患者提供个性化的建议、定制化的治疗方案，并持续监测康复进展，使医疗服务更加以患者为中心。三、根据数据收集发现，大语言模型具有多模态融合能力：多模态医疗模型（如LLaVA-Med）能整合视觉和语言信息，根据生物医学图像生成诊断描述，提高诊断准确性。

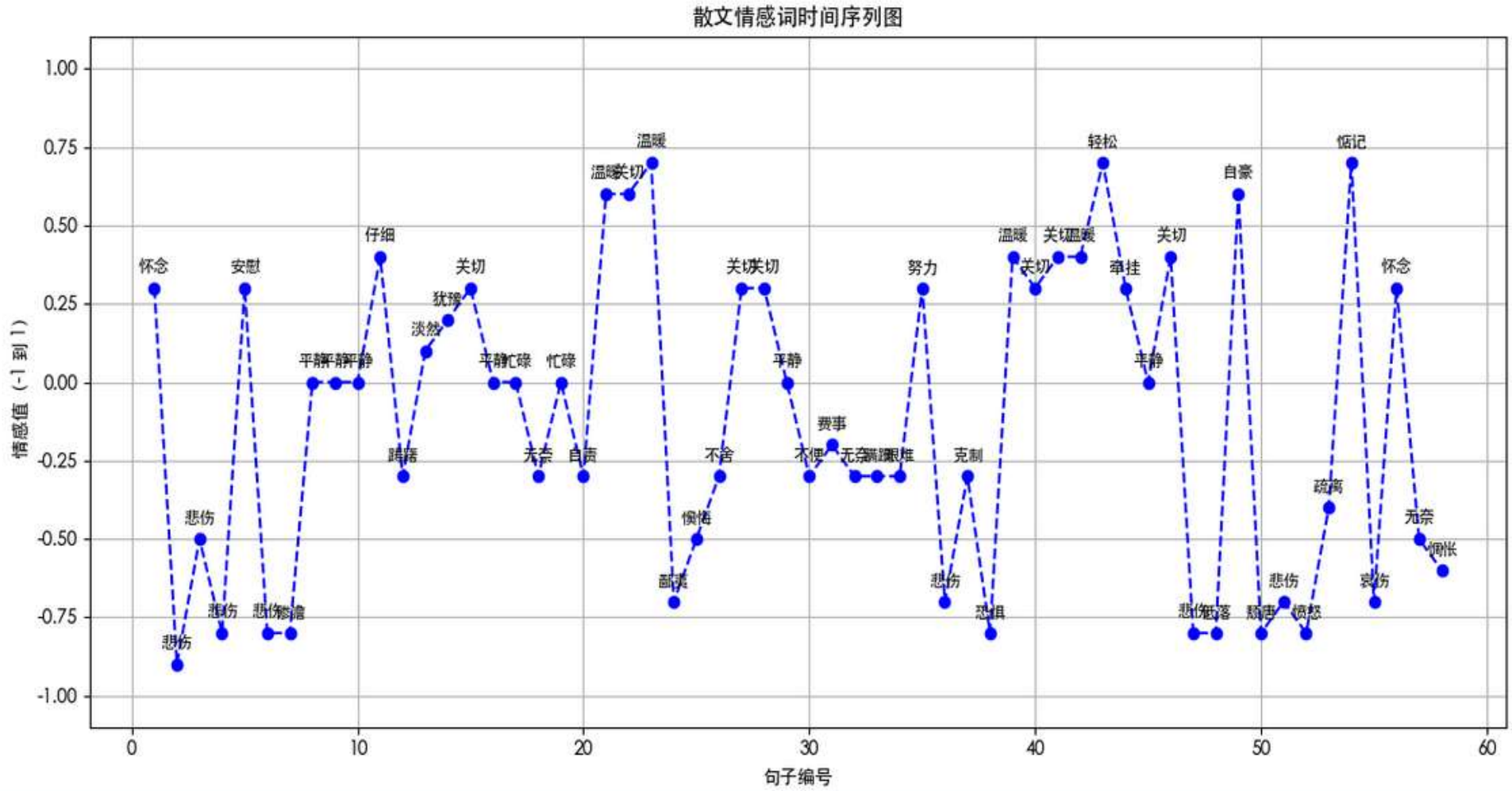
缺点：一、不能完全替代人类。二、模型可解释性不足：“模型可解释性”是医疗应用中的关键挑战，缺乏透明的决策过程会影响医生和患者的信任度。三、伦理与责任问题：医疗大语言模型面临"数据隐私保护"、"算法公平性"和"模型可解释性"等伦理挑战，需要严格遵循相关法律法规。

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？

答：大模型情感世界的理解还可以运用营销或者市场销售方面，通过采集顾客购买时的情绪以及评论和反馈中情绪的体现，可以得到顾客层面上的商品的优点与缺点，对于生产商和企业来说都具有重要的意义。

代码运行4

```
plt.title("散文情感词时间序列图")
plt.xlabel("句子编号")
plt.ylabel("情感值（-1 到 1）")
plt.ylim(-1.1, 1.1)
plt.grid(True)
plt.tight_layout()
plt.show()
```



代码运行4续

```
[5]: for item in emotion_results:
         print(f"句子{item[0]}: 情感词: {item[2]}, 情感值: {item[3]}")
```

```
句子1: 情感词: 怀念, 情感值: 0.3
句子2: 情感词: 悲伤, 情感值: -0.9
句子3: 情感词: 悲伤, 情感值: -0.5
句子4: 情感词: 悲伤, 情感值: -0.8
句子5: 情感词: 安慰, 情感值: 0.3
句子6: 情感词: 悲伤, 情感值: -0.8
句子7: 情感词: 惨淡, 情感值: -0.8
句子8: 情感词: 平静, 情感值: 0.0
句子9: 情感词: 平静, 情感值: 0.0
句子10: 情感词: 平静, 情感值: 0.0
句子11: 情感词: 仔细, 情感值: 0.4
句子12: 情感词: 踌躇, 情感值: -0.3
句子13: 情感词: 淡然, 情感值: 0.1
句子14: 情感词: 犹豫, 情感值: 0.2
句子15: 情感词: 关切, 情感值: 0.3
句子16: 情感词: 平静, 情感值: 0.0
句子17: 情感词: 忙碌, 情感值: 0.0
句子18: 情感词: 无奈, 情感值: -0.3
句子19: 情感词: 忙碌, 情感值: 0.0
句子20: 情感词: 自责, 情感值: -0.3
句子21: 情感词: 温暖, 情感值: 0.6
句子22: 情感词: 关切, 情感值: 0.6
句子23: 情感词: 温暖, 情感值: 0.7
```

简要做代码运行总结分析

答：该代码实现了对文本《背影》的情感分析：首先读取并按句号、感叹号分句，然后调用大语言模型API逐句分析情感词与情感值，提取结果后通过Matplotlib绘制情感值随句子变化的时间序列图，直观展示文本情感动态趋势。

南京工業大學
NANJING TECH UNIVERSITY

02 关联数据分析

吴志祥

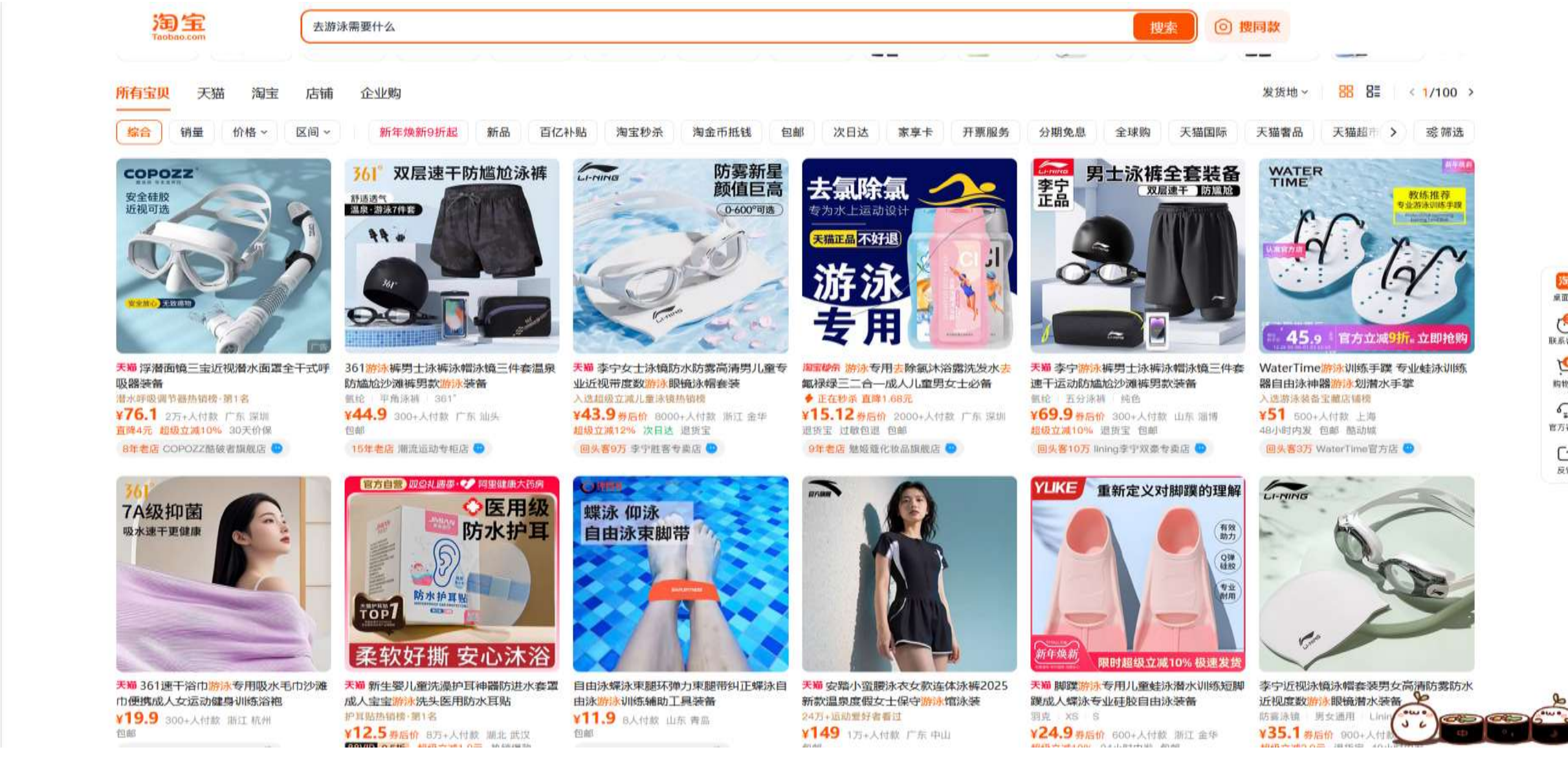
18205185639

cnwzx2012@njtech.edu.cn

南京工業大學
NANJING TECH UNIVERSITY

第六讲作业

第六讲1、淘宝实际产业案例（去游泳需要什么）分析



1、淘宝实际产业案例（去游泳需要什么）分析续

淘宝 Taobao.com

去游泳需要什么 | 搜索 | 搜同款

综合 | 销量 | 价格 | 区间 | 新年焕新9折起 | 新品 | 百亿补贴 | 淘宝秒杀 | 淘金币抵钱 | 包邮 | 次日达 | 家享卡 | 开票服务 | 分期免息 | 全球购 | 天猫国际 | 天猫奢品 | 天猫超市 | 筛选

天猫 游泳专用去氯沐浴露洗发水二合一儿童成人套装男女士专用除氯去盐
品牌新客补贴，当日有效
¥25.01 首单价 300+人付款 广东 珠海
直降1元 超级立减3元 淘金币已抵0.89元
9年老店 水理芙旗舰店

天猫 泳镜防水防雾高清男士有度数近视游泳眼镜泳帽成人女男款潜水装备
竞速泳镜 | 男女通用 | 捷虎
¥15.9 11人付款 山东 青岛
退货宝 包邮
回头客1万 捷佳吉运动专营店

对本次搜索体验满意吗
非常满意
满意
感觉一般
不满意
非常不满意
不感兴趣，关闭

天猫 李宁泳裤男款平角专业游泳装备男全套游泳裤防尴尬泳衣套装成人
氨纶 | 条纹 | 抽绳
¥25.9 券后价 1000+人付款 山东 淄博
超级立减10% 退货宝 包邮
回头客10万 lining李宁双豪专卖店

天猫 游泳专用去氯洗发水沐浴露二合一儿童男女士除氟绿绿洗护必备用品
品牌新客补贴，当日有效
¥15.3 首单价 500+人付款 广东 深圳
直降1元 淘金币已抵0.5元 退货宝
9年老店 魅姬蔻化妆品旗舰店

天猫 新年焕新 361孙杨同款游泳裤男士泳裤泳帽泳镜三件套防尴尬泡温泉裤男装备
694万+运动爱好者看过
¥37.9 券后价 20万+人付款 浙江 杭州
超级立减10元 热销爆款 次日达
回头客50万 361度游泳旗舰店

户外换衣服遮挡神器男士可穿吸水洗澡游泳海边围裙裙裤无档夏季棉
1000mm(含)-1500mm(不含) 需要搭建
¥12.93 43人付款 河北 石家庄
店铺满15减1 退货宝 包邮
清水创意小铺

天猫 脚蹼游泳成人自由泳蛙泳硅胶短脚蹼男女儿童专业轻便潜水训练蛙鞋
羽克 | XS | S
¥39.9 100+人付款 浙江 金华
包邮
回头客10万 来鸟运动专营店

天猫 李宁泳裤男士平角五分泳裤速干泳镜泳帽三件套防尴尬专业游泳套装
聚酯纤维 | 条纹 | 抽绳
¥29.9 800+人付款 山东 淄博
退货宝 包邮
回头客10万 lining李宁双豪专卖店

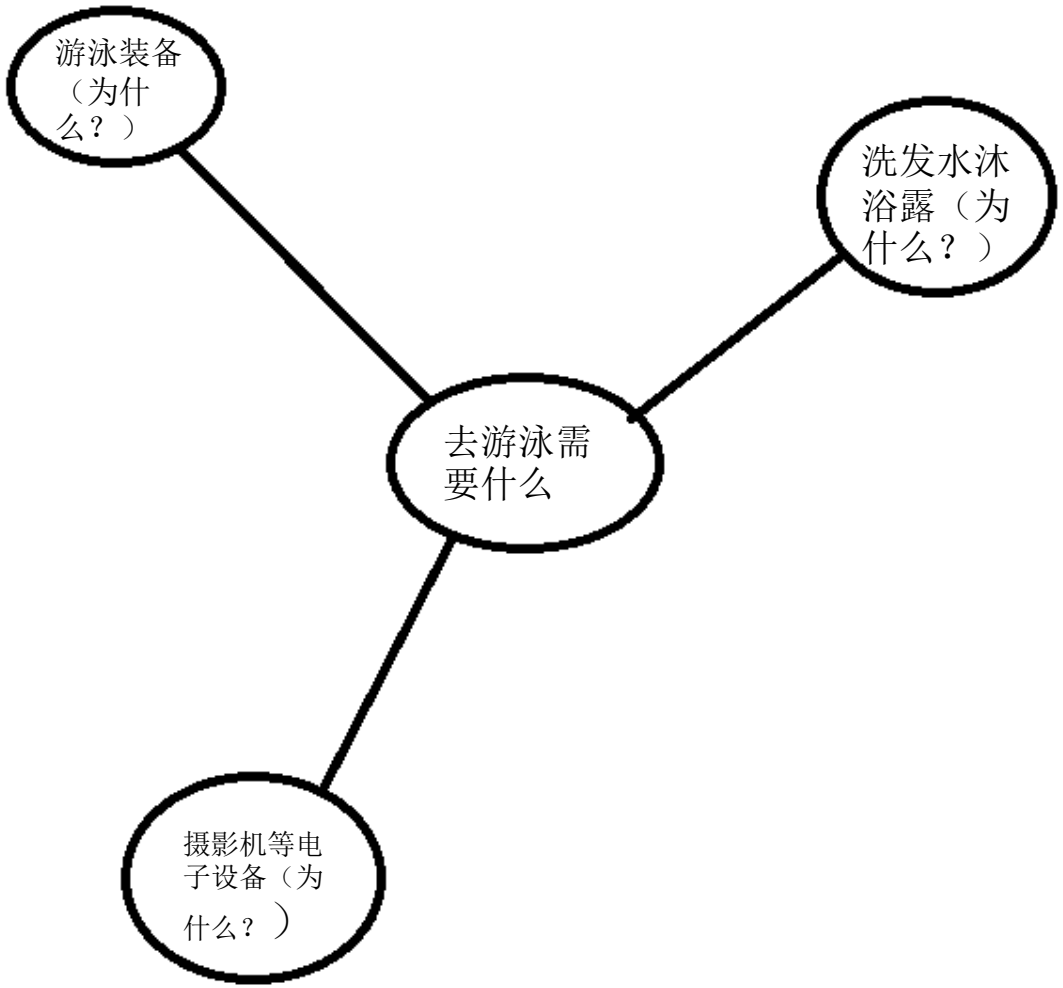
天猫 李宁泳衣泳裤套装泳帽泳镜三件套男士防尴尬宽松双层游泳装备新款
氨纶 | 五分泳裤 | 纯色
¥69.9 券后价 800+人付款 山东 淄博
超级立减15% 退货宝 包邮
回头客10万 lining李宁双豪专卖店

天猫 运动相机vlgo高清4K头盔骑行摄像水下摩托车自行车记录仪胸前固定
品牌新客补贴，当日有效
¥356.02 首单价 1000+人付款 广东
直降4元 超级立减20元 热销爆款
FETANA旗舰店

夏季男士野室外更衣神器沙滩户外游泳衣遮挡布换衣罩可裹可穿浴巾
1000mm(含)-1500mm(不含) 需要搭建
¥3.47 券后价 52人付款 江西 赣州
店铺满6减5 包邮
佛克斯精品百货

桌面版 | 联系客服 | 购物车 | 官方客服 | 反馈 | 回顶部

第六讲1、淘宝实际产业案例（去游泳需要什么）分析续



当我输入去游泳需要什么时，淘宝的“阿里商品大脑”智能导购先会通过“游泳”关键字进行检索，然后会联想到游泳需要合适的装备，例如泳衣、泳裤以及潜水镜等，同时也会考虑到用户游泳过程中可能会有记录游泳过程或者拍摄水下的需求，于是会推荐摄影机等电子设备，也会考虑到用户游完泳后的清洁需求，于是推送洗发水沐浴露等清洁物品。大致分为以下三个方面：

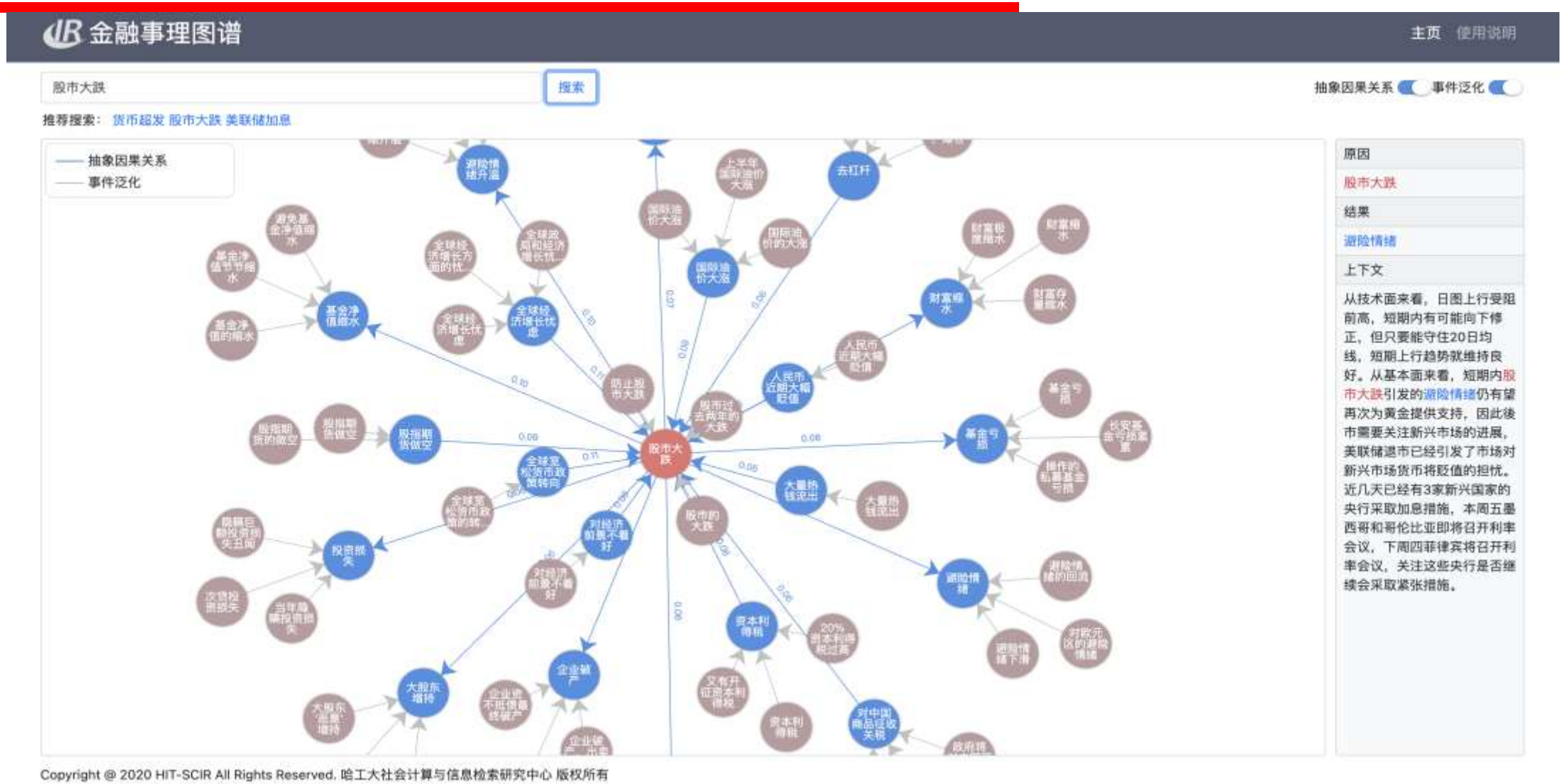
一、游泳前，用户游泳的装备：泳衣、泳裤以及潜水镜。

二、游泳中，用户可能会具有记录和拍摄以及查看时间的需求：摄像机、防水手表等。

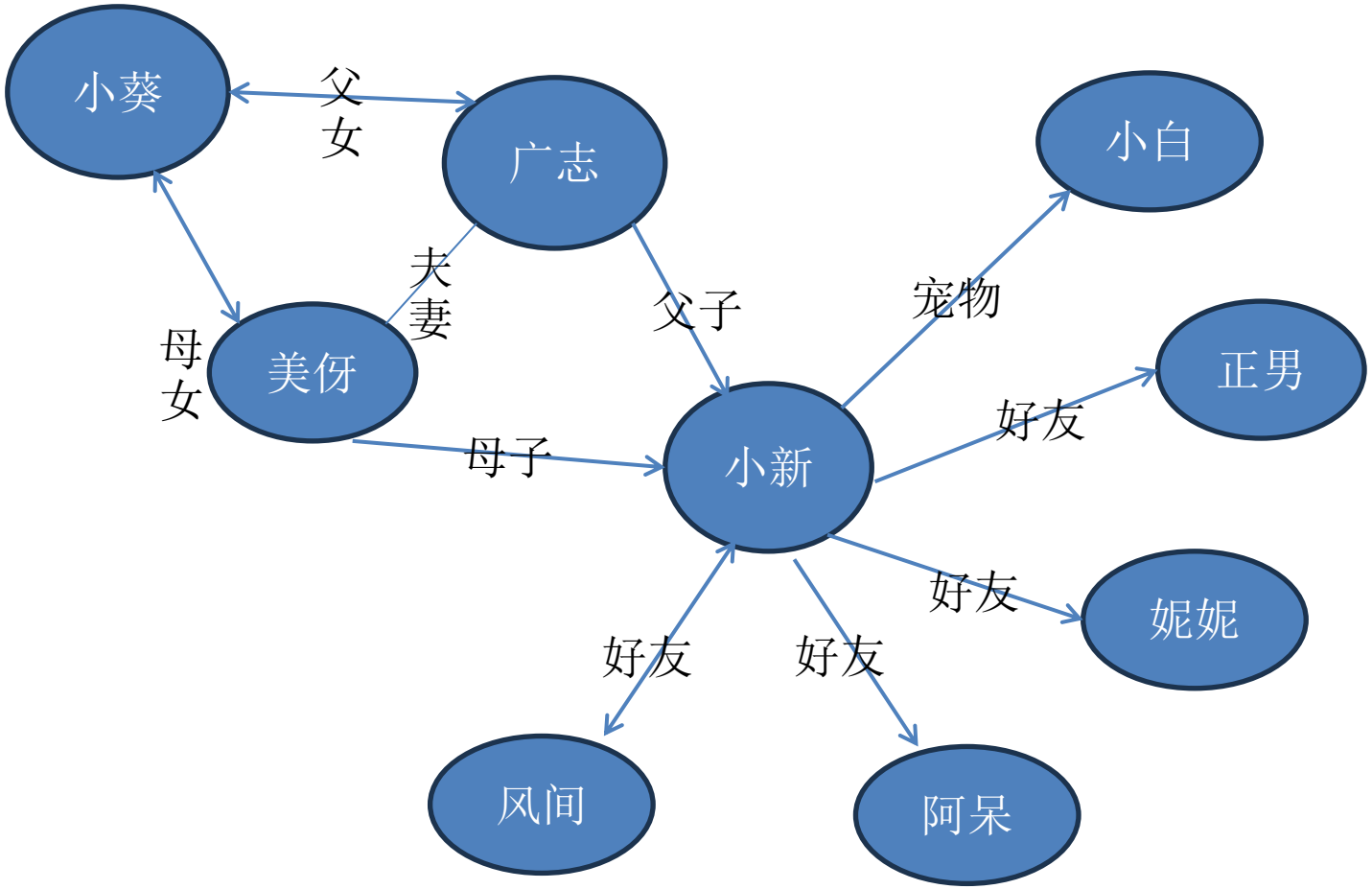
三、游泳后，用户清洁身体的需求：洗发水沐浴露等清洁必需品。

总结，“阿里商品大脑”智能推送大致是通过分析关键字以及分阶段理解用户需求来推荐商品。

第六讲（2）1、使用PPT中知识图谱链接平台(哈尔滨工业大学-金融事理图谱http://eeg.8wss.com/)



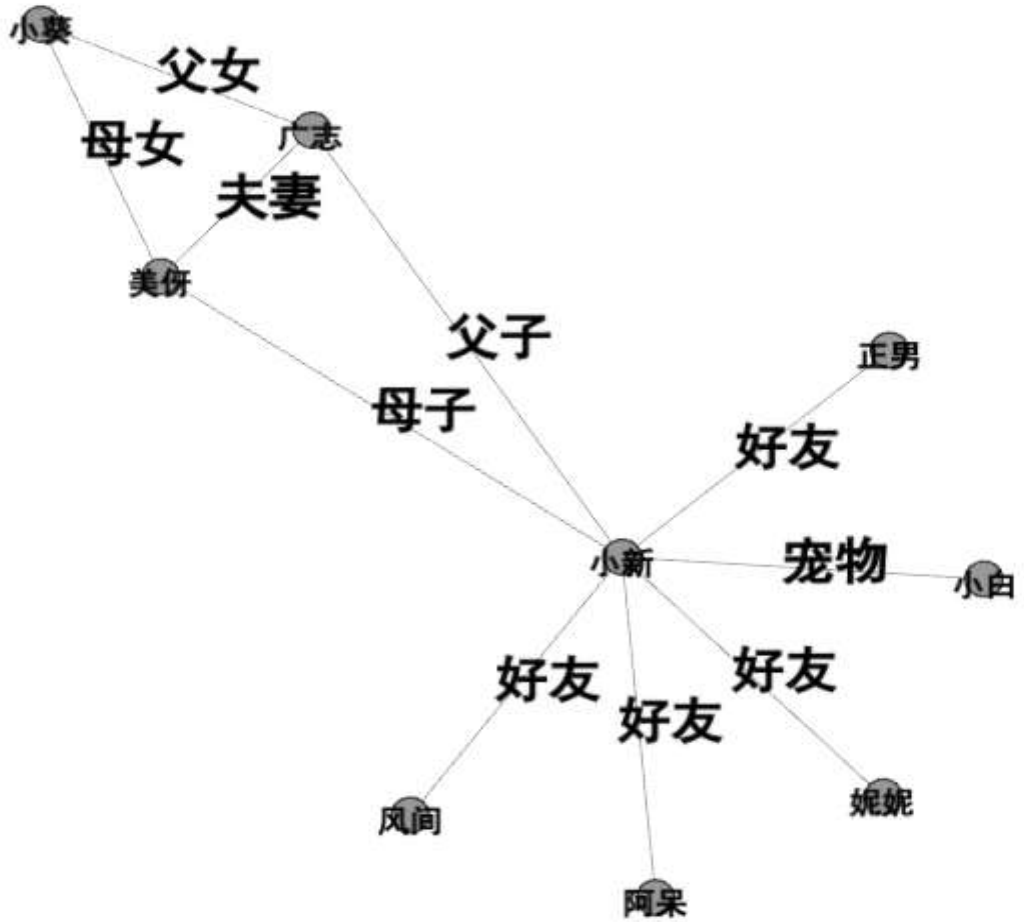
第六讲（2）2、使用白板建模绘制《蜡笔小新》人物关系“知识图谱”



解释：《蜡笔小新》主要讲述了小新一家的温馨搞笑日常故事，本图谱的绘制主要通过作品中的人物关系进行绘制，绘制了主要人物的关系

类别	关系描述
核心角色	小新 —— 故事主角，所有关系围绕他展开
家庭成员	父亲（广志）、母亲（美伢）、姐姐（小葵）——构成完整家庭结构
宠物	小白 —— 忠诚的伙伴，情感寄托
朋友圈	正男、妮妮、风间、阿呆 —— 表现儿童社交与成长

第六讲（2）3、使用echarts中的关系图，绘制作业2《蜡笔小新》中的“知识图谱”。



第六讲（2）4、使用Neo4j（可在线版本），编程绘制一款（简单）知识图谱