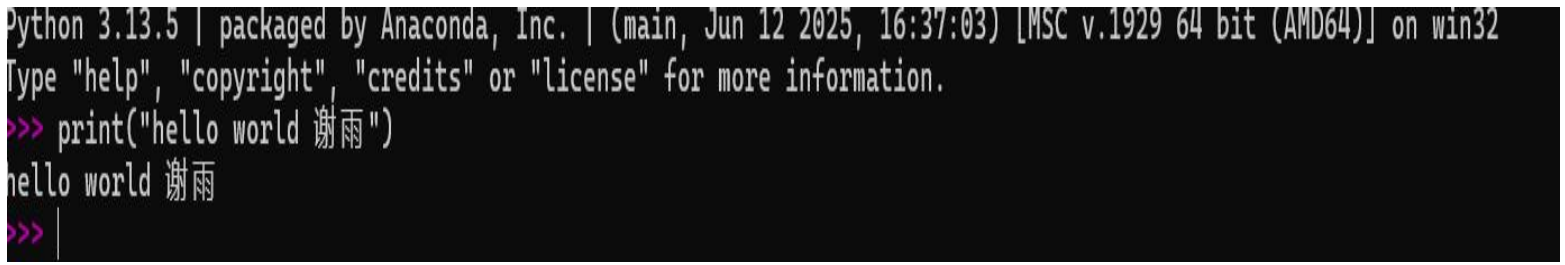
第一讲

1-2


THULAC：一个高效的中文词
THULAC：一个高效的中文词法分析工具包
欢迎使用THULAC中文分词工具包demo系统
周末去公园散步时，刚好赶上樱花开得最盛的时节。粉白的花瓣被风一吹，簌簌落在肩头，空气里都是淡淡的甜香。不远处有小朋友追着泡泡跑，笑声脆生生的，连坐在长椅上晒太阳的老爷爷，都忍不住跟着笑眯了眼。原来最治愈的时光，从来都藏在这些寻常的小美好里呀。
[测试 Try]
周末_t 去_v 公园_n 散步_v 时_t ，_w 刚好_d 赶上_v 樱花_n 开_v 得_u 最_d 盛_a 的_u 时节_n 。_w 粉白_a 的_u 花瓣_n 被_p 风_n 一_m 吹_v ，_w 簌簌_o 落_v 在_p 肩头_n ，_w 空气_n 里_f 都_d 是_v 淡淡_a 的_u 甜香_n 。_w 不_d 远处_n 有_v 小朋友_n 追_v 着_u 泡泡_n 跑_v ，_w 笑声_n 脆生生_a 的_u ，_w 连_p 坐_v 在_p 长椅_n 上_f 晒太阳_v 的_u 老爷爷_n ，_w 都_d 忍不住_v 跟着_v 笑眯_v 了_u 眼_n 。_w 原来_d 最_d 治愈_v 的_u 时光_n ，_w 从来_d 都_d 藏_v 在_p 这些_r 寻常_a 的_u 小_a 美好_a 里_f 呀_y 。_w
词性解释
n/名词 np/人名 ns/地名 ni/机构名 nz/其它专名
m/数词 q/量词 mq/数量词 t/时间词 f/方位词 s/处所词
v/动词 vm/能愿动词 vd/趋向动词 a/形容词 d/副词
h/前接成分 k/后接成分 i/习语 j/简称
r/代词 c/连词 p/介词 u/助词 y/语气助词
e/叹词 o/拟声词 g/语素 w/标点 x/其它
版权所有：清华大学自然语言处理与社会人文计算实验室
Copyright：Natural Language Processing and Computational Social Science Lab, Tsinghua University


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 谢雨")
hello world 谢雨
>>>

001

```
seg_list = jieba.cut("南京大学生都爱南京市长江大桥")

print('*'.join(seg_list))
```

南京*大学生*都*爱*南京市*长江大桥

```
for seg in seg_list:
    if seg not in stopwords:
        final += seg+'*'

print (final)   # 做实体抽取的时候，停用词表很管用
```

使用*词表*之后*效果*就*好看*很多*什么*之类*词*就*不见*

```
jieba.load_userdict('dict.txt')

seg_list = jieba.cut("吴志祥是南京工业最好大学青年教师，经济与管理学院的，他对那种二次元小魔仙是无感的，这怎么行？python看上去还有点人性")

print('%'.join(seg_list))
```

吴志祥%是%南京工业最好大学%青年教师%，%经济%与%管理%学院%的%，%他%对%那种%二次元小魔仙%是%无感%的%，%这%怎么%行%？%python%看上去%还%有点人性

质量,不大,好

```
ss = jieba.cut('质量不大好')

print(",".join(ss))
```

质量,不大好

```
s1 = SnowNLP(u"吴志祥是南京工业大学青年教师，他对那种二次元小魔仙是无感的，这怎么行？")

print(",".join(s1.words))  # 因为snownlp擅长处理英文
```

吴,志祥,是,南京,工业,大学,青年,教师,，,他,对,那种,二次元,小,魔仙,是,无,感,的,，,这,怎么,行,？

002

[的 ， 了 ， 是 ， 啊 ， 、 ， ， ， 。 ， 停用]

生成 + 代码 + Markdown

```python
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潜艇总体研究设计所研究员
```

Python

```python
final = ''
```

Python

```python
for seg in seg_list_huang:
    if seg not in stopwords:
        final+= seg+'/'
```

Python

```python
print(final)
```

Python

```
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/
```

003

```python
print("特定词汇词频统计结果：")
for word in target_words:
    print(f"'{word}': {word_counts[word]}次")
```

Python

```
特定词汇词频统计结果：
'数字化': 2次
'智能化': 3次
'安全': 2次
```

```python
# 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")
```

Python

```
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
```

004

```
    else:
        print(f"请求失败，状态码：{response.status_code}")
        print(response.text)
```

Python

```
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",
    "免疫抑制"
  ],
  "方法": [
    "单细胞RNA测序",
    "scRNA-seq",
    "细胞亚群聚类",
    "轨迹分析",
    "pseudotime推断",
    "细胞间通讯网络"
  ],
  "工具": [
    "Seurat",
    "Monocle3",
    "CellChat"
  ],
  "疾病与样本": [
    "非小细胞肺癌",
    "肿瘤样本"
...
    "个体化免疫治疗"
  ]
}
```
```

4

该研究围绕基于关键词的学术文本聚类展开，核心是探究聚类集成方法对学术文本分类性能的影响，以及关键词抽取方法、关键词个数这两个因素的作用。

聚类集成方法可显著提升学术文本聚类性能。研究对比了以 K-means 为基础的聚类集成、以增量聚类为基础的聚类集成，与单一的 K-means、增量聚类算法，发现聚类集成方法的 F1 值普遍更高，且经 T 检验验证，性能差异具有显著性；同时，在不同关键词抽取方法和关键词个数下，聚类集成方法的稳定性也更好。

关键词个数与聚类集成性能呈正相关。随着关键词个数从 5 个增加到 60 个，学术文本聚类的性能逐步提升，能更充分地体现文本主旨；且在关键词个数较少时，聚类集成方法的优势更突出，可缓解关键词数量不足对分类的影响。

第二讲
1

关键词总数大幅增长，研究维度持续丰富

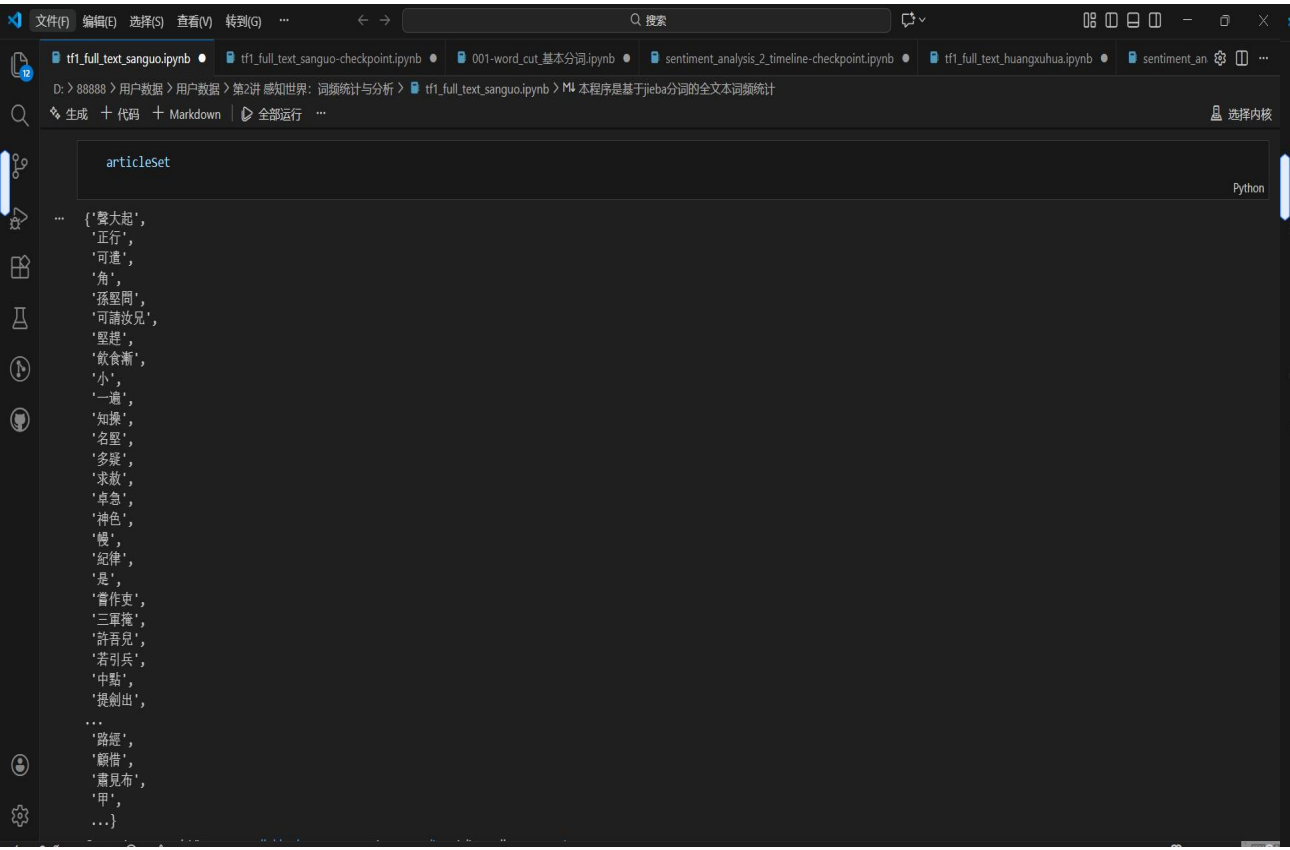
环保相关主题文献的关键词总数十年间增长显著，2024 年较 2014 年平均增长超 2 倍

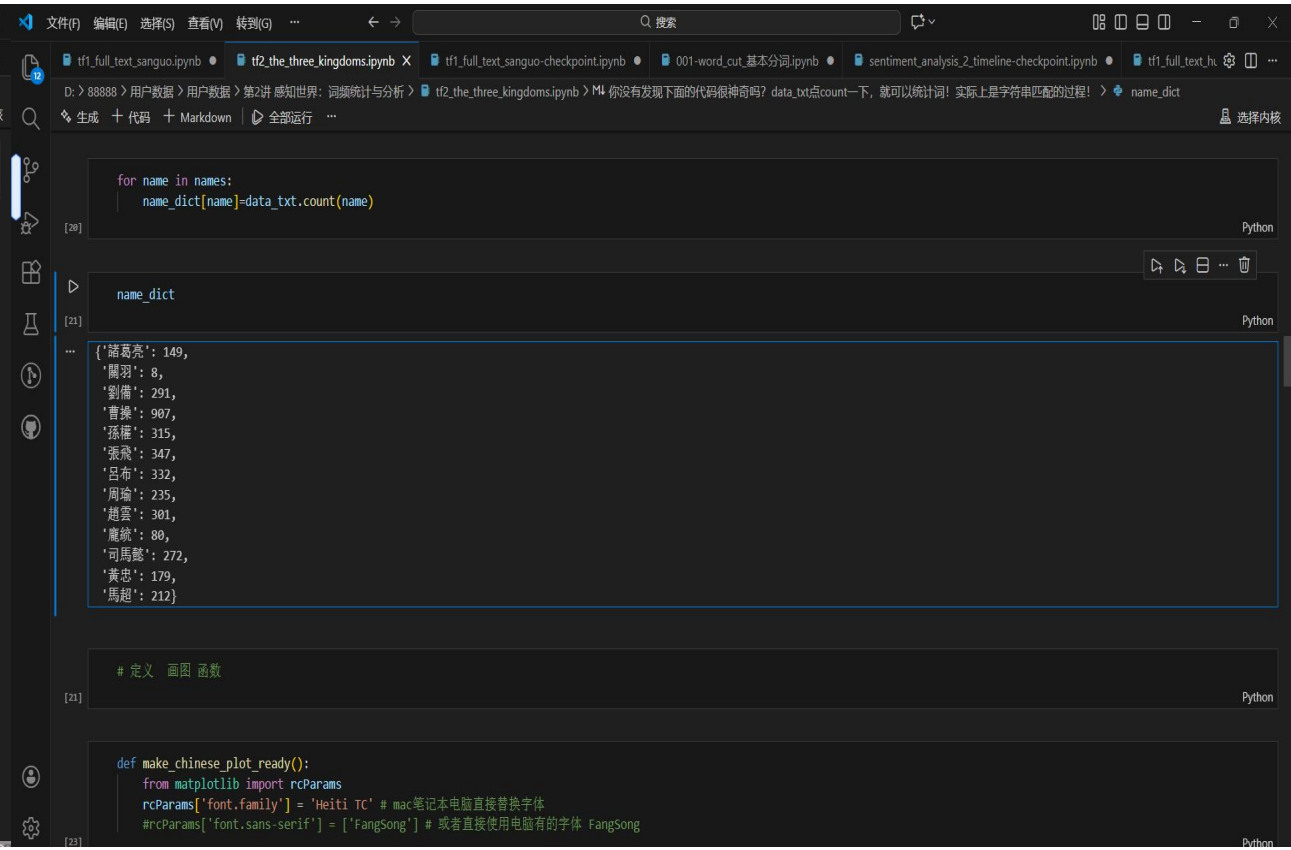
2014 年前后，“污染治理”“PM2.5”“污水处置” 等传统治污类关键词占比达 40%-50%，是绝对核心。

早期关键词多集中于单一治理场景，关联性较弱；2024 年关键词形成多维度聚类网络，“协同发展”“生态产品”“绿色经济” 等跨领域关键词突现强度最高，成为连接不同研究方向的核心纽带。

关键词关联度提升，反映环保研究从 “单点突破” 转向 “系统协同” 的整体趋势。

2


articleSet
Python
{'鬱太起',
'正行',
'可遣',
'角',
'孫堅問',
'可請汝兄',
'堅趕',
'飲食漸',
'小',
'一遍',
'知操',
'名堅',
'多疑',
'求赦',
'卓急',
'神色',
'幔',
'紀律',
'是',
'嘗作吏',
'三軍掩',
'許吾兒',
'若引兵',
'中黔',
'提劍出',
...
'路經',
'顧惜',
'肅見布',
'甲',
...}

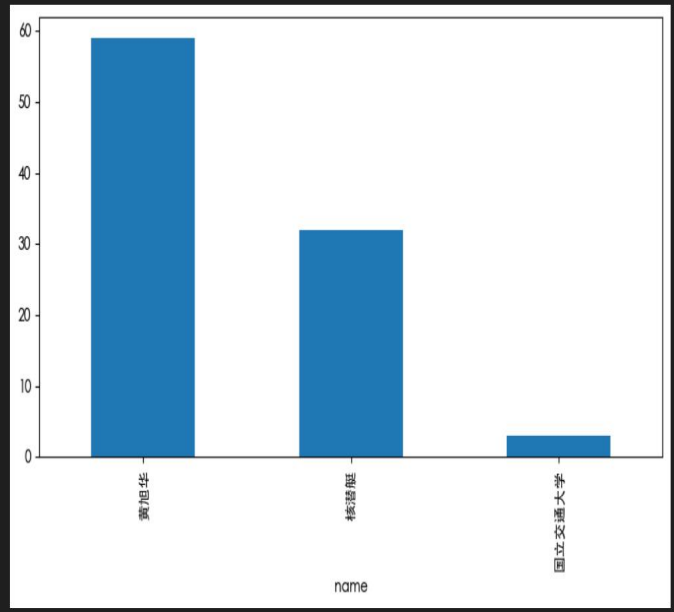

for name in names:
name_dict[name]=data_txt.count(name)
[20]
Python
name_dict
[21]
Python
{'諸葛亮': 149,
'關羽': 8,
'劉備': 291,
'曹操': 907,
'孫權': 315,
'張飛': 347,
'呂布': 332,
'周瑜': 235,
'趙雲': 301,
'龐統': 80,
'司馬懿': 272,
'黃忠': 179,
'馬超': 212}
定义 画图 函数
[21]
Python
def make_chinese_plot_ready():
from matplotlib import rcParams
rcParams['font.family'] = 'Heiti TC' # mac笔记本电脑直接替换字体
#rcParams['font.sans-serif'] = ['FangSong'] # 或者直接使用电脑有的字体 FangSong
[23]
Python

3

```
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)
('第一代', 8)
('介绍', 8)
('主要', 8)
('科学', 8)
('思想', 7)
('人生', 7)
('及其', 7)
('历史', 7)
('传记', 7)
...
('节点', 3)
('叙述', 3)
('计划', 3)
('直接', 3)
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings...
```

```
draw_dict(terms_dict)
```



4

这篇论文用谷歌图书、谷歌学术的数字化资源研究杰出物理学家的科学声誉，发现牛顿、爱因斯坦等虽离世，影响力却持续数世纪。全球范围看，20 世纪中期后爱因斯坦在学术界声誉超牛顿，但存在 “群体内偏好”，比如英式英语书籍中牛顿更受关注，爱因斯坦则在美式英语、德语书籍中提及度更高。共现分析显示，牛顿的声誉与万有引力定律、微积分等相关，爱因斯坦则与相对论、量子理论相关。研究还列出 21 世纪知名度前 20 的物理学家，爱因斯坦、普朗克、牛顿居前三。论文认为谷歌相关工具可作替代计量工具衡量科学家的社会影响力，也指出语料库语言偏见、姓名消歧误差等局限性，建议未来拓展研究领域

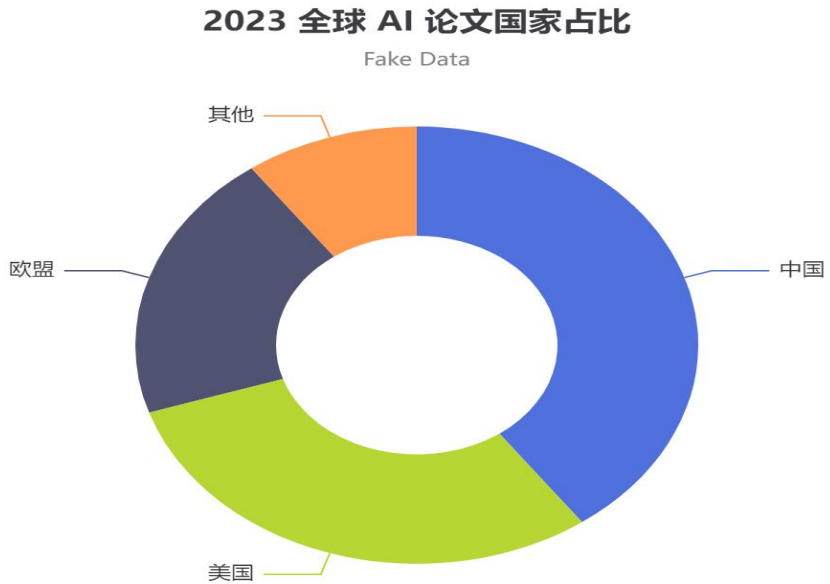
第三讲
1

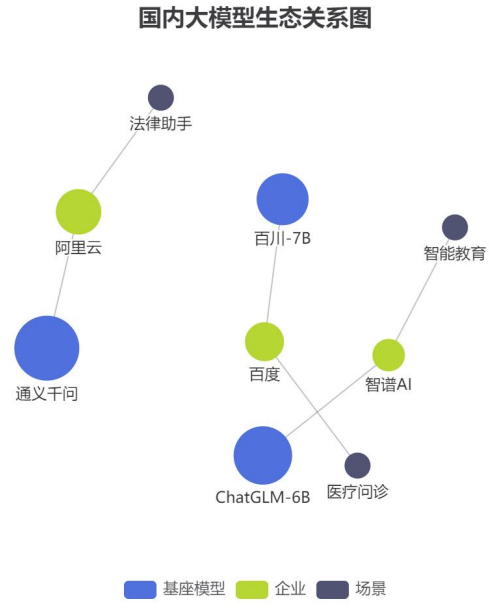
即梦词云

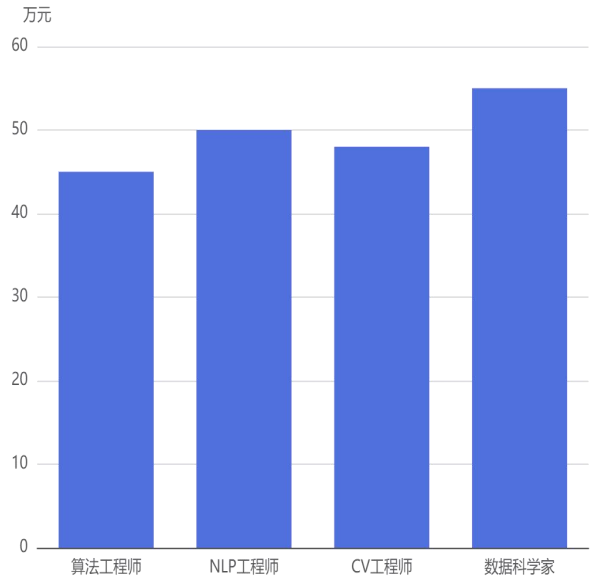
该词云图围绕“人工智能发展”主题，提取了当前AI领域最具代表性的关键词。图中“人工智能”作为核心词汇居中放大，周围环绕着“深度学习”“大模型”“AIGC”等高频热词，体现了技术演进的焦点。整体配色采用蓝绿渐变，象征科技与未来感，突出AI技术的前沿性与广泛影响力。

2

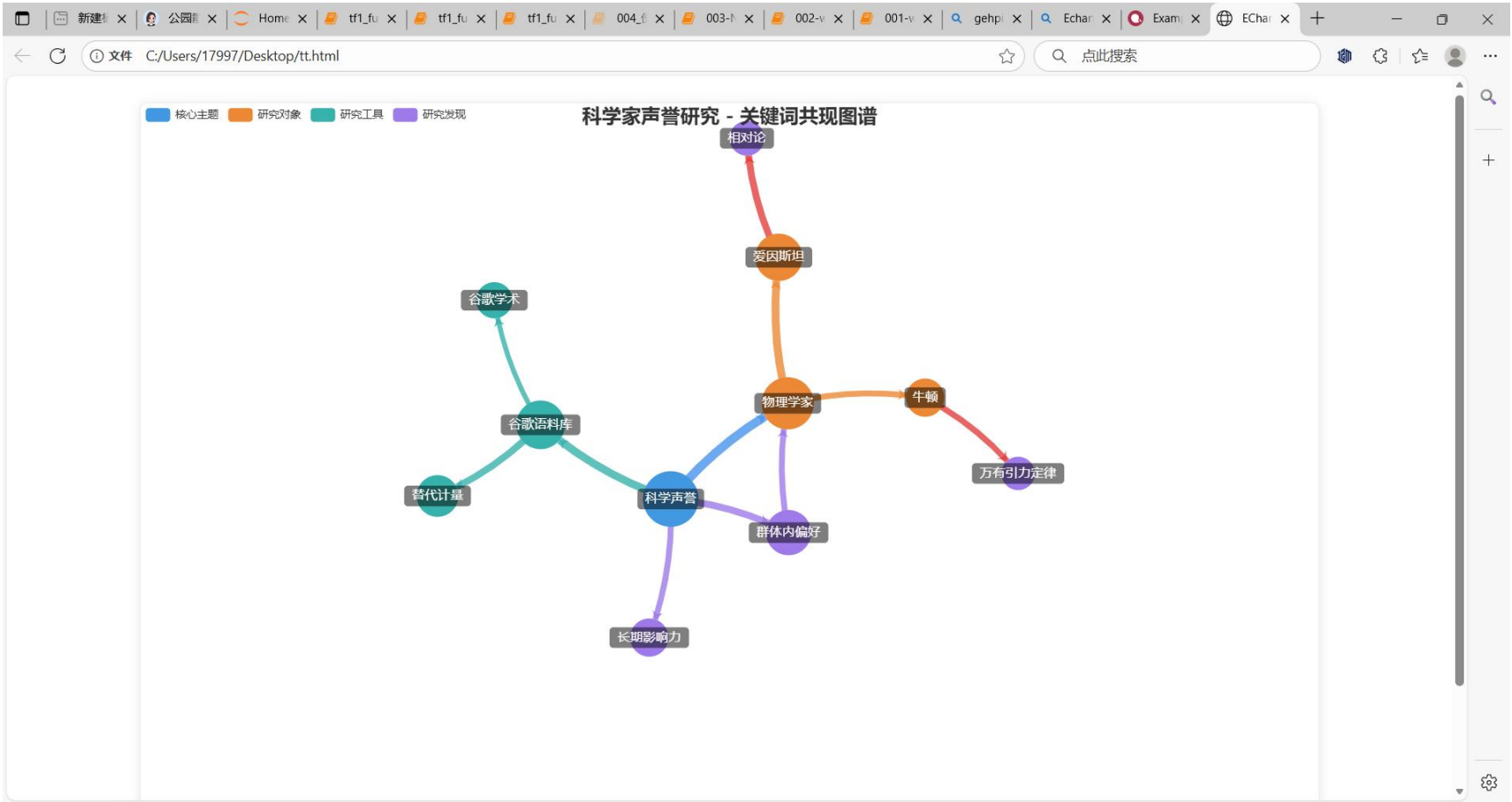
条形图：AI 岗位年薪 45–55 万，模型越新越值钱。
饼图：中国 AI 论文量全球 4 成，领先美欧。
关系图：基座—企业—场景三层链，大模型生态一目了然。







3


核心主题
研究对象
研究工具
研究发现
科学家声誉研究 - 关键词共现图谱
相对论
爱因斯坦
谷歌学术
物理学家
牛顿
谷歌语料库
万有引力定律
替代计量
科学声誉
群体内偏好
长期影响力

4

文件(F) 编辑(E) 选择(S) 查看(V) 转到(G) ··· 搜索

sanguo.ipynb ● | tf2_the_three_kingdoms.ipynb | tf1_full_text_huangxuhua.ipynb D:\88888\... | tf2_the_huangxuhua.ipynb | wordcloud_huangxuhua-checkpoint.ipy

D: > 88888 > 用户数据 > 用户数据 > 第3讲 感知世界：词云与可视化 > 第3讲 感知世界：词云与可视化 > .ipynb_checkpoints > wordcloud_cn-checkpoint.ipynb > M↓ 词云 > M↓ 给词云加上

生成 + 代码 + Markdown | 全部运行 Restart 清除所有输出 | Jupyter Variables Outline ···

接下来，你还能想到哪些比较实用的改进呢?

给词云加上特定的形状

```
big_pic = imread("D:\\88888\\用户数据\\用户数据\\第3讲 感知世界：词云与可视化\\第3讲 感知世界：词云与可视化\\.ipynb_checkpoints\\bg2-check
```

[105] ✓ 0.0s

```
# 替换原 font_path 参数，使用 Windows 系统自带宋体的完整路径
wordcloud_cn_x = WordCloud(
    mask=big_pic,
    width=500,
    height=300,
    background_color='white',
    scale=2.5,
    font_path="C:/Windows/Fonts/simsun.ttc"  # 重点修改这一行
).generate(text_cn_word)
plt.imshow(wordcloud_cn_x)
plt.axis('off')
```

[106] ✓ 1.9s

```
··· (np.float64(-0.5), np.float64(2559.5), np.float64(2414.5), np.float64(-0.5))
```

```
wordcloud_cn_x.to_file('yes_minister_x.jpg')
```

[107] ✓ 0.1s

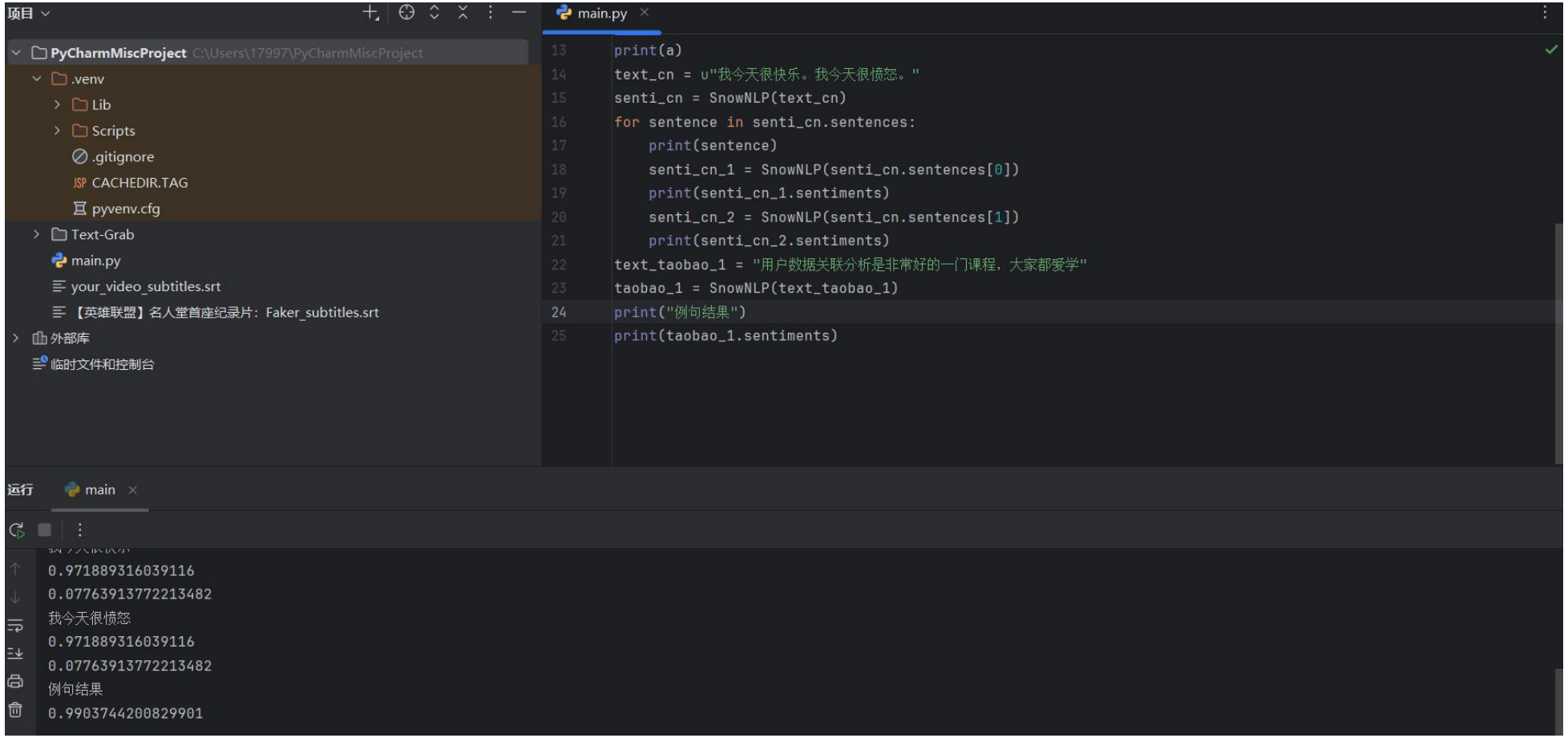
```
··· <wordcloud.wordcloud.WordCloud at 0x1dd6daf91d0>
```

第四讲

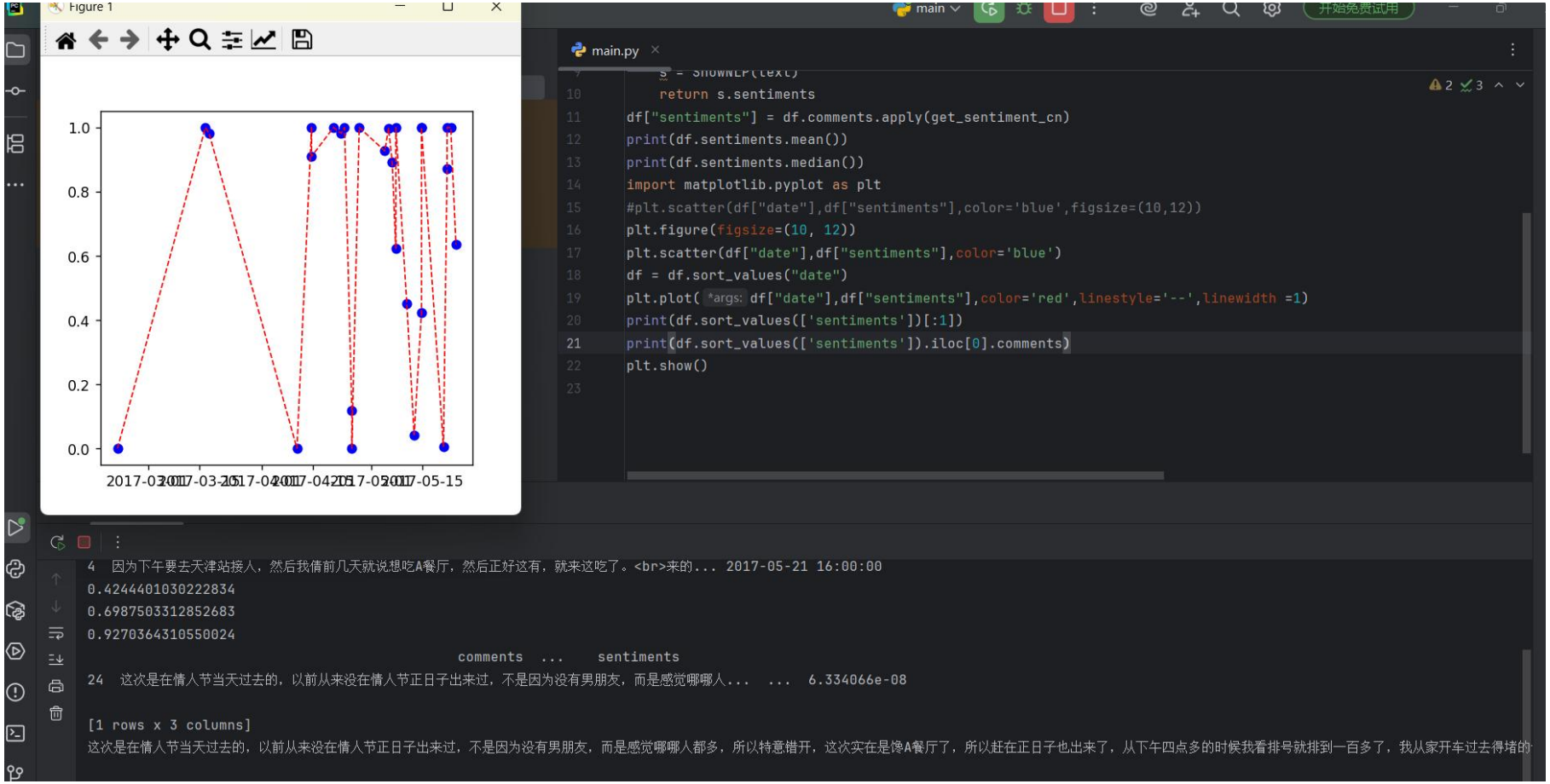
- 1


https://hanlp.hankcs.com/demos/sentiment.html?text=这款+Python+数据分析库的+API+设计得太人性化了，几行代码就能完成数据清洗和可视化，帮我高效搞定了课程大作业的数据分析部分，强烈...
在线演示
HanLP
安装
演示
文档
书籍
引用
论坛
博客
API
搜索
中文分词
词性标注
命名实体识别
依存句法分析
成分句法分析
语义依存分析
语义角色标注
抽象意义表示
指代消解
语义文本相似度
文本风格转换
关键词短语提取
抽取式自动摘要
请输入一段中文文本:
这款 Python 数据分析库的 API 设计得太人性化了，几行代码就能完成数据清洗和可视化，帮我高效搞定了课程大作业的数据分析部分，强烈推荐给入门学习者!
78/1000
情感分析
此页内容
简介
调用方法
创建客户端
情感分析
本地调用
多语种支持
情感极性
负面
正面
非常负面
非常正面
0
-0.5
0.5
-1
1
+0.95
情感极性

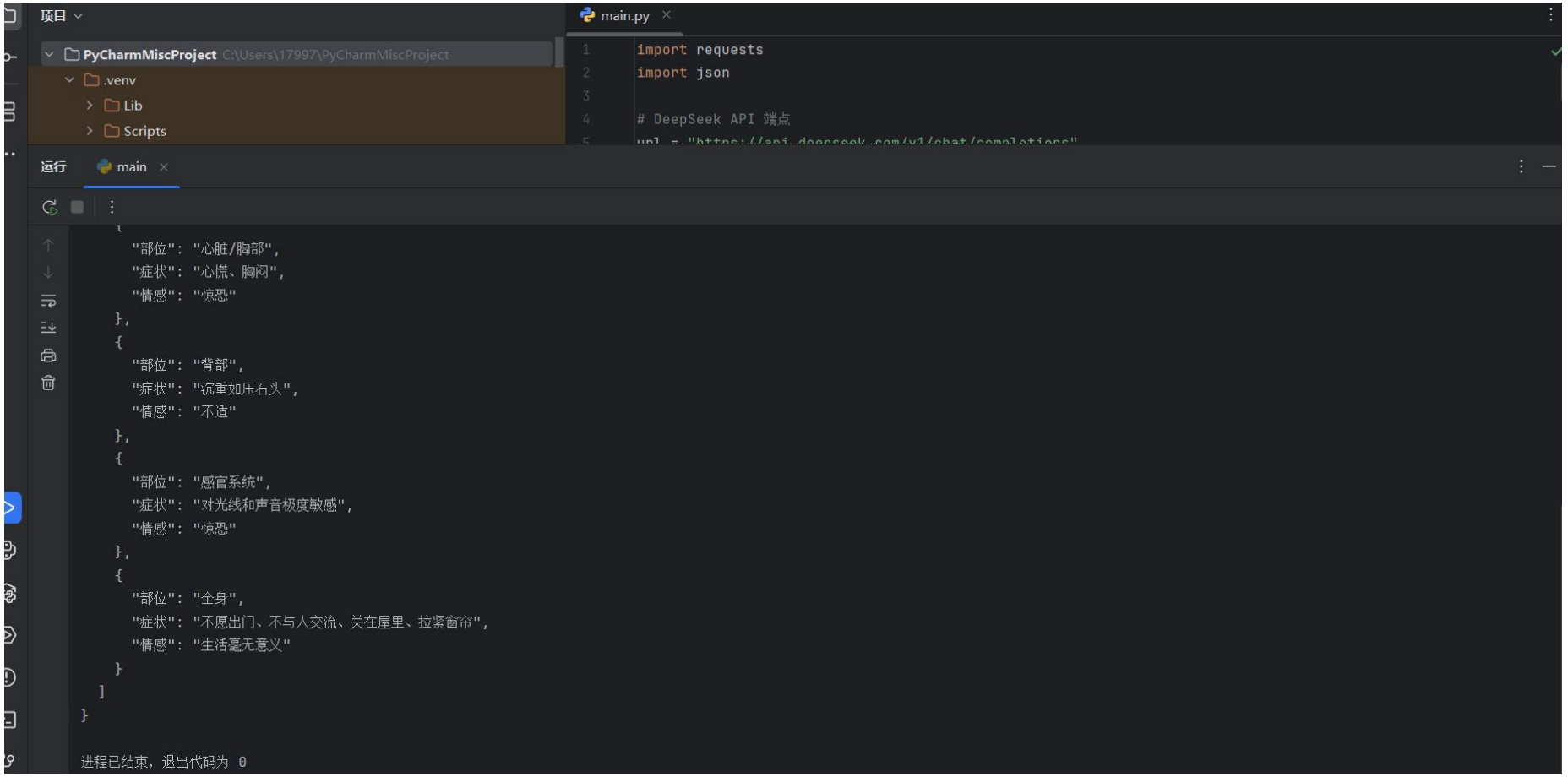
2 sentiment_analysis_1



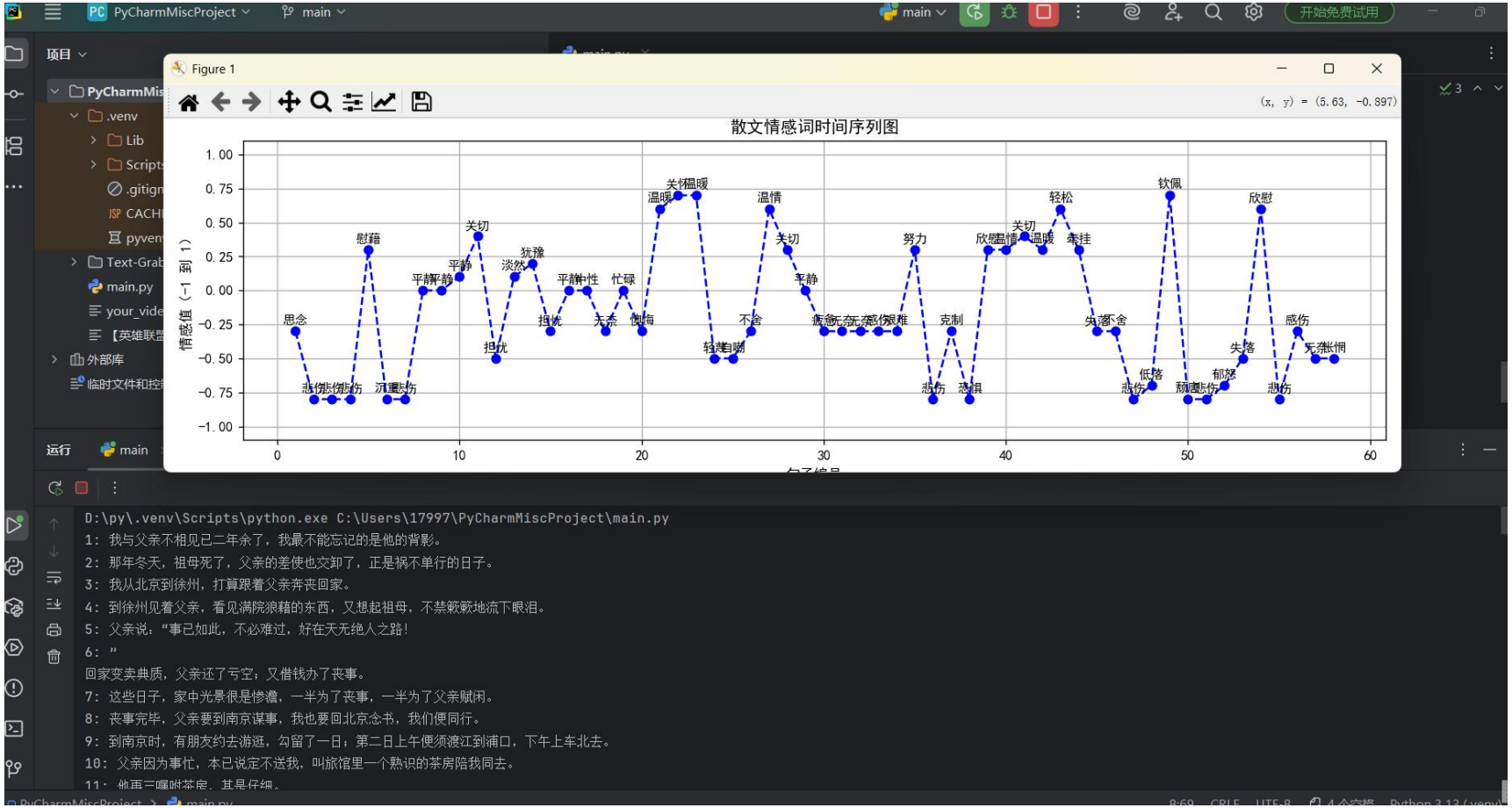
sentiment_analysis_2



sentiment_analysis_3



sentiment_analysis_4



总结分析

- 四段代码的功能都是对文本进行情感分析，在运行时发现打印图表时电脑加载比较慢。在将代码移动到pycharm中运行时常常遇到缩进错误等问题，需要手动调整代码缩进。总的来看使用这几段python代码进行情感分析的过程中虽有问题，但都能够很快解决并得出不错的结果，可以看出python在情感分析功能方面具有很好的作用

第六讲（1）美团大脑知识图谱生态构建分析

核心定位与生态规模定位

美团大脑是本地生活服务领域的超级知识中枢，以“全域数据整合 + AI 技术赋能”为核心，打通人、店、商品、场景的关联，支撑美团从“外卖平台”向“综合性生活服务超级 App”升级。

核心生态规模（截至 2025 年最新数据）

知识覆盖：33 类核心概念（含餐饮、酒店、零售、出行等）

实体量级：30 亿 + 业务实体（商户、商品、用户、景点等）

关联强度：1000 亿 + 三元组（如“用户 - 偏好 - 川菜”“商户 - 提供 - 到店服务”）

数据底座：40 亿 + 累计用户评价 + 即时零售 / 无人配送场景动态数据

最新生态构建进展

1. 技术架构升级：从 “静态图谱” 到 “全域协同大脑”

融合美团 BERT 自研模型与多模态感知技术，实现 “文本理解 + 场景识别” 双驱动

打通低空无人机、地面无人车、楼宇配送机器人的数据链路，构建 “空地楼” 全域决策网络

升级 Listwise 排序模型，在 MRR@10 指标上保持行业领先，支撑复杂搜索需求解析

2. 场景生态拓展：从 “餐饮外卖” 到 “全场景即时服务”

关联 13 万 + 可配送商品，实现 “关键词搜索→需求识别→履约匹配” 全链路支撑（如搜索 “口罩” 自动匹配附近药店 + 无人配送）

支持 长句搜索，精准拆解场景、人群、品类需求

为校园、园区、机场等提供定制化图谱服务，适配无人配送路径规划

3. 数据生态激活：从 “被动存储” 到 “主动赋能”

通过 NLP 技术解析 40 亿 + 用户评价，提取菜品、服务、环境等维度偏好，反哺商户经营优化

动态更新实体关联：实时同步商户营业时间、商品库存、配送范围等信息，保证图谱时效性

典型应用案例

案例背景

用户搜索 “疫情期间适合遛娃的户外景点”，需同时满足 “场景安全”“亲子适配”“即时可达” 三大需求。

图谱赋能流程

需求拆解：通过知识图谱识别 “疫情安全”“遛娃”“户外景点”“即时可达” 四大核心实体

关联匹配：调用 景点 - 属性 - 安全评级，景点 - 适配人群 - 儿童，景点 - 配送范围 - 3 公里内”等三元组

履约衔接：匹配就近无人车配送站点，提供 “景点门票购买 + 亲子套餐预订 + 即时物资配送” 一体化服务

案例效果

搜索准确率提升 40%，长句需求满足率达 85%

无人配送履约效率提升 25%，核心场景配送时效压缩至 30 分钟内

生态评价与展望

优势评价

AI 团队与业务团队一体化架构，避免技术浪费，如无人配送数据快速反哺图谱优化

从被动响应需求 到 主动预判，显著降低服务决策成本

现存挑战

跨场景数据治理复杂度高，部分领域（如医疗健康）实体关联精度待提升

无人配送场景的动态数据更新频率，需匹配图谱实时计算能力

个性化推荐具有负面效果风险，需平衡用户偏好与探索性需求

未来展望

深化 具身智能 + 知识图谱 融合，提升无人配送场景的动态决策能力

开放部分图谱能力给第三方商户，提供 用户画像 - 商品优化 - 履约效率”定制化解决方案

拓展 本地生活 + 城市服务 关联，如将交通、政务等实体纳入图谱，打造城市级服务中枢

总结与评价

美团大脑的最新生态构建，核心是以知识图谱为纽带，打通 数据 - 技术 - 场景 - 履约 的通路。既支撑了 C 端用户的复杂需求，也能够支持 B 端商户的数字化经营和无人配送的规模化落地。未来若能持续提升跨领域实体关联精度与开放生态建设，有望成为生活服务领域的基础设施中枢。

优点：贴合外卖、到店消费等日常场景，搜店、点餐更精准，还能帮商家算备货量、提经营建议，功能多样且易于使用。

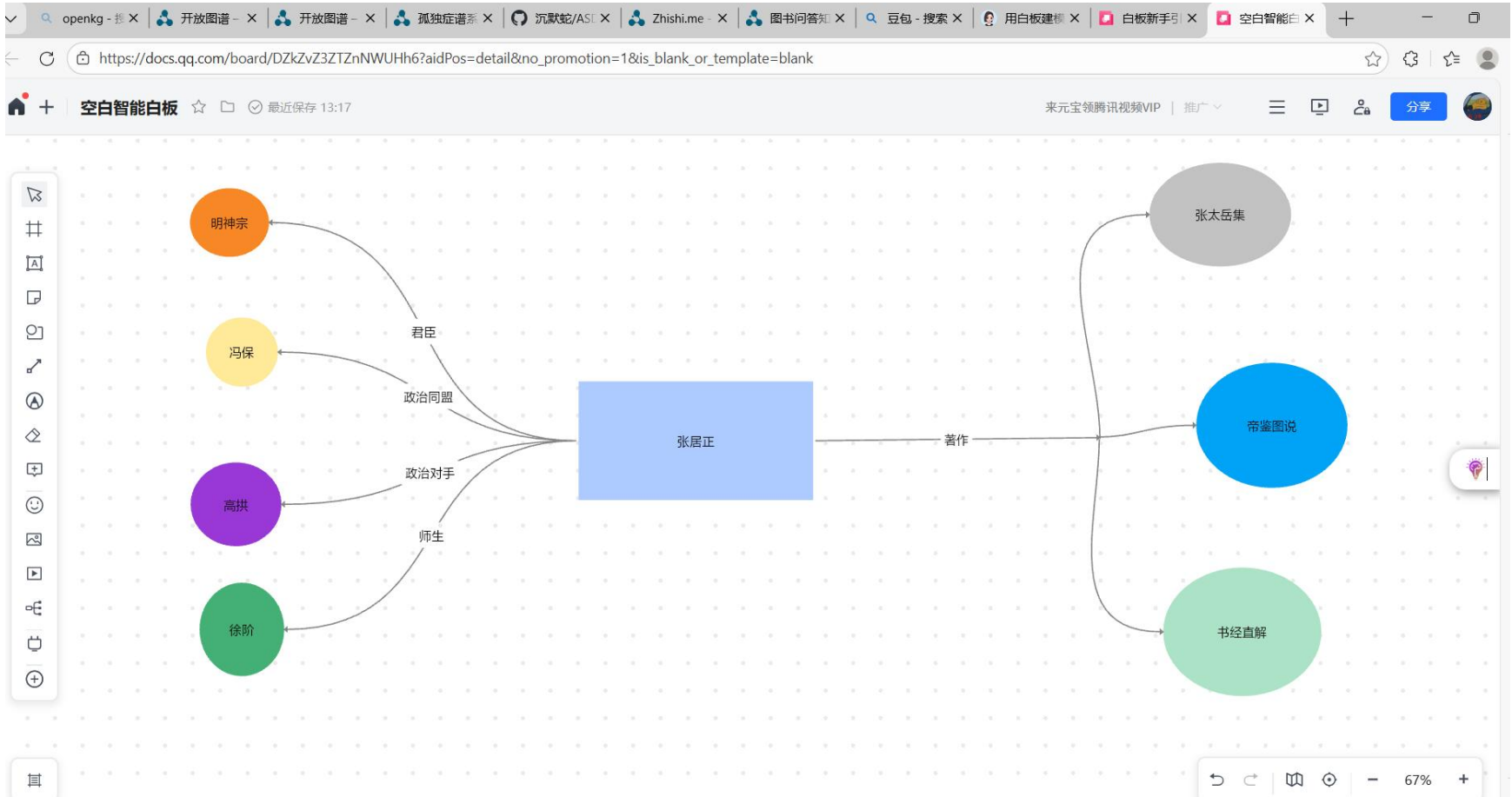
不足：跨领域能力有待完善，还面临用户对于数据安全和隐私保护可靠性的质疑。

第六讲（续）

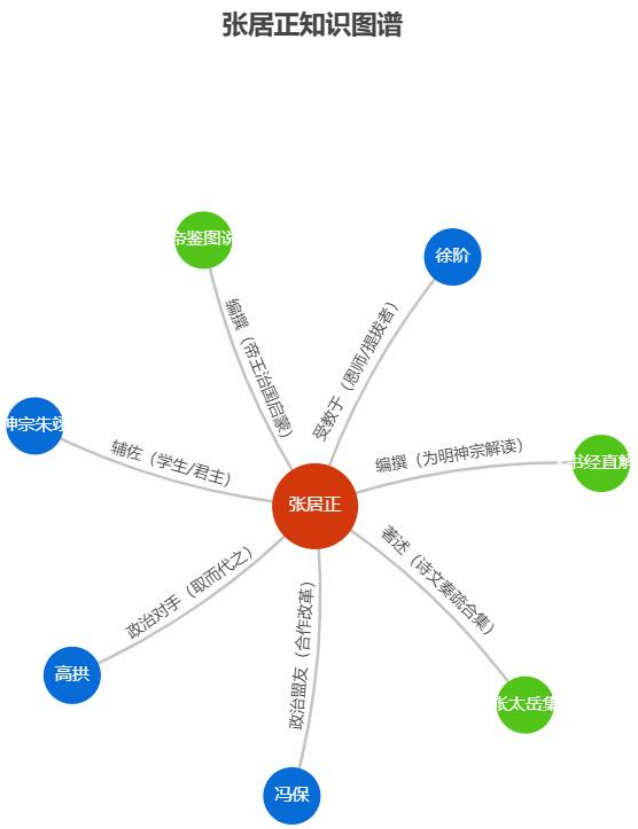
1


openkg - 搜索
开放图谱 – 开放知识图
开放图谱 – 开放知识图
孤独症谱系障碍知识库
沉默蛇/ASDKB
Zhishi.me - 数据集 - 开
图书问答知识图谱 - 数
https://github.com/SilenceSnake/ASDKB
README.md
更新 README.md
3 years ago
阅读
CC-BY-SA-4.0 许可证
AsdKB：中文自闭症谱系障碍早期筛查与诊断知识库
(Text)
DSM-5
Diagnostic Standard
(Structured Knowledge)
SNOMED CT
The global language of healthcare
好大夫在线
https://m.haodf.com
国家统计局
National Bureau of Statistics
(Tables)
Physicians and Hospitals
家庭医生在线
familydoctor.com.cn
AsdKB
Diseases and Symptoms
ICD-10
(Text)
Intervention Information
Screening Tools
The National Clearinghouse on Autism Evidence & Practice
NCAEP • BRIDGING SCIENCE AND PRACTICE
(Tables)
ALSOLIFE
中国自闭症评估干预平台
CDC Centers for Disease Control and Prevention
OCALI
(Tables)
AUTISM CANADA AUTISME
发行作品
未发布任何版本
套餐
没有发布任何软件包

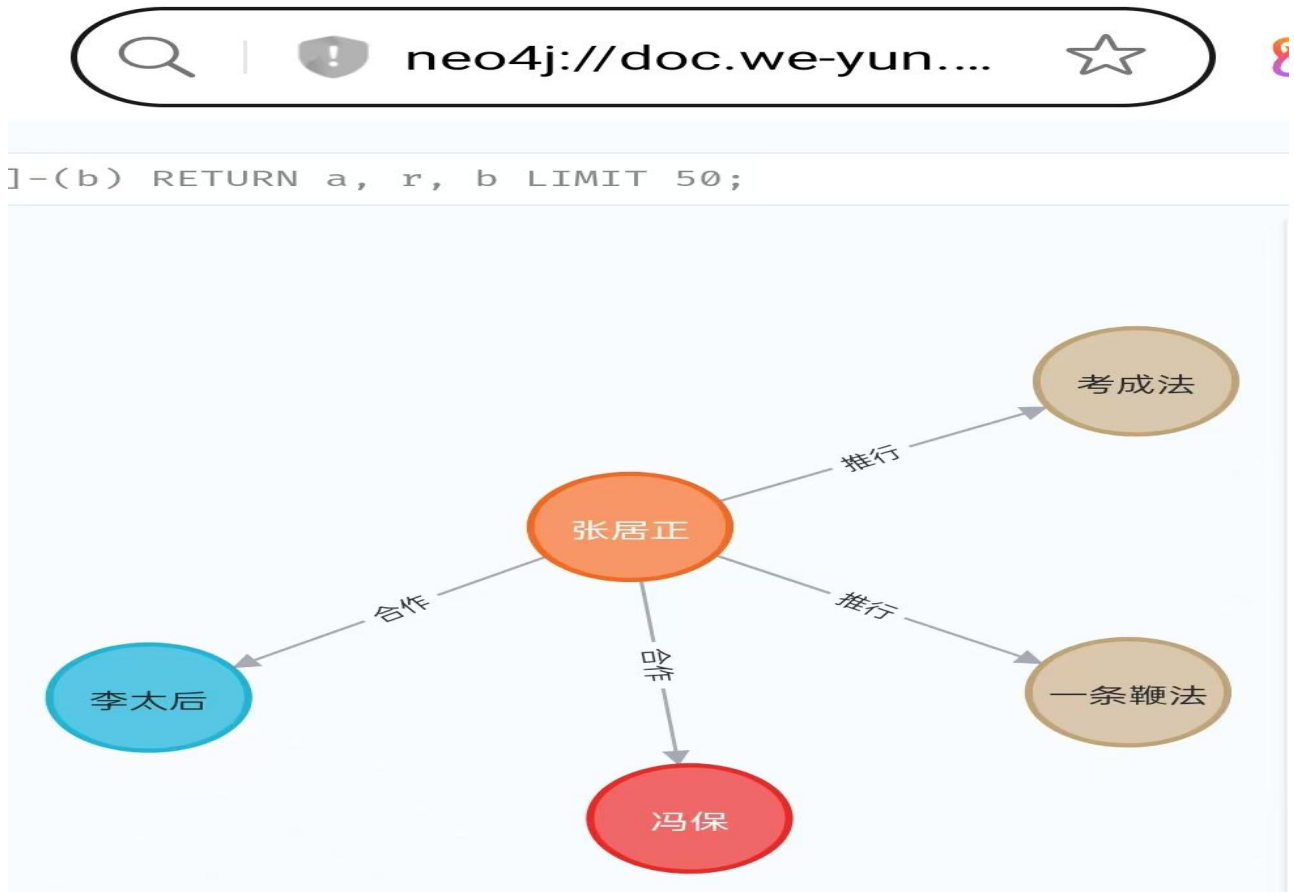
2


空白智能白板
最近保存 13:17
分享
https://docs.qq.com/board/DZkZvZ3ZTZnNWUHh6?aidPos=detail&no_promotion=1&is_blank_or_template=blank
明神宗
冯保
高拱
徐阶
君臣
政治同盟
政治对手
师生
张居正
著作
张太岳集
帝鉴图说
书经直解
67%

3


张居正知识图谱
帝鉴图说
徐阶
编撰（帝王治国启蒙）
受教于（恩师/提拔者）
神宗朱翊
辅佐（学生/君主）
编撰（为明神宗解读）
书经直解
张居正
著述（诗文奏疏合集）
政治对手（取而代之）
政治盟友（合作改革）
高拱
张太岳集
冯保

4(neo4j)


neo4j://doc.we-yun....
]-(b) RETURN a, r, b LIMIT 50;
考成法
推行
张居正
合作
推行
合作
李太后
一条鞭法
冯保