用户数据采集与关联分析（结课作业）

程一娇 信管2301

目录

Contents

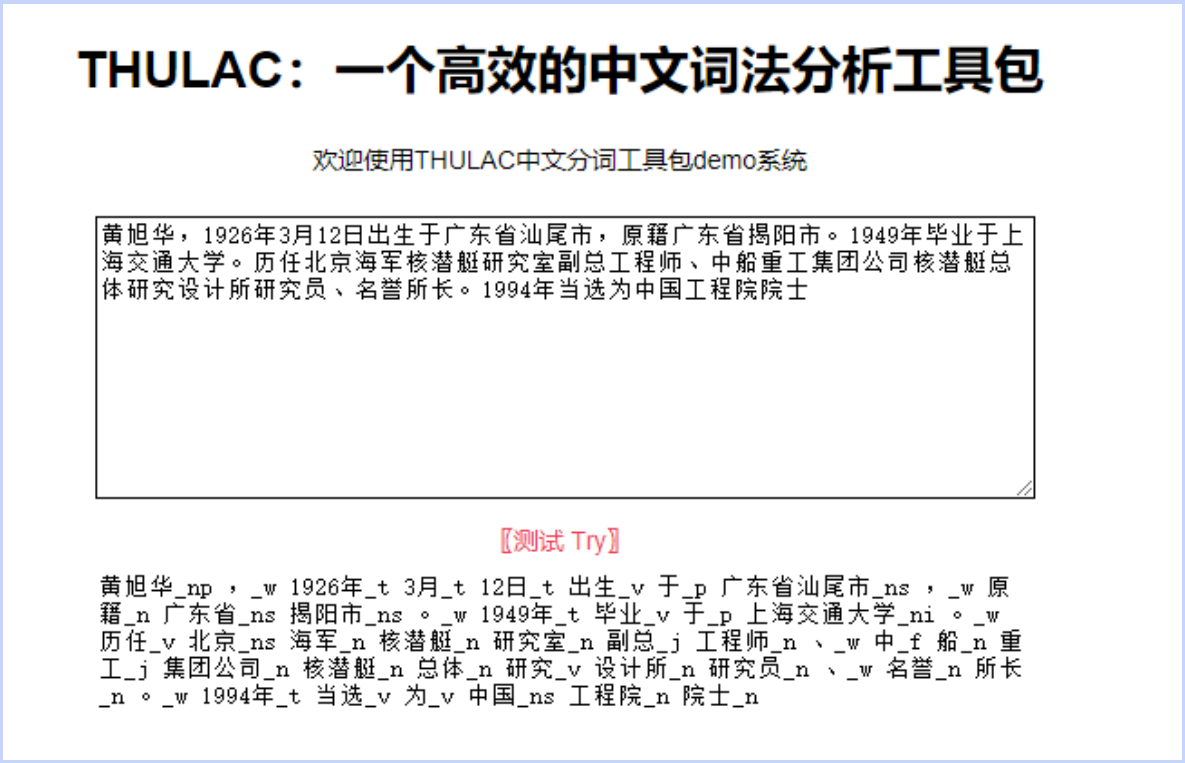
感知世界：课程导言与分词

01

1、使用在线NLPIR分词系统演示分词

分词系统简介

介绍了NLPIR、微词云、清华分词三大在线分词系统，它们在学术文本、短文本、批量处理场景下各有优势。NLPIR适合学术文本，微词云适合短文本，清华分词适合批量处理。



2 、 Anaconda环境搭建与Hello World

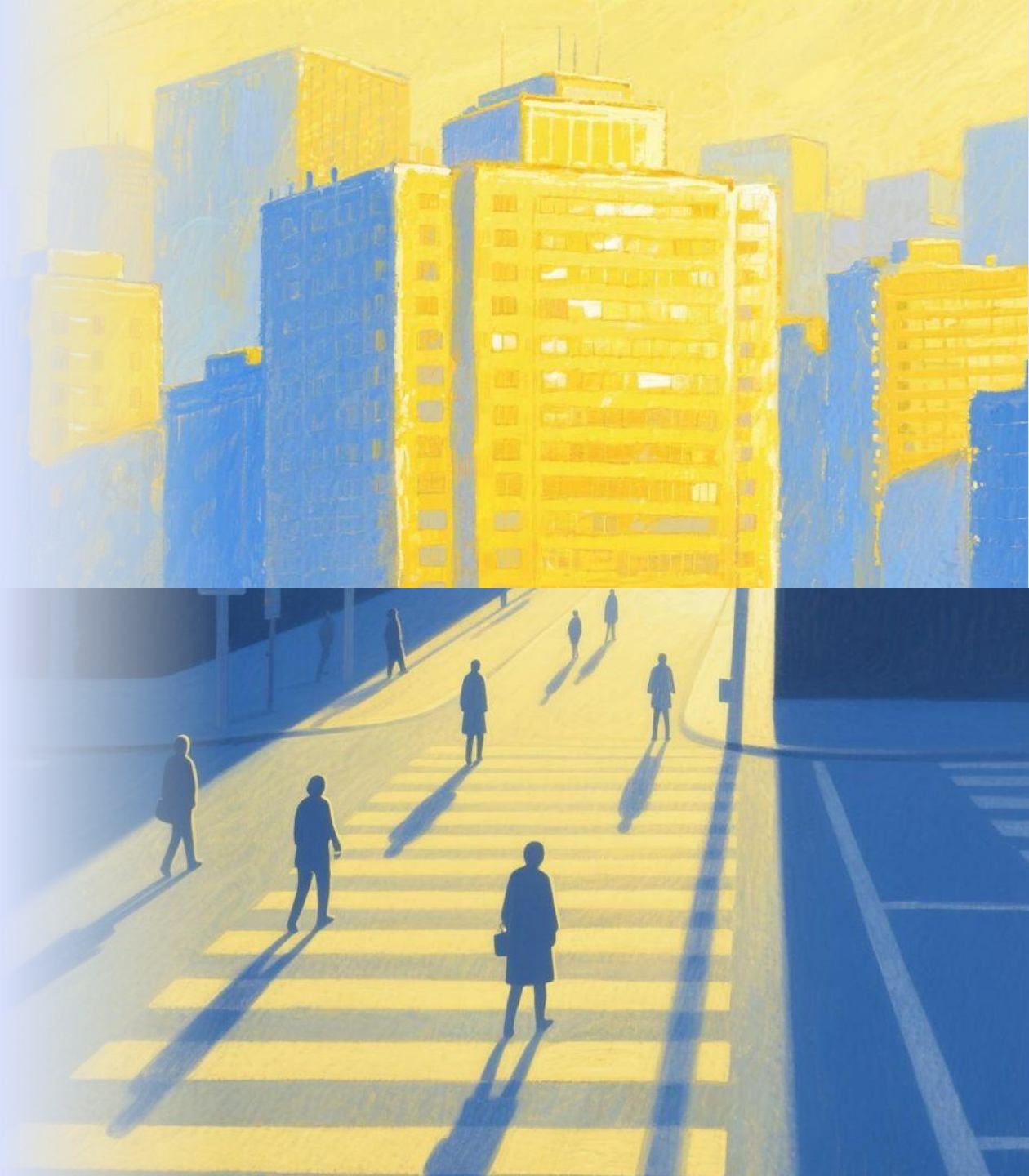
Anaconda环境搭建

介绍了Anaconda的下载、安装、虚拟环境创建和Jupyter启动的完整流程。强调了选择Python3.x的重要性以及勾选“Add to PATH”的时机，确保环境搭建顺利。

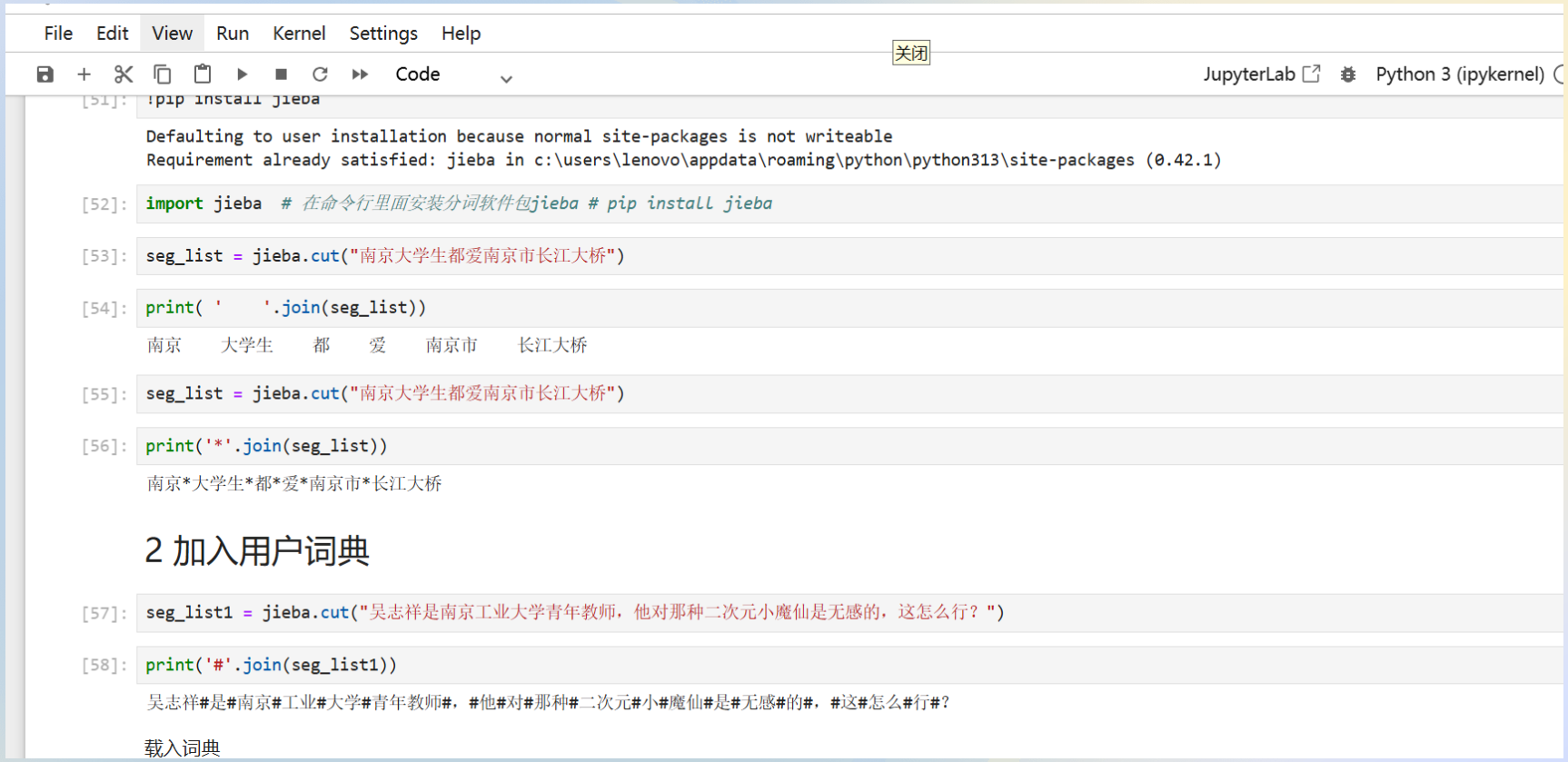
Hello World代码示例

```
[3]: print("Hello World. Hello “程一娇”")

     Hello World. Hello “程一娇”
```



3、课后作业001-004运行



4、基于关键词的学术文本聚类集成研究_张颖怡

研究背景与问题

学术文献数量快速增长，如何高效、自动地进行学科分类成为研究热点。
传统聚类方法（如K-means、增量聚类）存在性能瓶颈，聚类集成方法被提出以提升分类效果。
本文重点探讨：
聚类集成方法能否提升文本聚类性能?
关键词抽取方法对聚类结果有何影响?
关键词数量如何影响聚类性能?

研究方法

数据集：采用ACM计算机学科分类体系中的8个子数据集，涵盖40个研究领域。
关键词抽取：比较四种无监督方法：
TF-ISF、CSI、ECC、TextRank
抽取关键词数量：5, 10, ..., 60个
聚类方法：
基础聚类：K-means、增量聚类
聚类集成：ECKM（基于K-means）、ECIC（基于增量聚类）
评估指标：使用F1值评估聚类性能。

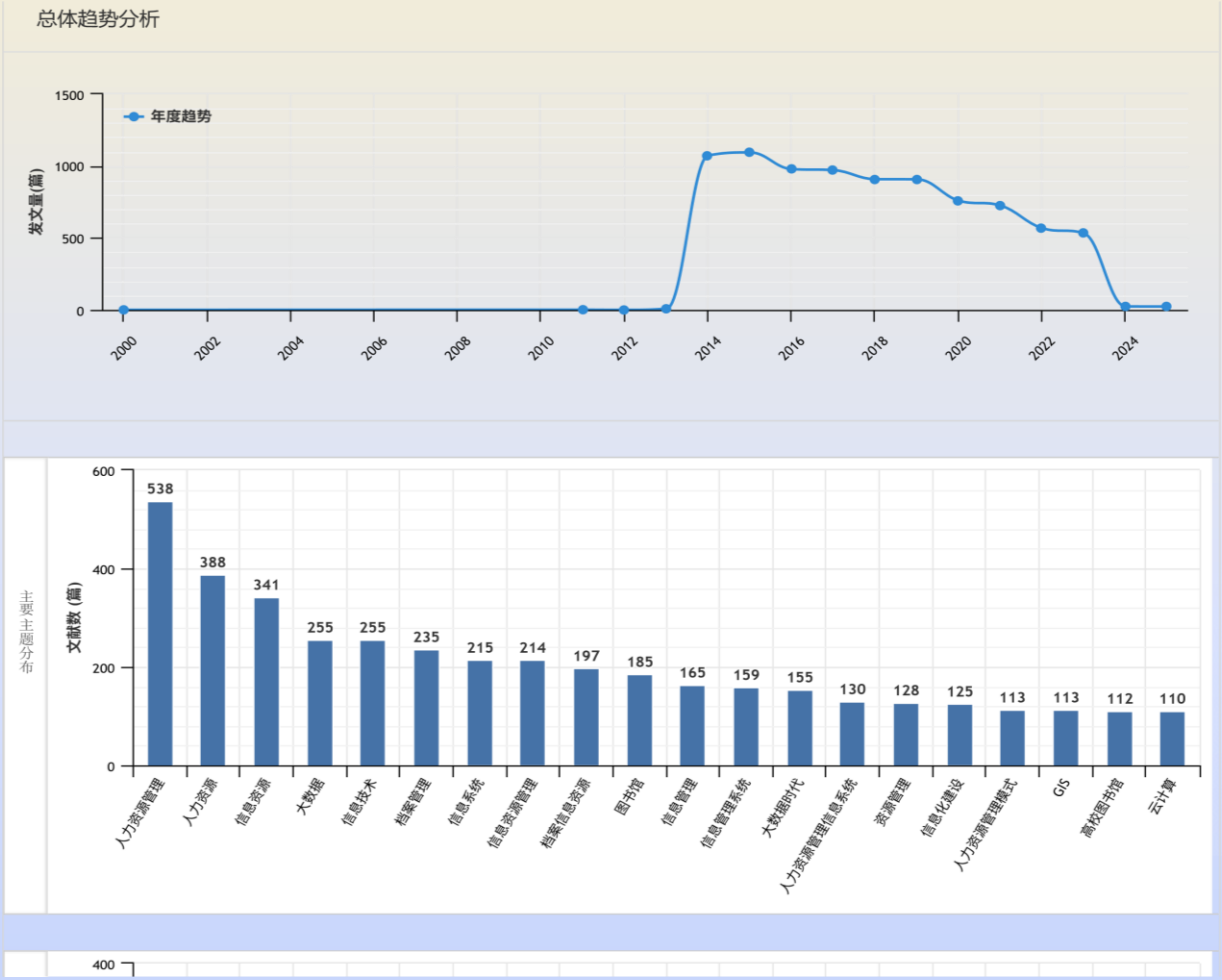
主要结论

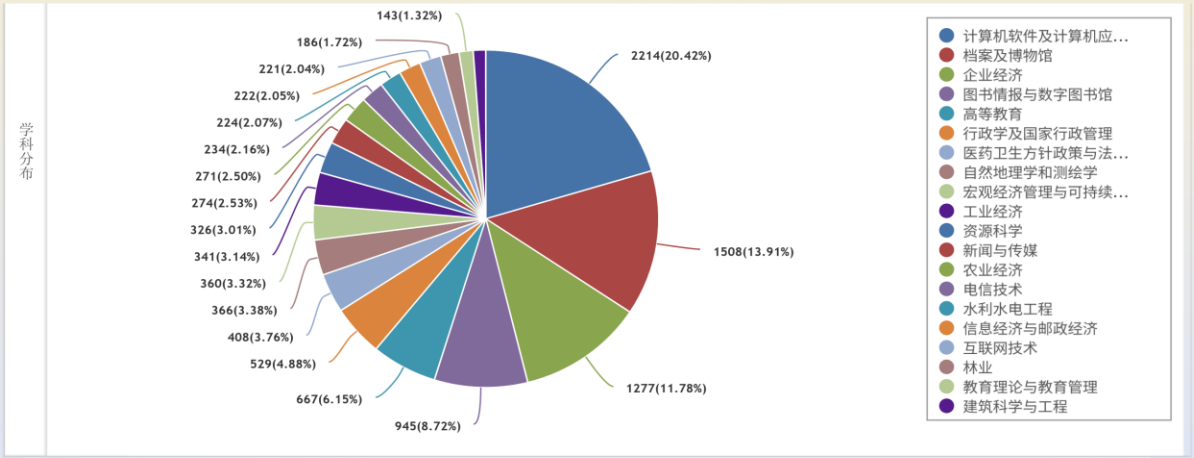
✅ 聚类集成显著提升性能：ECKM和ECIC在多数数据集上优于基础聚类方法，尤其在学科类别差异大的数据中表现更佳。
✅ 关键词抽取方法影响显著：TextRank表现最佳，CSI表现较差。
✅ 关键词数量越多，性能越好：随着关键词数量增加，聚类性能整体提升。
✅ 聚类集成对关键词数量的鲁棒性更强：尤其在关键词较少时，集成方法仍能保持较高性能。

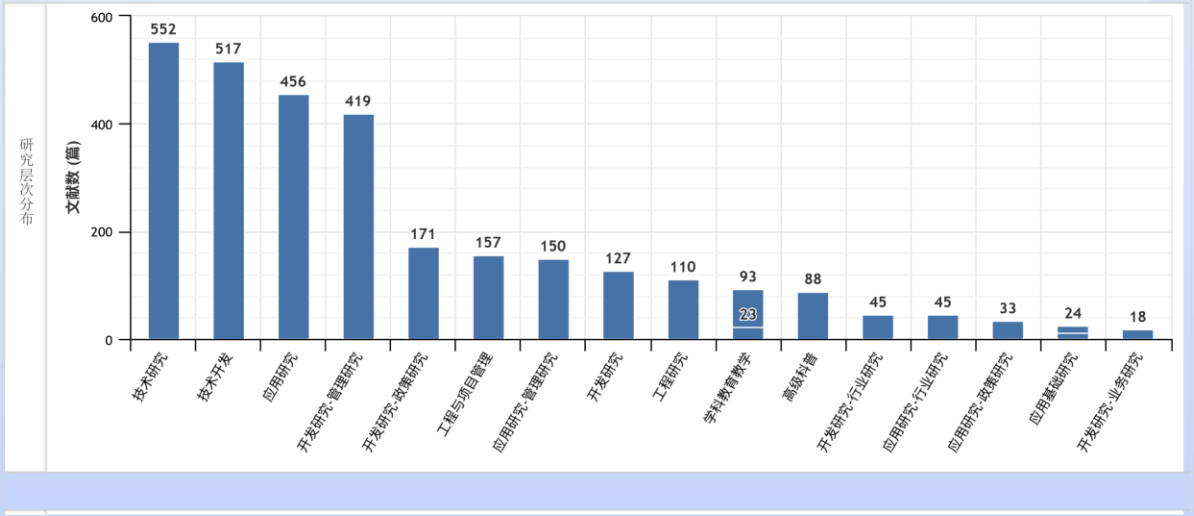
感知世界：词频统计与分析

02

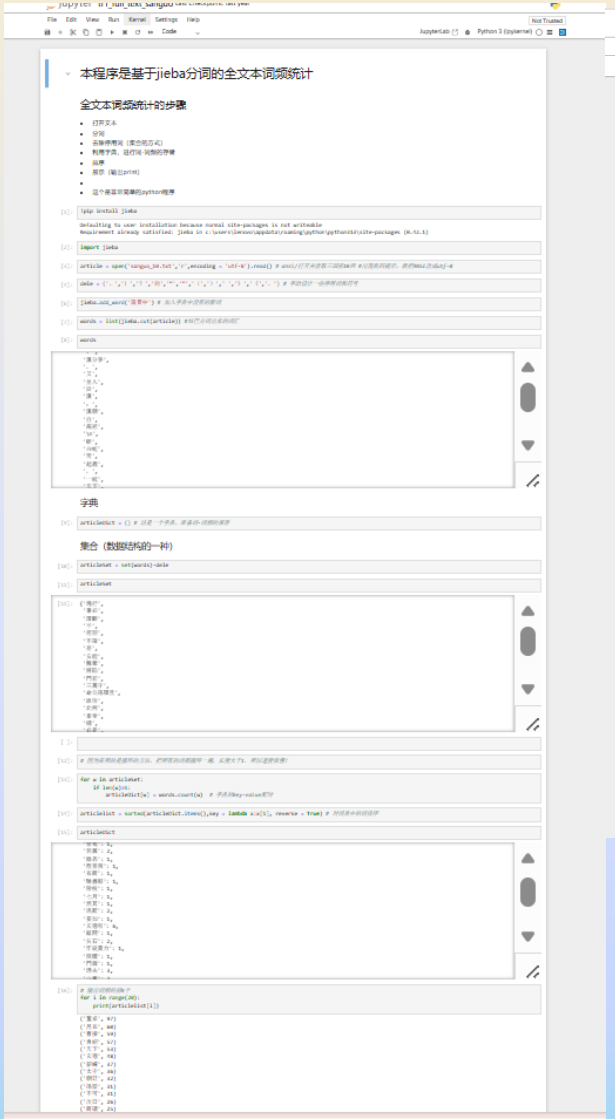
1、信息资源管理十年主题演化

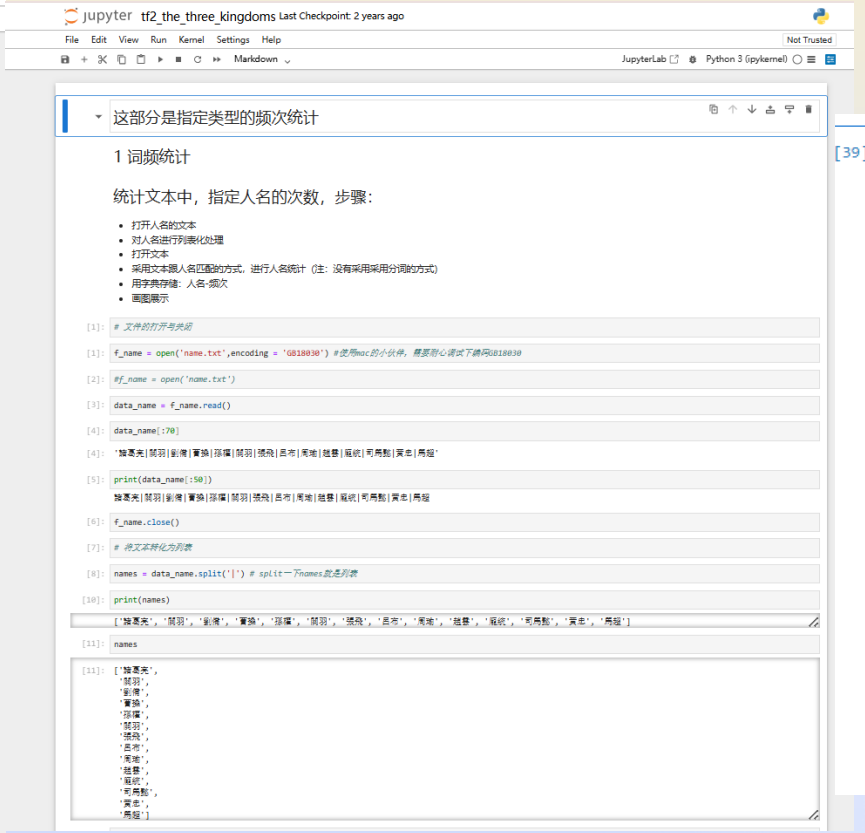


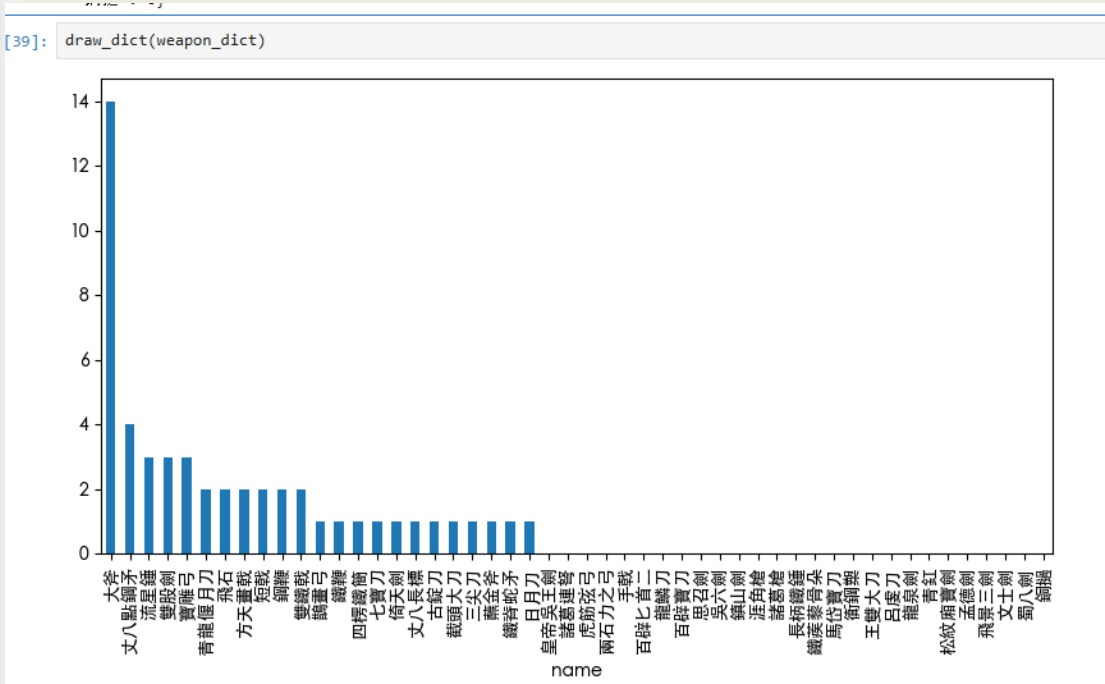




2、完成ppt中的程序运行，包括全文词频统计，指定类型词频统计







3、黄旭华传记文本词频透视

词频统计结果

黄旭华传记文本中，核潜艇、采集、学术、资料、工作等词汇位列词频前五，反映了传记的核心内容和主题。

单词	词性	次数	条数	词频	TF-IDF
核潜艇	名词	32	9	0.029331	0.017659
采集	动词	29	6	0.026581	0.020121
学术	名词	22	8	0.020165	0.013063
资料	名词	21	5	0.019248	0.015859
工作	名动词	15	7	0.013749	0.00961
成长	动词	14	8	0.012832	0.008313
小组	名词	14	6	0.012832	0.009714
院士	名词	13	7	0.011916	0.008329
进行	动词	13	8	0.011916	0.007719
研制	名动词	12	6	0.010999	0.008326
工程	名词	11	8	0.010082	0.006532
访谈	动词	10	4	0.009166	0.008278
专业	名词	9	5	0.008249	0.006797
技术	名词	9	6	0.008249	0.006244
科学	名词	8	6	0.007333	0.005551
介绍	动词	8	6	0.007333	0.005551
思想	名词	7	5	0.006416	0.005286
传记	名词	7	3	0.006416	0.006416
人生	名词	7	6	0.006416	0.004857
研究	名动词	6	4	0.0055	0.004967
大学	名词	6	4	0.0055	0.004967
过程	名词	6	6	0.0055	0.004163
传主	动词	6	2	0.0055	0.006187
要求	动词	6	4	0.0055	0.004967
时间	名词	6	5	0.0055	0.004531
历史	名词	6	3	0.0055	0.0055
成就	名词	6	4	0.0055	0.004967

```
[15]: # 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)
('进行', 13)
('院士', 13)
('专业', 13)
('我国', 12)
('研制', 12)
('技术', 12)
('工程', 11)
('访谈', 10)
('介绍', 8)
('科学', 8)
('第一代', 8)
('主要', 8)
('历史', 7)
('传记', 7)
('思想', 7)
('及其', 7)
('人生', 7)
('一生', 6)
('过程', 6)
('设计', 6)
('传主', 6)
('按照', 6)
('成就', 6)
('研究', 6)
('要求', 6)
('实物', 5)
('先后', 5)
('求学', 5)
('精神', 5)
('实现', 5)
('重点', 5)
('黄旭', 5)
('重要', 4)
('事件', 4)
('保密', 4)
('其中', 4)
('照片', 4)
```

4、Long live the scientists Tracking the scientific 阅读总结

研究问题：科学家如何“活”在书里?

通过提问“牛顿与爱因斯坦谁更常被提起？”引入研究问题，探讨科学家声望在时间长河中的演变。

方法：用“名字被提及的次数”丈量声望

谷歌图书3600万册、学术文献9100万篇的宏大语料，为研究提供了丰富的数据基础，确保结果的可靠性。

主要发现

伟大的头脑从未被遗忘

即使已逝世几个世纪，牛顿的名字在书籍中依然活跃。科学家的物理生命有限，但其知识影响可延续数百年。

“自己人”偏好明显

牛顿在英文书籍（尤其是英式英语）中始终更受青睐；而爱因斯坦在德文和美式英语书籍中影响力更大。科学声望也逃不开“地域亲近性”。

他们因何而被记住?

牛顿：万有引力定律（15.2%）、微积分（7.9%）是最常被关联的贡献。

爱因斯坦：相对论（28.1%）和量子理论（16.9%）构成了他声望的核心。

但值得注意的是，仍有大量书籍谈论他们的哲学思想、生平轶事——科学家的影响力早已溢出纯学术范畴。

物理学界的“声望排行榜”

基于21世纪书籍提及频次，排名前五的物理学家依次是：

爱因斯坦 > 普朗克 > 牛顿 > 帕斯卡 > 伽利略。

霍金位列第六，其公众影响力甚至超过法拉第等经典人物。

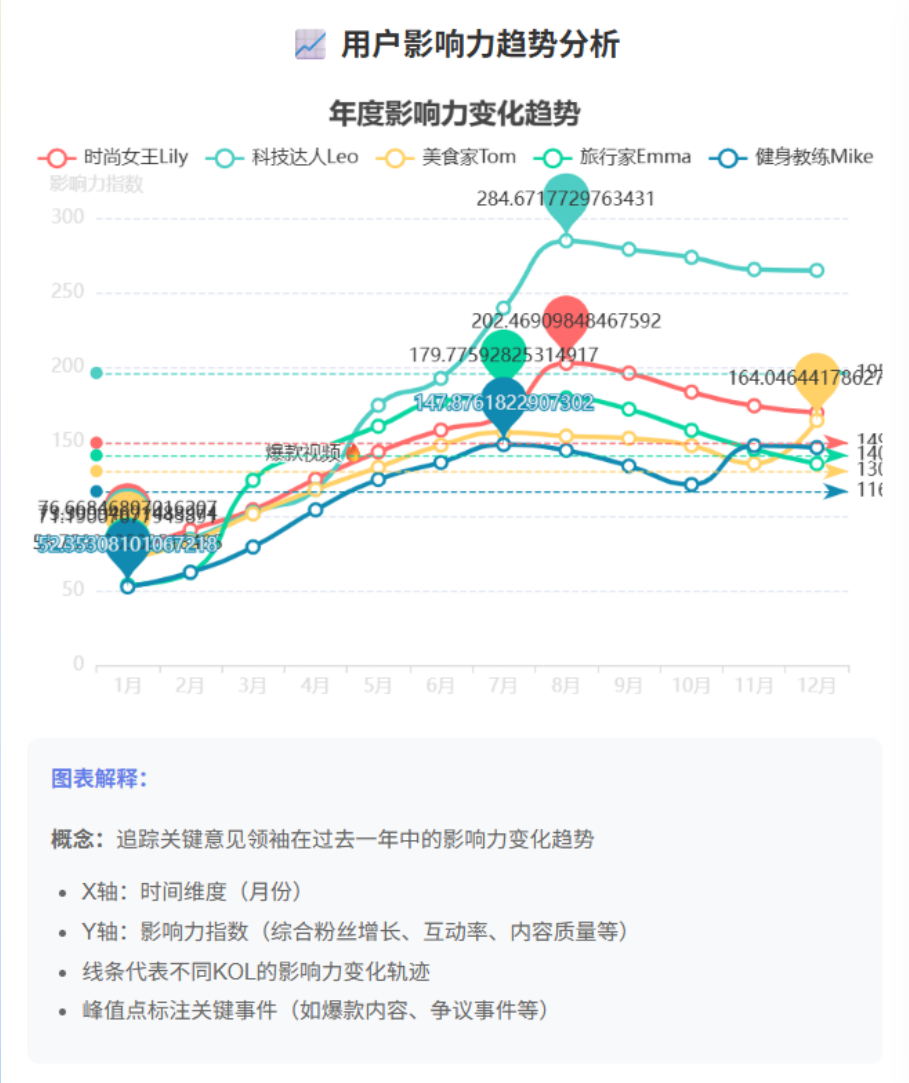
感知世界：词云与可视化

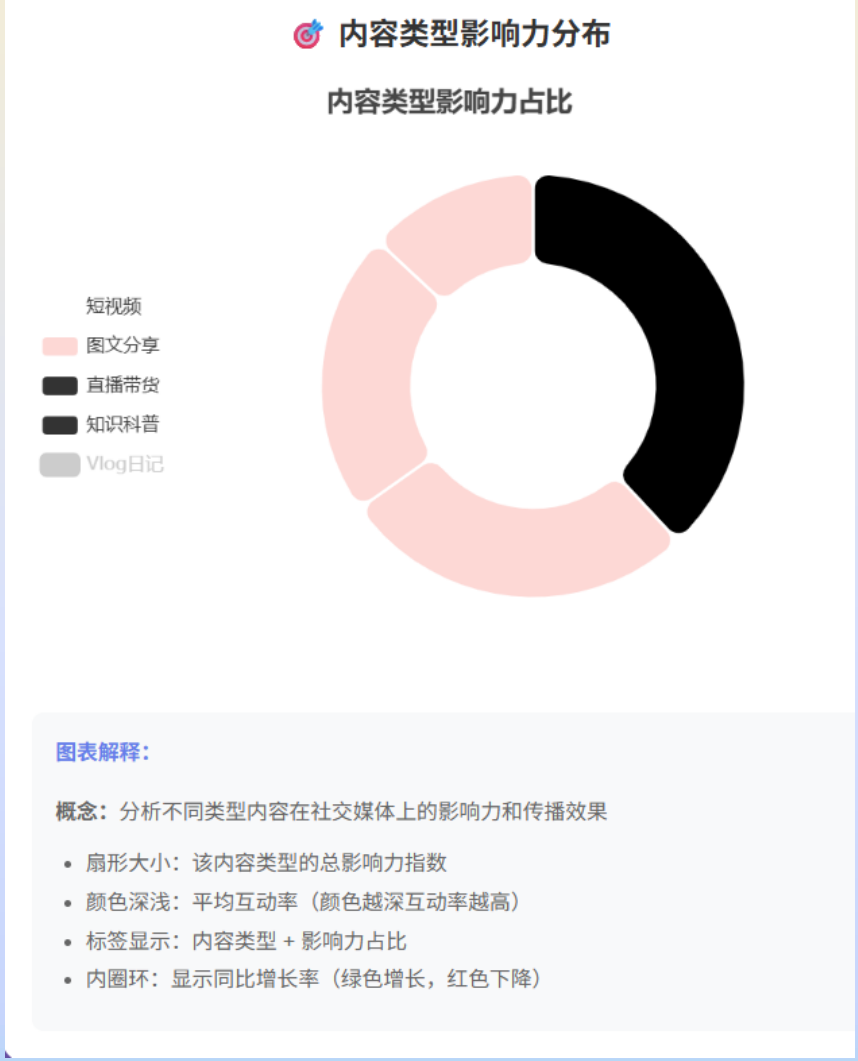
03

1、词云美学与释义写作

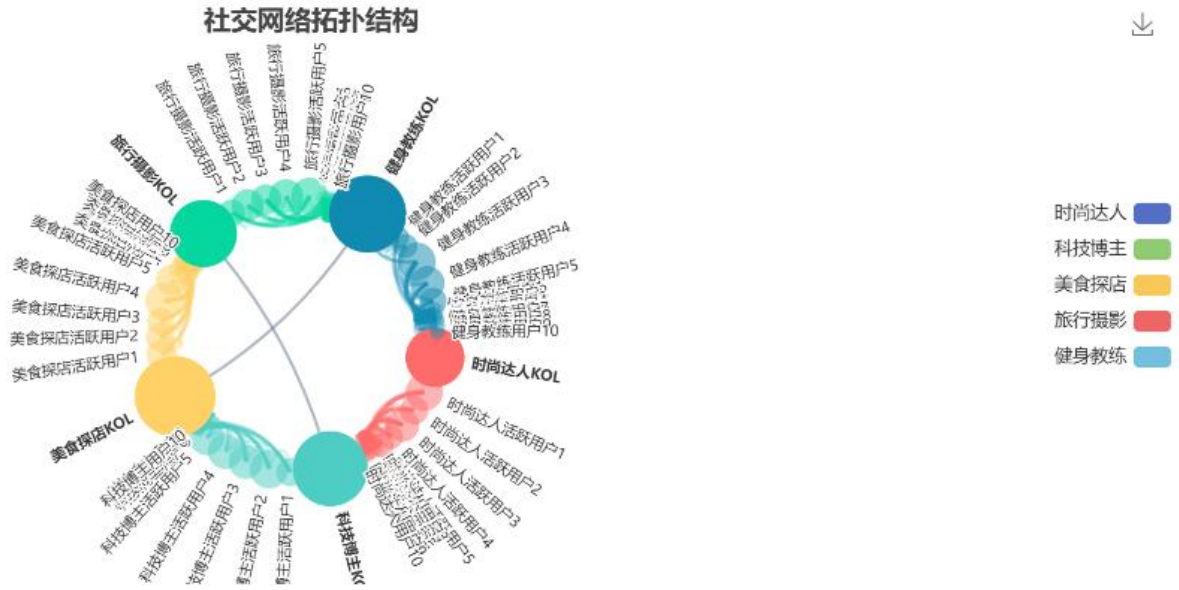
这个词云图聚焦美国的贸易政策，以“美国”“贸易”“关税”“贸易战”为核心，展现了特朗普政府加征关税、实施制裁等行为，不仅引发国际争端、遭到多国反对与批评，也对全球经济、相关国家（如中国、日本）及美国国内领域造成了多方面影响，同时体现出其政策被认为破坏规则、带有保护主义色彩的争议性。

2、ECharts关系图概念化设计





2、ECharts关系图概念化设计

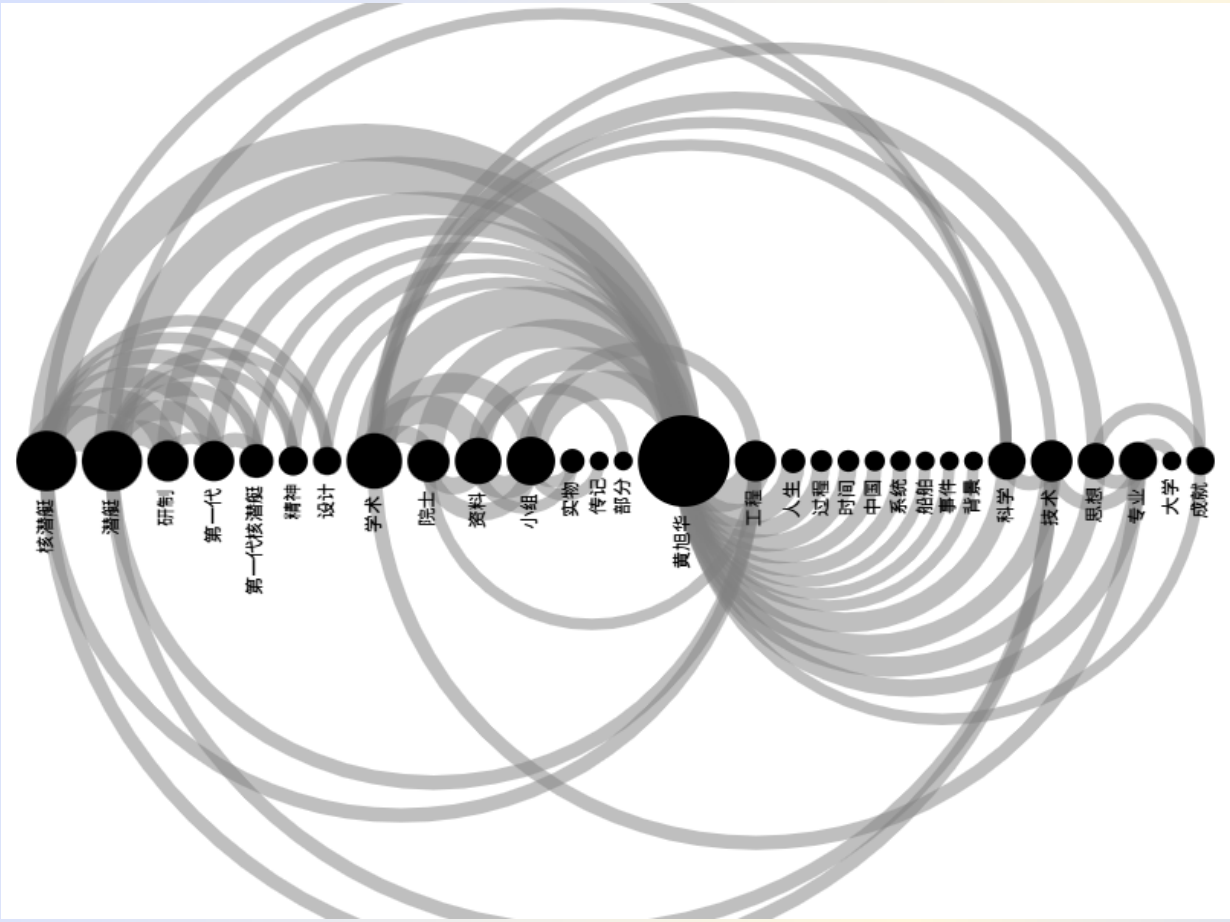


图表解释：

概念：展示社交媒体意见领袖与其粉丝之间的互动关系和影响力传播路径

- **节点说明：**
  - 大节点：意见领袖（KOL），节点大小代表粉丝数量
  - 中节点：活跃用户，参与度较高的普通用户
  - 小节点：普通用户，节点大小代表互动频率
- **连接线说明：**
  - 实线：双向关注关系
  - 虚线：单向关注关系
  - 线条粗细：互动频率和强度

3、使用RAWGraphs绘制图谱



4、采用给的程序，实现一段科学家文本的词云图绘制

感知世界：词云与可视化

04

1、对文本情感进行分析

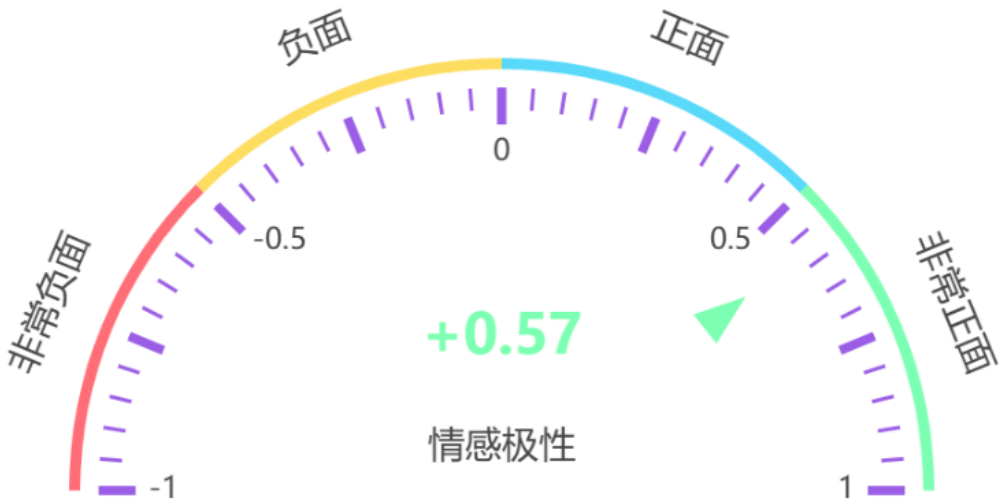
请输入一段中文文本：

我与父亲不相见已二年余，最不能忘记的是他的背影。
那年冬天，祖母去世，父亲失业，我们一同回家办丧事。丧事完毕，父亲送我回北京念书。在浦口车站，他忙着照看行李，又坚持要穿过铁道，去那边的月台给我买橘子。
我看见他穿着黑布大马褂，蹒跚地走到铁道边，慢慢探身下去。他用两手攀着上面，两脚再向上缩，肥胖的身子向左微倾，显出努力的样子。这时我看见他的背影，我的泪很快地流下来了。
等他回来，将橘子放在我的大衣上，嘱咐我路上小心，便离开了。我望着他的背影混入人群，眼泪又来了。
近年来父亲老境颓唐。最近他来信说身体平安，只是膀子疼痛，大去之期不远矣。我读到此处，在泪光中，又看见那肥胖的、青布棉袍黑布马褂的背影。唉！我不知何时再能与他相见！

314/1000

情感分析

情感极性



2、（1）sentiment_analysis_1运行

从淘宝或者其他平台上找一些评论的句子来试一试，它的情感值是多少。

[37]:
```
# 例如
text_taobao_1 = "多次购买了，这个黄桃燕麦酸奶味道不错，日期也很新鲜，值得购买"
```

[38]:
```
taobao_1 = SnowNLP(text_taobao_1)
```

[39]:
```
taobao_1.sentiments
```

[39]:
```
0.9977383610796604
```

[44]:
```
text_taobao_2 = "这个酸奶味道太酸，而且口感不丝滑，怪怪的，不好喝"
```

[45]:
```
taobao_2 = SnowNLP(text_taobao_2)
```

[46]:
```
for sentence in taobao_2.sentences:
    print(sentence)
```

```
这个酸奶味道太酸
而且口感不丝滑
怪怪的
不好喝
```

[47]:
```
taobao_2.sentiments
```

[47]:
```
0.985551898059871
```

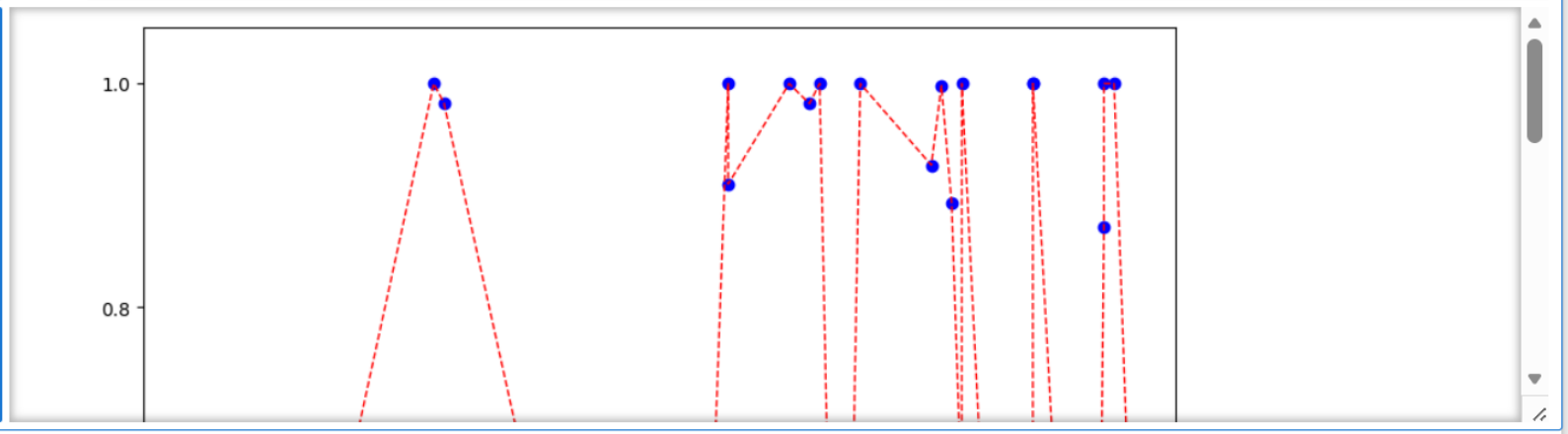
（2）sentiment_analysis_2运行

jupyter sentiment_analysis_2_timeline Last Checkpoint: 2 years ago

File Edit View Run Kernel Settings Help Not Trusted

Code JupyterLab Python 3 (ipykernel)

[18]:
```
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()
```



[16]:
```
plt.savefig('timeline.png') # 看不到？改一改？
```

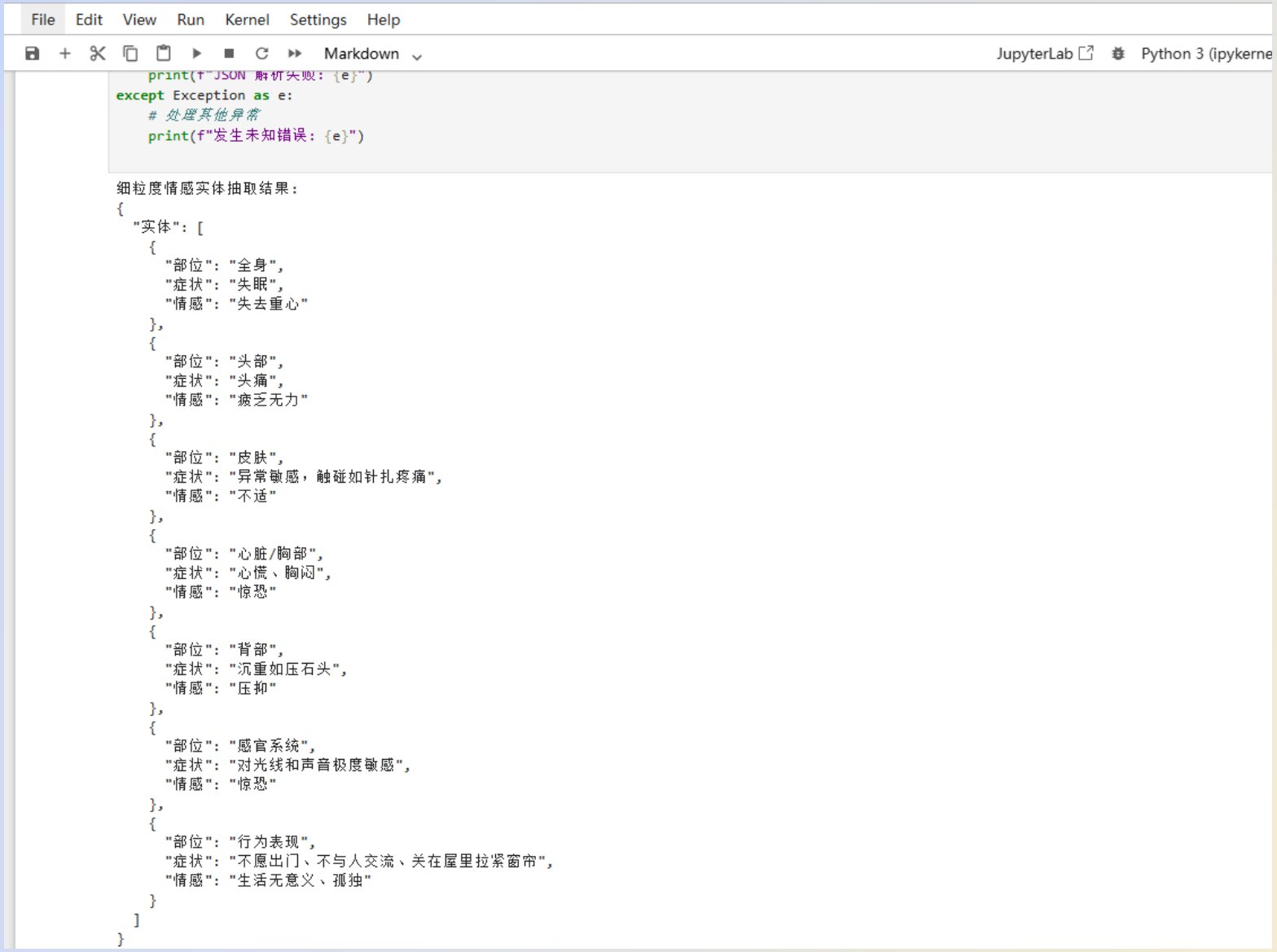
```
<Figure size 640x480 with 0 Axes>
```

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：

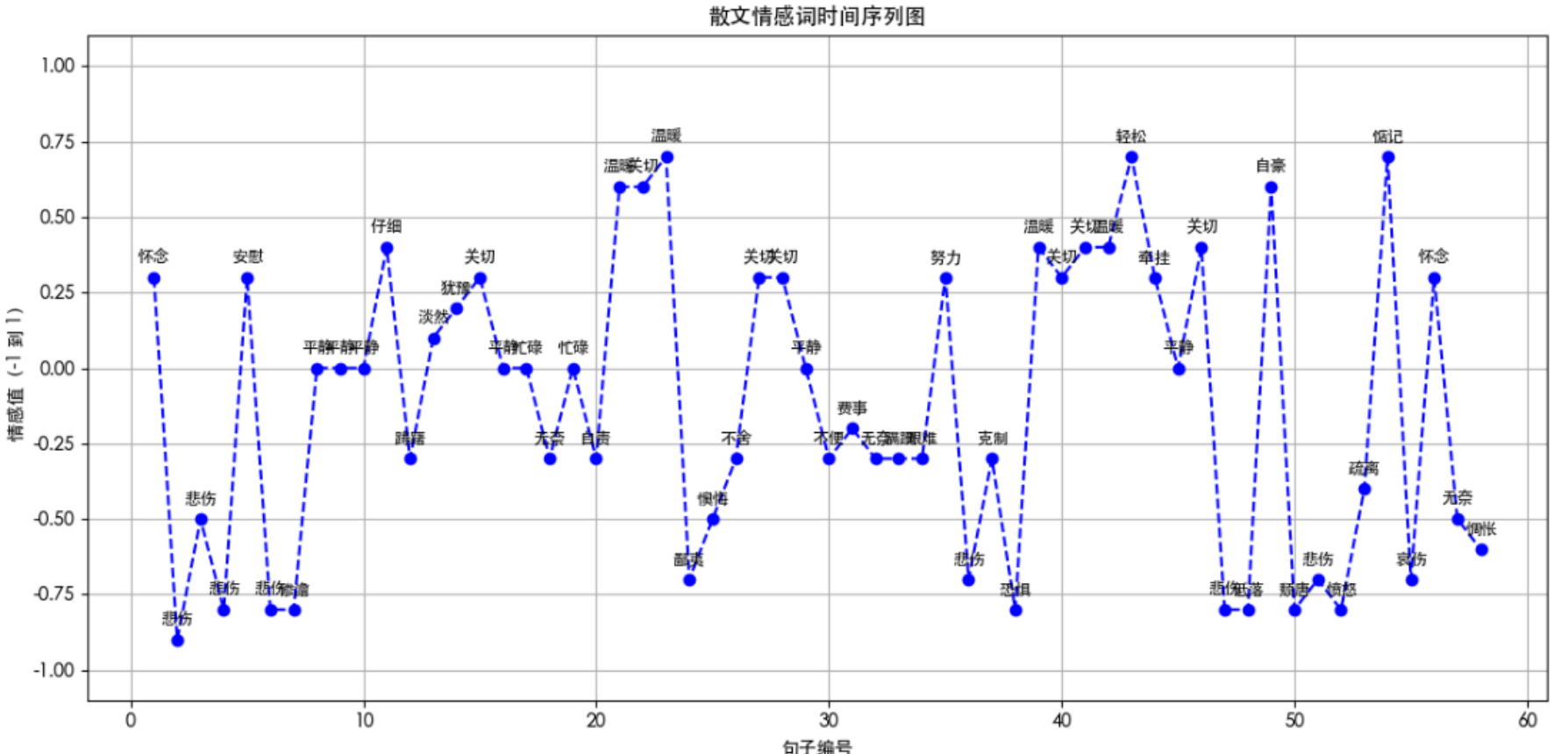
（3）sentiment_analysis_3运行



（4）sentiment_analysis_4运行

Code　　JupyterLab　Python 3 (ipykernel)

```
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()
```



美团多模态菜品知识图谱生态构建：用AI理解“吃的学

05

1. 项目背景与战略目标

业务痛点

美团作为本地生活服务平台，菜品是最基础的供给单元。但传统依赖文本的菜品知识挖掘，在烧烤、火锅等品类准确率不足64%，食材覆盖率仅约68%，难以支持精细化运营。

核心任务

构建一个能够系统化、细粒度理解菜品的知识图谱，从“是什么”深入到“由什么构成、有什么特点”，为搜索、推荐、运营提供智能基础。

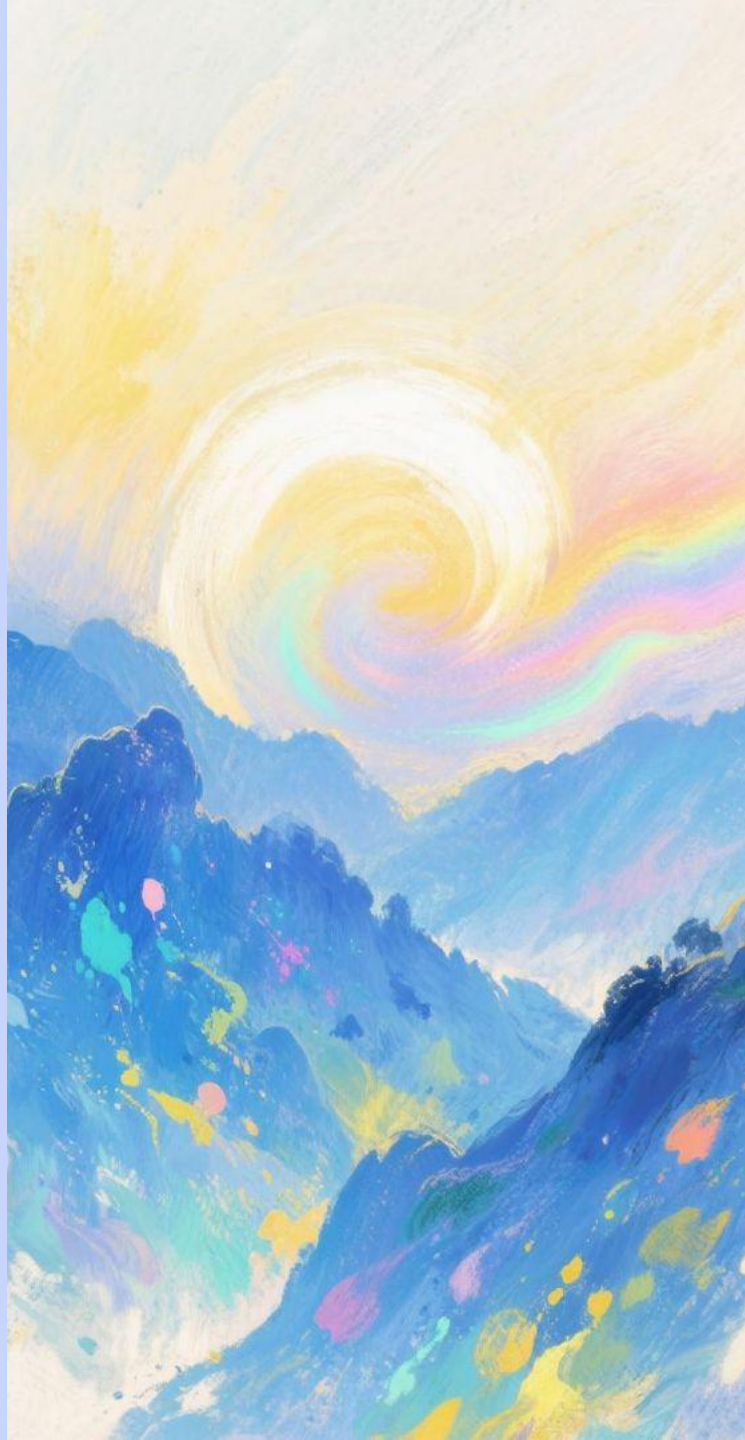
最新进展

美团与天津大学合作，于2024年在国际权威期刊发表了跨模态食材级数据集CMIngre的研究，并以此为核心技术，推动菜品知识图谱进入多模态信息融合的新阶段。

2. 核心技术突破：从“看图识字”到到“精准理解”

- 构建的核心是解决菜品信息稀疏、视觉复杂、常识依赖强的难题，其技术路径体现了从感知到认知的深化

技术挑战	传统方案局限	美团多模态创新方案
信息稀疏	仅分析菜名、文本，信息有限。	**文本+视觉多源信息融合**：不仅用NLP抽取菜谱文本中的食材，更用CV识别图片中的食材及其空间位置。
食材识别难	通用物体检测模型对形态多变的食材（如切碎的葱花、融化的芝士）效果差。	**构建专用跨模态数据集CMIngre**：包含8001张高质量“菜品图-食材列表”对，标注了**95,290个食材边界框**，为算法提供精准“教材”。
常识推理缺	知道“红烧肉”成分是“肉”，但不知道这里的“肉”特指“带皮五花肉”。	**知识挖掘与显式推理**：从海量菜谱中统计食材搭配概率，结合烹饪常识进行推理，将泛化标签转化为精准知识。



3. 生态构建与业务应用

技术能力通过“多模态知识图谱构建流程”形成标准化生产力，并融入业务闭环

构建流程

文本侧：命名实体识别抽取食材、口味、烹饪方法等。
图像侧：目标检测模型识别图像中的食材区域。
知识融合与关联：对齐图文信息，并通过共同实体关联不同菜品，形成全景式图谱。

业务赋能与价值

搜索与推荐：当用户搜索“清爽”时，能精准召回含有黄瓜、薄荷等视觉可识别清爽食材的菜品，极大改善长尾搜索体验。
标准化与品控：为“相同菜品识别”等场景提供关键理解信息，使识别错误率从20.38%大幅降低至2.3%，提升运营效率。
知识沉淀：将菜品知识准确率提升至96.52%，覆盖率提升至87.01%，构建了可复用的数字化资产。

4. 评价与启示

评价

创新性：该项目不仅是工程实践，更是前沿研究（CMIngre数据集）与产业需求结合的典范，解决了中餐数字化理解中“细粒度”和“多模态”两大核心挑战，走在领域前列。

实用性：技术直接服务于核心业务指标（搜索体验、运营效率），形成了“数据-算法-应用-价值”的清晰闭环，ROI明确。

前瞻性：该图谱是美团构建“生活服务超级大脑”的关键一环。从理解菜品（吃）出发，未来可无缝扩展至理解零售商品（买），模式具备强可扩展性。

启示

垂直领域知识图谱的深度取决于对业务的解构粒度。美团证明，从“菜品”深入到“食材”是提升服务智能化的关键路径。

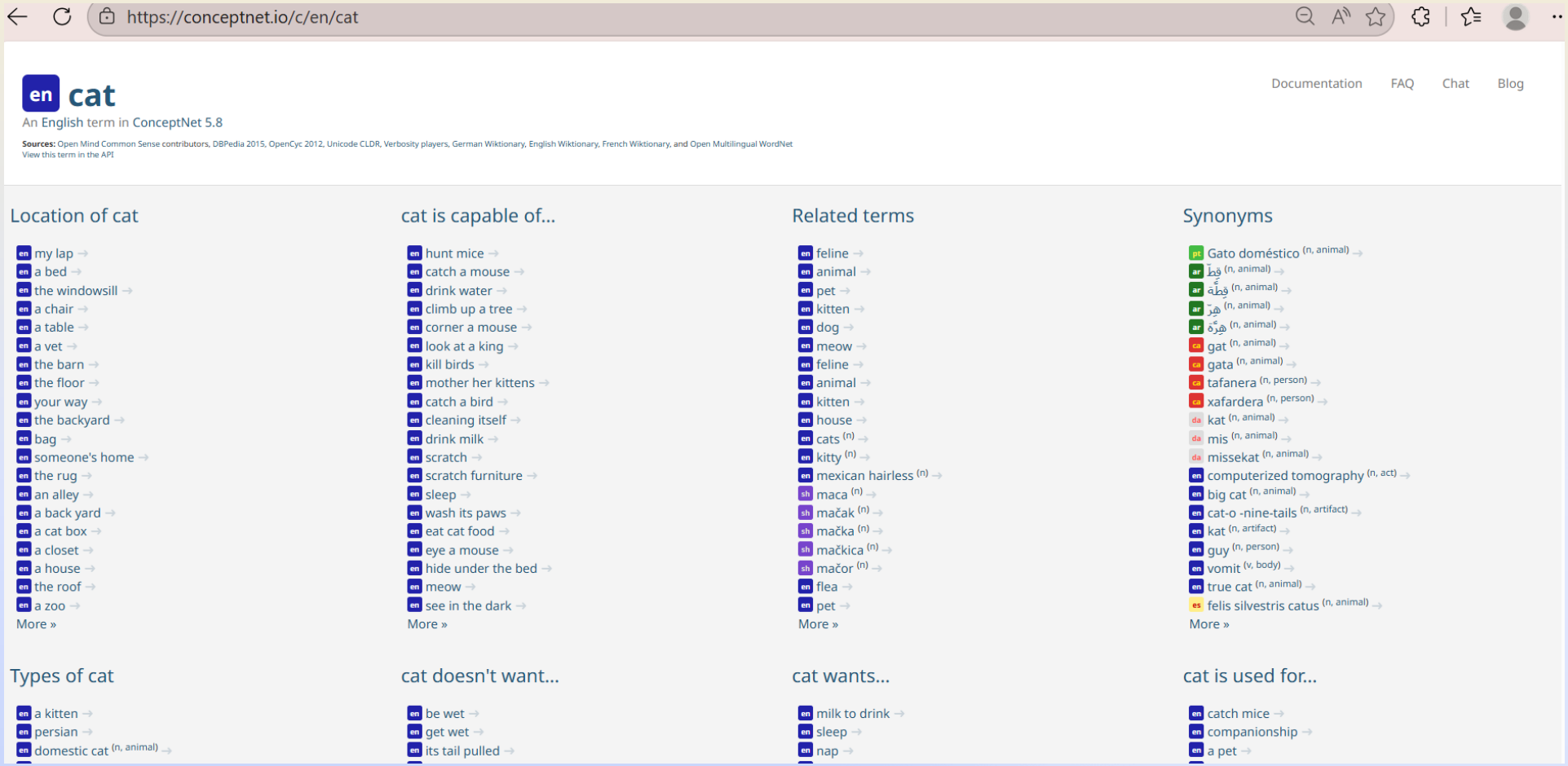
多模态是突破文本信息瓶颈的必然选择，尤其在高频、非标准的消费领域。

“产-学-研”协同是攻克前沿难题、建立技术壁垒的有效模式。

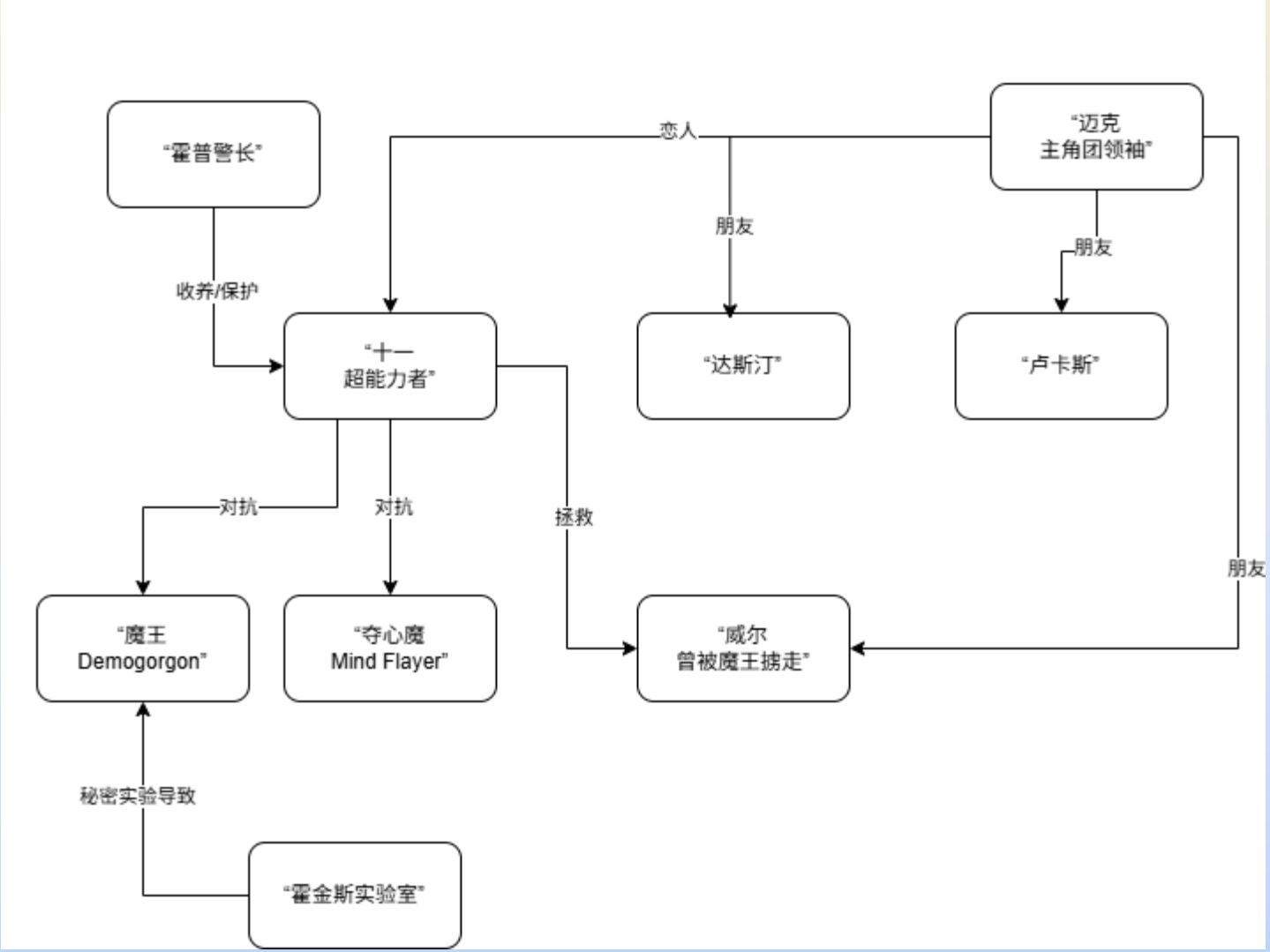
组织世界-语义关联与知识图谱

06

1、ConceptNet链接平台，检索



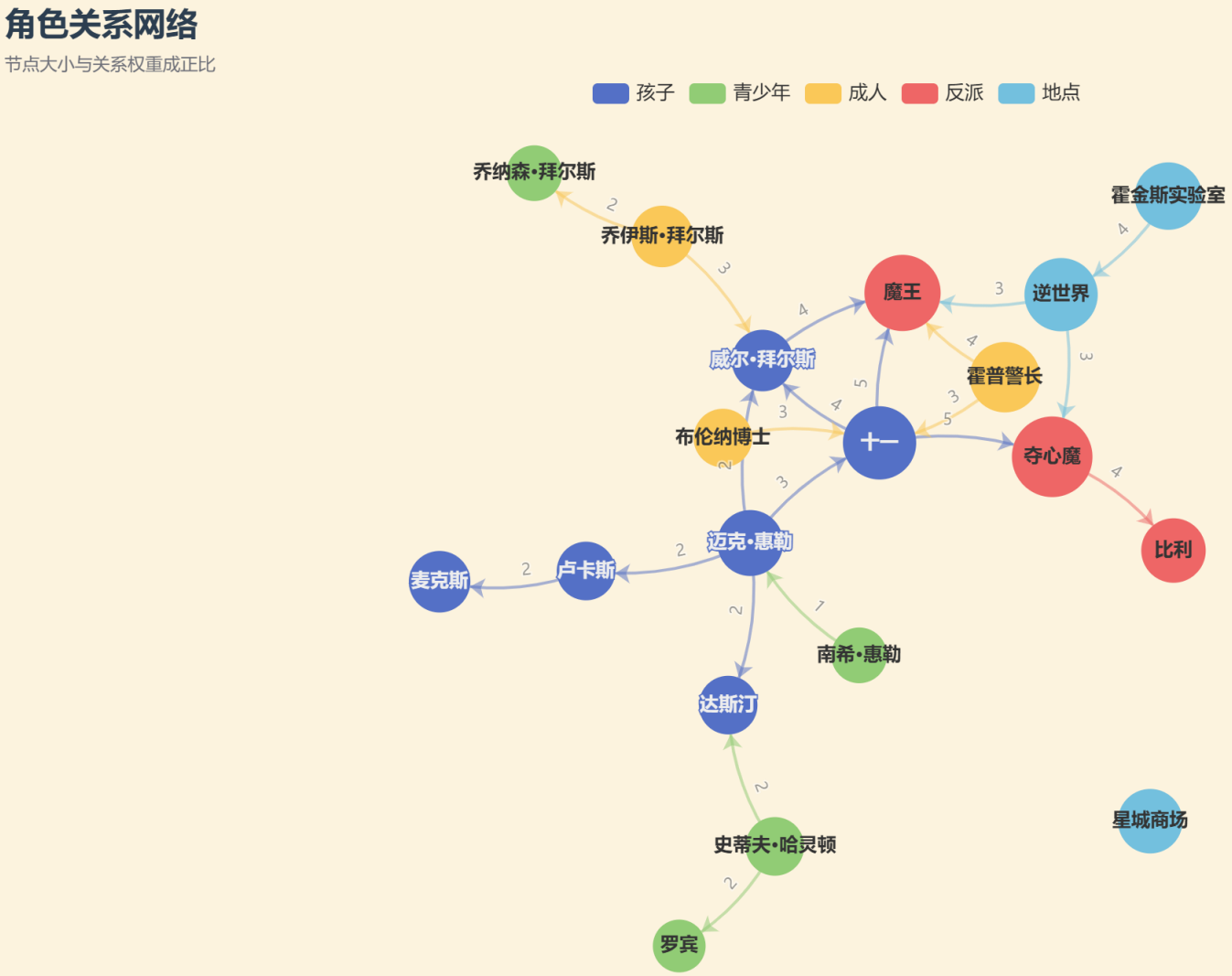
2、使用白板建模绘制 怪奇物语“知识图谱” 人物关系，



这个知识图谱梳理了《怪奇物语》的核心关系：霍普警长收养保护超能力者“十一”，“十一”与主角团领袖迈克是恋人，迈克和达斯汀、卢卡斯、曾被掳走的威尔为朋友，达斯汀参与拯救威尔；同时“十一”要对抗由霍金斯实验室实验催生的魔王、夺心魔，清晰呈现了主角团的人际联结与正邪冲突脉络。

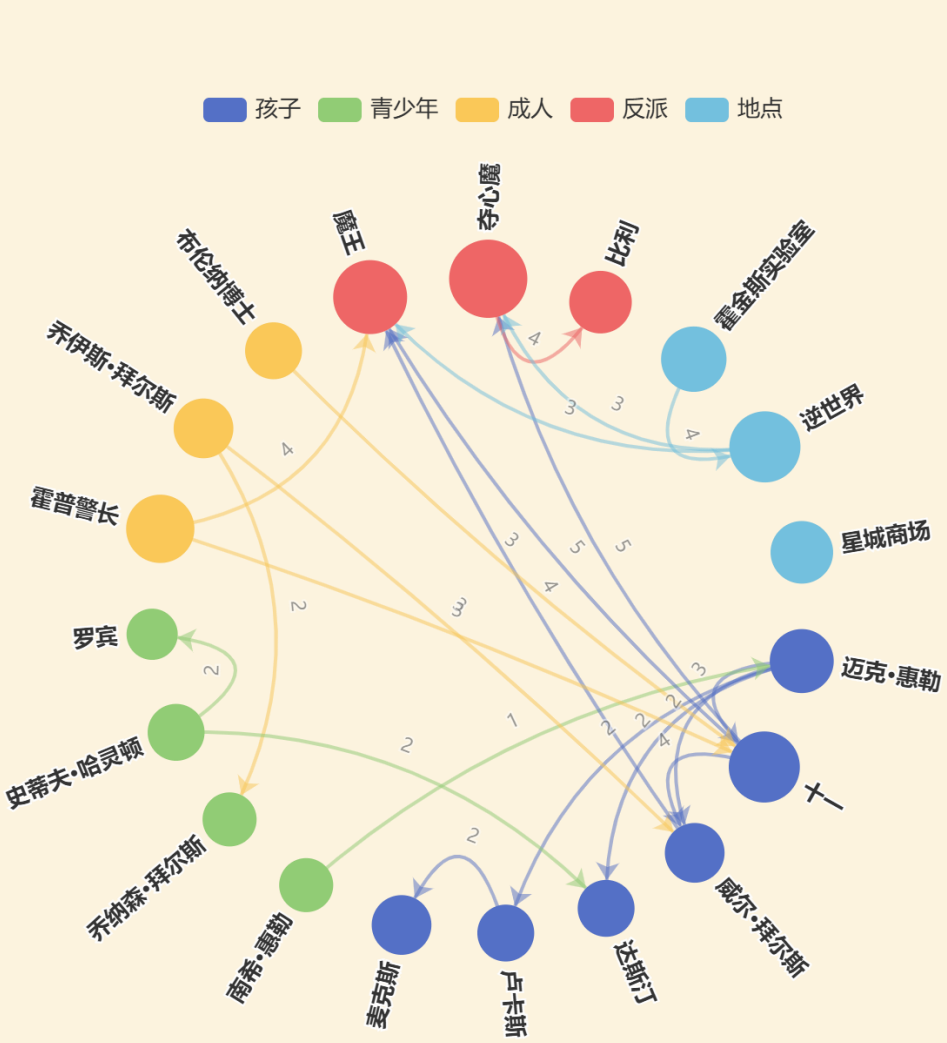
3、使用echarts中的关系图，绘制怪奇物语“知识图谱”

- 力引导布局



3、使用echarts中的关系图，绘制怪奇物语“知识图谱”

- 环形布局



4、使用Neo4j，编程绘制怪奇物语知识图谱