数据分析作业

信管2302 杨振凯 202321054054

第一讲 课程导言与分词

1、2

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

输入黄旭华，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。中文

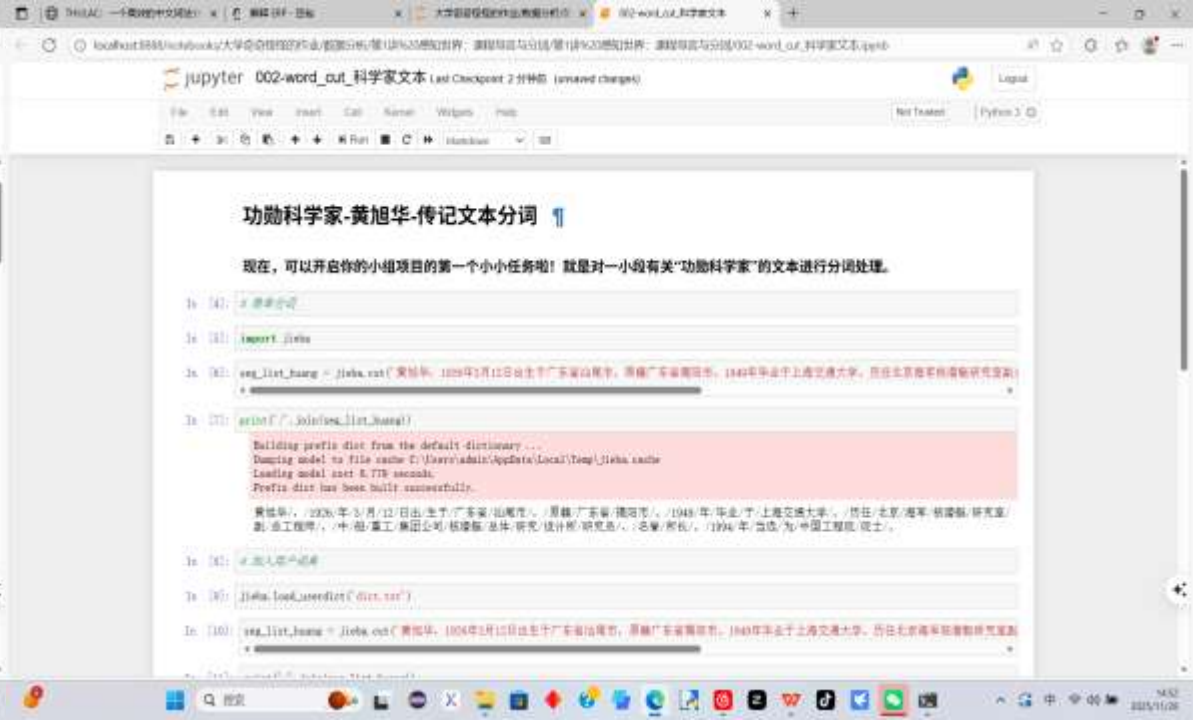
[测试 Try]

输入_v 黄旭华_np ，_w 1926年_t 3月_t 12日_t 出生_v 于_p 广东省汕尾市_ns ，_w 原籍_n 广东省_ns 揭阳市_ns 。_w 1949年_t 毕业_v 于_p 上海交通大学_ni 。_w 历任_v 北京_ns 海军_n 核潜艇_n 研究室_n 副总_j 工程师_n 、_w 中_f 船_n 重工_j 集团公司_n 核潜艇_n 总体_n 研究_v 设计所_n 研究员_n 、_w 名誉_n 所长_n 。_w 1994年_t 当选_v 为_v 中国_ns 工程院_n 院士_n 。_w 中文_n

In [1]:
```
print("hello world.hello 杨振凯")
```

hello world.hello 杨振凯

3




```
In [12]: # 2. 定义要统计的特殊词汇
target_words = ['数字化', '智能化', '安全']

In [14]: # 统计词频
word_counts = Counter(words)

In [15]: # 输出特定词汇的词频统计结果
print("特定词汇词频统计结果：")
for word in target_words:
    print(f"'{word}': {word_counts[word]}次")
```

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

```
In [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
```

```
    result = response.json()
    try:
        entities = result['choices'][0]['message']['content']
        print("提取到的实体和专业术语：")
        print(entities)
    except KeyError:
        print("无法解析API响应，原始响应：")
        print(result)
else:
    print(f"请求失败，状态码：{response.status_code}")
    print(response.text)
```

```
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",
    "免疫抑制信号通路"
  ],
  "方法": [
    "单细胞RNA测序",
    "scRNA-seq",
    "细胞亚群聚类",
    "轨迹分析",
    "pseudotime推断",
    "细胞间通讯网络构建",
    "动态识别方法"
  ],
  "工具": [
    "Seurat",
    "Monocle2",
    "CellChat"
  ],
  "生物学术语": [
```
```

4

阅读总结：

《基于关键词的学术文本聚类集成研究》针对学术文献海量增长（年增 3%）、人工分类耗时费力的现状，聚焦学术文本自动类别划分的迫切需求，核心探究了聚类集成是否能提升基于关键词的学术文本聚类性能、关键词抽取方法与个数对聚类集成效果的影响三大问题。研究采用 ACM CCS2012 数据集（含 3506 篇论文、40 个领域、8 个子集，人工标注确保准确性），通过 TF-ISF、CSI、ECC、TextRank 四种无监督方法抽取 5~60 个关键词，以 K-means 和增量聚类（IC）为基准，对比 ECKM（K-means 基础）、ECIC（增量聚类基础）两种聚类集成方法的性能，采用准确率、召回率及 F1 值作为评估指标。实验结果显示，聚类集成方法性能显著高于基准方法（T 检验 P<0.001）且稳定性更强，其中 TextRank 关键词抽取效果最佳（F1 值> 0.5）、CSI 效果最差（F1 值 < 0.4），且关键词个数越多聚类性能越好，关键词较少时集成方法优势更突出。综上，聚类集成优于单一聚类，学术文本自动分类应优先选用 TextRank 抽取关键词并尽可能抽取较多关键词；未来可扩展中文学术文本验证，并结合关键词与引文信息进一步提升性能。

基于关键词的学术文本聚类集成研究

张颖怡[1,2]，章成志[1,2]，陈　果[1]

（1. 南京理工大学信息管理系，南京　210094；2. 中国科学技术信息研究所，北京　100038）

摘　要　文本聚类是一种无监督且高效的文本类别划分方法。从文本中抽取的关键词代表了文本主旨内容，基于关键词的文本聚类是当下主流方式之一。在学术文本聚类研究中，主要使用单一的聚类方法。目前，一部分提升聚类性能的方法被提出，聚类集成是其中之一。因此，根据聚类集成思想，本文开展了基于关键词的学术文本聚类研究。为分析聚类集成在学术文本聚类中的有效性，本文比较了非集成聚类算法与聚类集成算法的性能。同时，为分析关键词对聚类集成性能的影响，本文分析了不同关键词抽取方法和不同关键词个数下学术文本的聚类结果。实验结果表明，聚类集成算法能够提升学术文本聚类的性能。其中，当使用TextRank作为关键词抽取方法时，学术文本聚类结果较佳；随着关键词个数的增加，学术文本类别划分性能随之提升。

关键词　关键词抽取；文本聚类；主题划分；聚类集成

第二讲 词频统计

1

近 10 年（2014-2024 年），基于 CNKI 数据库文献的统计与主题挖掘（如 BERTopic 模型分析、高频关键词统计、跨学科标签占比测算等）显示，“信息资源管理” 主题演化呈现从 “传统单学科坚守” 到 “数智化跨学科融合”、从 “基础管理导向” 到 “战略应用导向” 的核心趋势，可按三个阶段清晰划分：

一、2014-2020 年：传统领域深耕与跨学科萌芽期

此阶段信息资源管理主题以传统图情档领域基础研究为核心，同时伴随跨学科研究的初步探索，整体呈现 “守正为主、创新为辅” 的特征。

二、2021-2023 年：学科更名驱动与技术赋能转型期

2022 年教育部将 “图书情报与档案管理” 一级学科更名为 “信息资源管理”（2023 年正式实施），叠加元宇宙、生成式 AI 等技术爆发，主题演化进入 **“学科扩容 + 技术驱动” 双轮驱动阶段 **，CNKI 文献呈现显著的 “新旧融合” 特征。

三、2024 年：数智化深化与战略导向聚焦期

基于 CNKI 2024 年文献的 BERTopic 模型聚类与热点评选（如 “2024 年度十大学术热点”），主题演化进入 **“技术深度渗透 + 国家战略响应” 的高质量发展阶段 **，核心主题高度聚焦 “数智” 与 “战略” 双维度。

2

```
In [1]: import jieba

In [2]: 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

In [3]: dele = {'，','：','；','？','的',' ','“','”','（','）',' ','》','《','。'} # 手动设计一些停用词和标点

In [4]: jieba.add_word('諸葛亮') # 加入字典中没有的新词
```

```
Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\admin\AppData\Local\Temp\jieba.cache
Loading model cost 0.527 seconds.
Prefix dict has been built successfully.
```

```
In [5]: words = list(jieba.cut(article)) #返回分词出来的词汇

In [6]: words
```

```
['偃月刃',
 '，',
 '又',
 'named',
 '於',
 '滿',
 '，',
 '滿斬',
 '自',
 '高祖',
 '\n',
 '斬',
 '白蛇',
 '而',
 '起義',
 '，',
```

```
In [1]: f_name = open('name.txt',encoding = 'GB18030') #使用mac的小伙伴，需要耐心调试下编码GB18030

In [2]: #f_name = open("name.txt")

In [3]: data_name = f_name.read()

In [4]: data_name[:70]
Out[4]: '諸葛亮|關羽|劉備|曹操|孫權|關羽|張飛|呂布|周瑜|趙雲|龐統|司馬懿|黃忠|馬超'

In [5]: print(data_name[:50])
諸葛亮|關羽|劉備|曹操|孫權|關羽|張飛|呂布|周瑜|趙雲|龐統|司馬懿|黃忠|馬超

In [6]: f_name.close()

In [7]: # 将文本转化为列表

In [8]: names = data_name.split('|') # split一下names就是列表

In [10]: print(names)
['諸葛亮', '關羽', '劉備', '曹操', '孫權', '關羽', '張飛', '呂布', '周瑜', '趙雲', '龐統', '司馬懿', '黃忠', '馬超']

In [11]: names
Out[11]: ['諸葛亮',
          '關羽',
          '劉備',
          '曹操',
          '孫權',
          '關羽',
          '張飛'
```

```
In [28]: f_weapon = open('weapon.txt',encoding = 'utf-8')

In [29]: data_weapon = f_weapon.read()

In [30]: print(data_weapon[:100])
```

青龍偃月刀
丈八點鋼矛
鐵脊蛇矛
涯角槍
諸葛槍
方天畫戟
長柄鐵錘
鐵蒺藜骨朵
大斧
藤金斧
三尖刀
截頭大刀
馬岱寶刀
古錠刀

3

```
In [1]: import jieba

In [2]: 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

In [3]: dele = {'。','！','？','的','“','”','（','）',' ','》','《','，'} # 手动设计一些停用词和符号

In [7]: jieba.add_word('国立交通大学') # 加入字典中没有的新词

In [8]: words = list(jieba.cut(article)) #结巴分词出来的词汇

In [9]: words

Out[9]: ['在',
 '核潜艇',
 '领域',
 '，',
 '我国',
 '已',
 '形成',
 '一套',
 '完整',
 '的',
 '研究',
 '、',
 '设计',
 '、',
 '试验',
 '、',
 '制造',
 '、',
 '测试',
```

```
In [6]: f_txt = open('科学家博物馆-黄旭华传记序言.txt',encoding = 'utf-8')

In [7]: data_txt = f_txt.read()

In [8]: f_txt.close()

In [10]: print(data_txt[:1000])
```

在核潜艇领域，我国已形成一套完整的研究、设计、试验、制造、测试的核潜艇产业体系，而且装备了一支具有极高战略威慑力的、成梯次配备的、已近实现战备巡逻的核潜艇部队。回顾我国核潜艇的发展历程，人们自然会想起以黄旭华为代表的五位两院院士及无数第一代核潜艇研制人员的皓首穷经、筚路蓝缕、无私奉献。正是他们所铸就的国之重器使我国彻底摆脱了超级大国的核讹诈，更使我们在民族复兴的道路上迈出了坚实的一步。

而今，由黄旭华院士等人所开创的核潜艇工程以令世人震撼的力量，继续承载着捍卫“中国梦”的伟大重任。

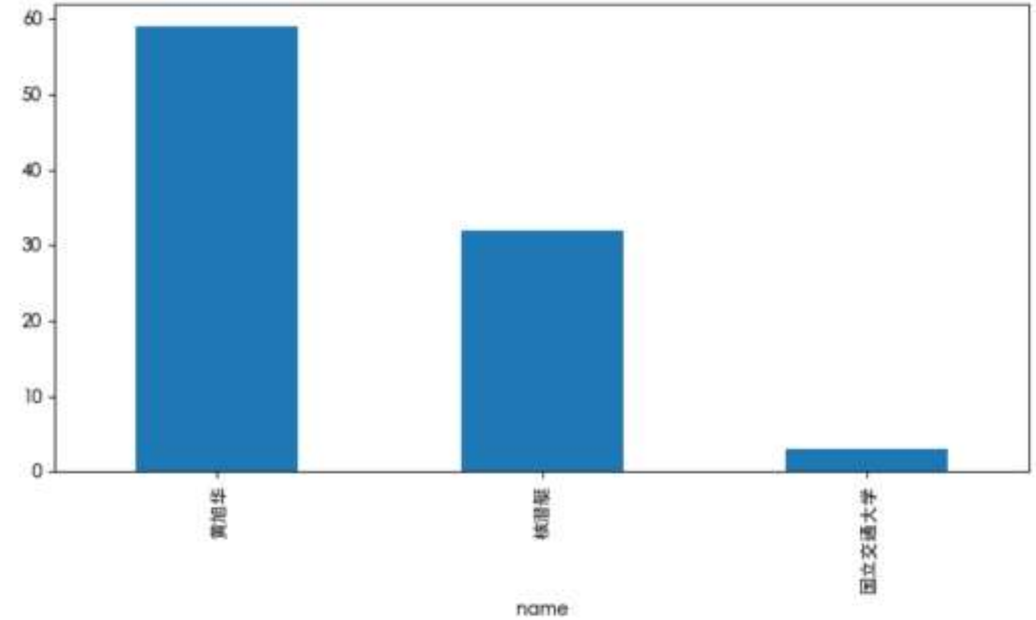
黄旭华是我国著名船舶专家、核潜艇研究设计专家、中国工程院首批院士、中国第一代核动力潜艇研制创始人之一。1924年2月24日，黄旭华出生于广东省汕尾市海丰县田墘镇，原籍广东省揭阳县。1949年，他毕业于国立交通大学造船系船舶制造专业，先后从事过民用船舶和军用舰艇的研究设计工作。1958年，黄旭华开始参与并领导我国第一代核潜艇的研究设计工作，先后出任第一代核潜艇副总设计师、第二任总设计师，历任中国船舶工业总公司及中船重工集团公司第七一九所副总工程师、副所长、所长、党委书记。黄旭华先后于1978年获全国科学大会奖、1982年获国防科工委二等奖，1986年被授予船舶工业总公司劳动模范，1989年被授予全国先进工作者，他参与完成的我国第一代核潜艇研制获1985年国家科学技术进步奖特等奖、导弹核潜艇研制获1996年国家科学技术进步奖特等奖。

黄旭华出生于以医为主、兼理农商之家，正直、勇敢、仁厚、坚毅的父母自小给予了他良好的道德与文化的熏陶。在历经了树基小学、作矶小学、聿怀中学、广益中学、桂林中学、教育部特设大学先修班的坎坷求学历程之后，他以优异的成绩进入了当时著名的国立交通大学，系统学习造船专业理论与技术，以期实现“科学强国”的报国理想。同期在地下党的培养下，历经风雨的洗礼成长为一名坚强的共产党员。

新中国成立后，经过党校系统培训学习，黄旭华在政治思想上逐步成熟。经过苏联军事舰船的转让仿制的锤炼，黄旭华在专业技术上也崭露头角。1958年，黄旭华因为政治素质过硬、专业技术精湛，成为开启“09工程的最初29位专业技术人员之一。从此将自己的一生献给了祖国的核潜艇事业。在核潜艇的研制过程中，黄旭华秉持“自力更生、艰苦奋斗、大力协同、无私奉献”的核潜艇精神，倡导以常规技术系统集成的科学理念，克服重重困

```
In [17]: %matplotlib inline

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



4

阅读总结：
该研究借助 Google Books（3600 万册数字图书）和 Google Scholar（9100 万篇学术文献）的语料库，以牛顿、爱因斯坦为核心案例，结合 234 位顶尖物理学家的样本，通过姓名提及频率统计、共现分析及多语言对比，探究了物理学家的科学声誉演化规律。研究发现，伟大科学家的声誉可跨越数世纪持续存在，且存在 “群体内偏好”—— 母语或本国语境中科学家更易获得认可（如英式英语中牛顿声誉长期高于爱因斯坦，美式英语和德语中爱因斯坦更受关注）；学术领域内，1948 年后爱因斯坦的声誉超越牛顿，牛顿的核心关联成就为万有引力定律和运动定律，爱因斯坦则以相对论（28%）和量子理论（16.9%）为主要标签；此外，Google Books 可作为替代计量工具（altmetrics），弥补传统引文计量难以衡量学术外社会影响的不足，但研究也存在语料库语言偏见、姓名歧义、无法区分提及性质等局限性。

Journal of Informetrics 12 (2018) 1089–1098

Contents lists available at ScienceDirect

Journal of Informetrics

journal homepage: www.elsevier.com/locate/joi

Long live the scientists: Tracking the scientific fame of great minds in physics

Guoyan Wang [a], Guangyuan Hu [b], Chuanfeng Li [c], Li Tang [d,*]

[a] Department of Science and Technology Communication and Policy, University of Science and Technology of China, 96 Jinzhai Road, Hefei, Anhui Province, 230026, China
[b] School of Public Economics and Public Administration, Shanghai University of Finance Economics, 777 Guoding Road, Shanghai, 200433, China
[c] Key Lab of Quantum Information, University of Science and Technology of China, 96 Jinzhai Road, Hefei, Anhui Province, 230026, China
[d] School of International Relations and Public Affairs, Fudan University, 220 Handan Road, Shanghai, 200433, China

ARTICLE INFO

Article history:
Received 15 February 2018
Received in revised form 21 August 2018
Accepted 23 August 2018

Keywords:
Scientific fame
Own-group preference
Google corpus
Altmetrics

ABSTRACT

This study utilizes global digitalized books and articles to examine the scientific fame of the most influential physicists. Our research reveals that the greatest minds are gone but not forgotten. Their scientific impacts on human history have persisted for centuries. We also find evidence in support of own-group fame preference, i.e., that the scientists have greater reputations in their home countries or among scholars sharing the same languages. We argue that, when applied appropriately, Google Books and Ngram Viewer can serve as promising tools for altmetrics, providing a more comprehensive picture of the impacts scholars and their achievements have made beyond academia.

Some say that a man dies three times. The first time is when his heart stops beating and he dies physically. The second is when people come to his funeral and his identity is erased from society. The third time is when nobody on the earth remembers him anymore. Then he is really dead.

"Dragon Raja" by Lee Yeongdo

1. Introduction

图书对人类历史与文明的档案价值

Books are the stepping stones to human progress. According to UNESCO (United Nations Educational, Scientific and Cultural Organization), the number of estimated published books in 2017 alone is up to 2.2 million.[1] Such a large collection is undoubtedly a rich archive of human history and civilization. Yet, as one of the most telling embodiments of knowledge stock and advancement, books have not captured sufficient attention in quantitative research evaluation.

Fortunately, with access to Google Books and the Google Books tool Ngram Viewer, scholars are now able to trace cultural evolution on a long time scale based on digitalized texts and trillions of words. This application of high-throughput data collection to study human culture can be traced back to Michel et al.(2011). In this pioneering study, the authors utilized the Google Books corpus and conducted text-based statistical analysis to trace cultural trends. That innovative research method soon captured academia's attention and was adopted in the arenas of digital history (Sternfeld, 2011), the history of science

研究图书的开始science，以及工具

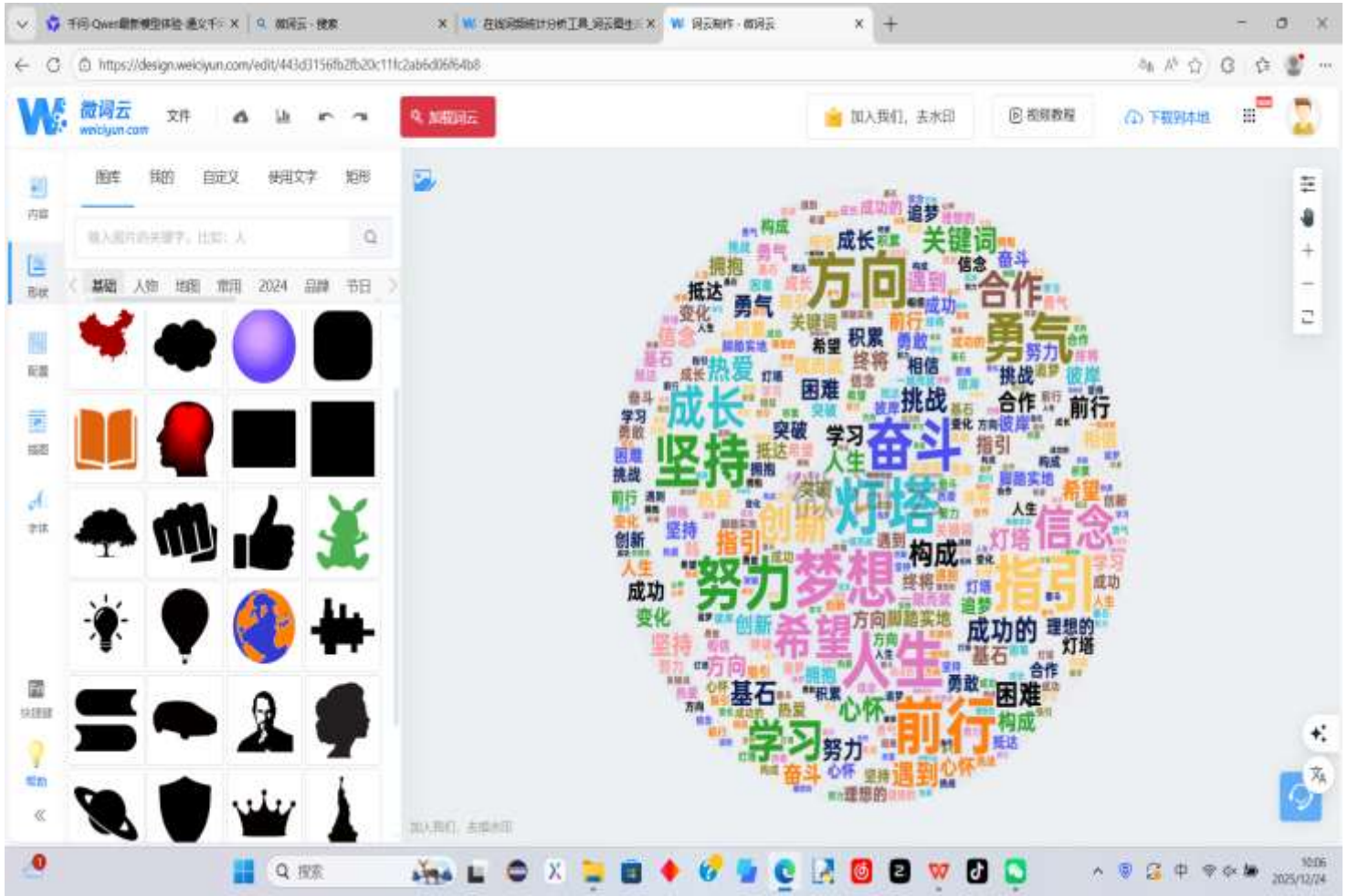
* Corresponding author.
E-mail address: litang@fudan.edu.cn (L. Tang).

[1] Data source: http://www.worldometers.info/books/. Accessed on January 18, 2018.

https://doi.org/10.1016/j.joi.2018.08.008
1751-1577/

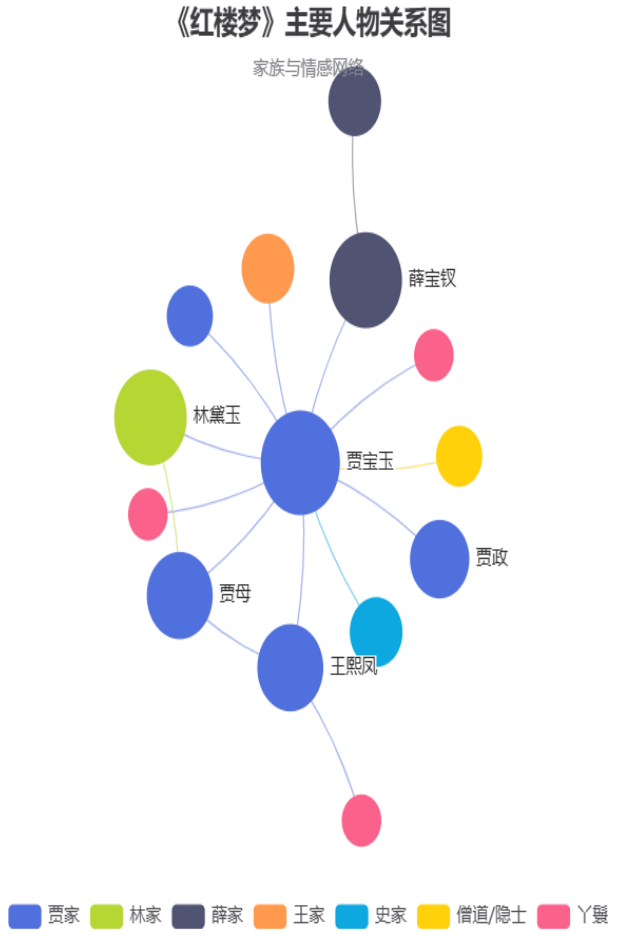
第三讲 词云与可视化

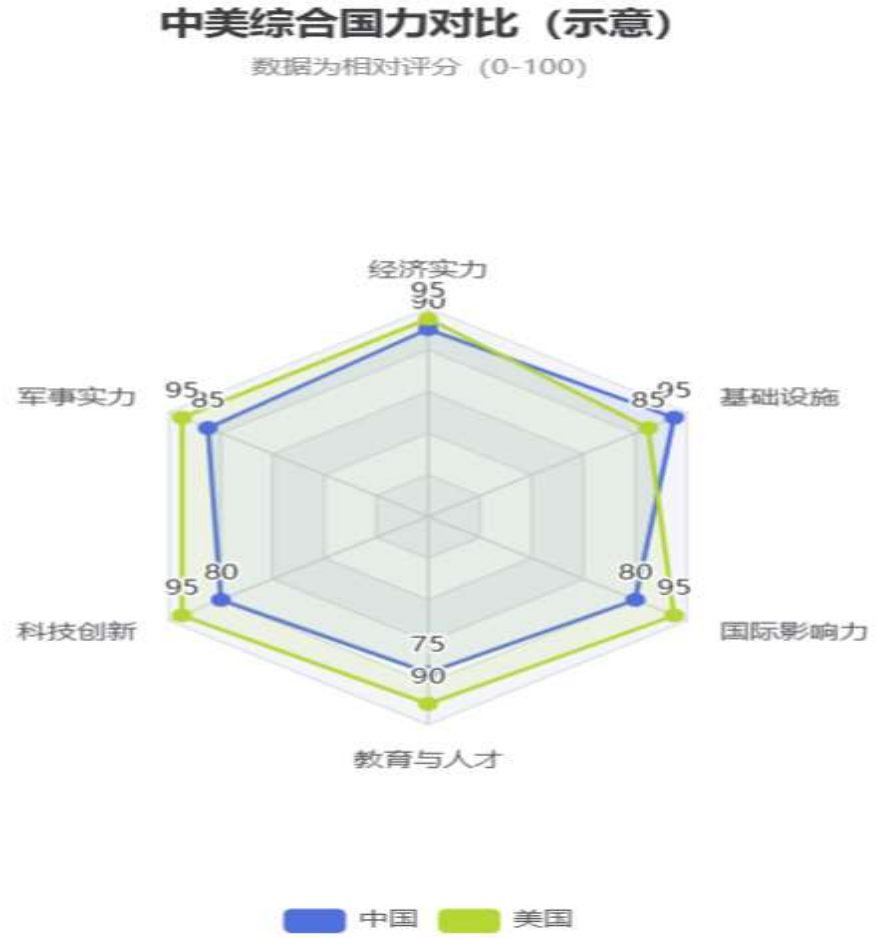
1.

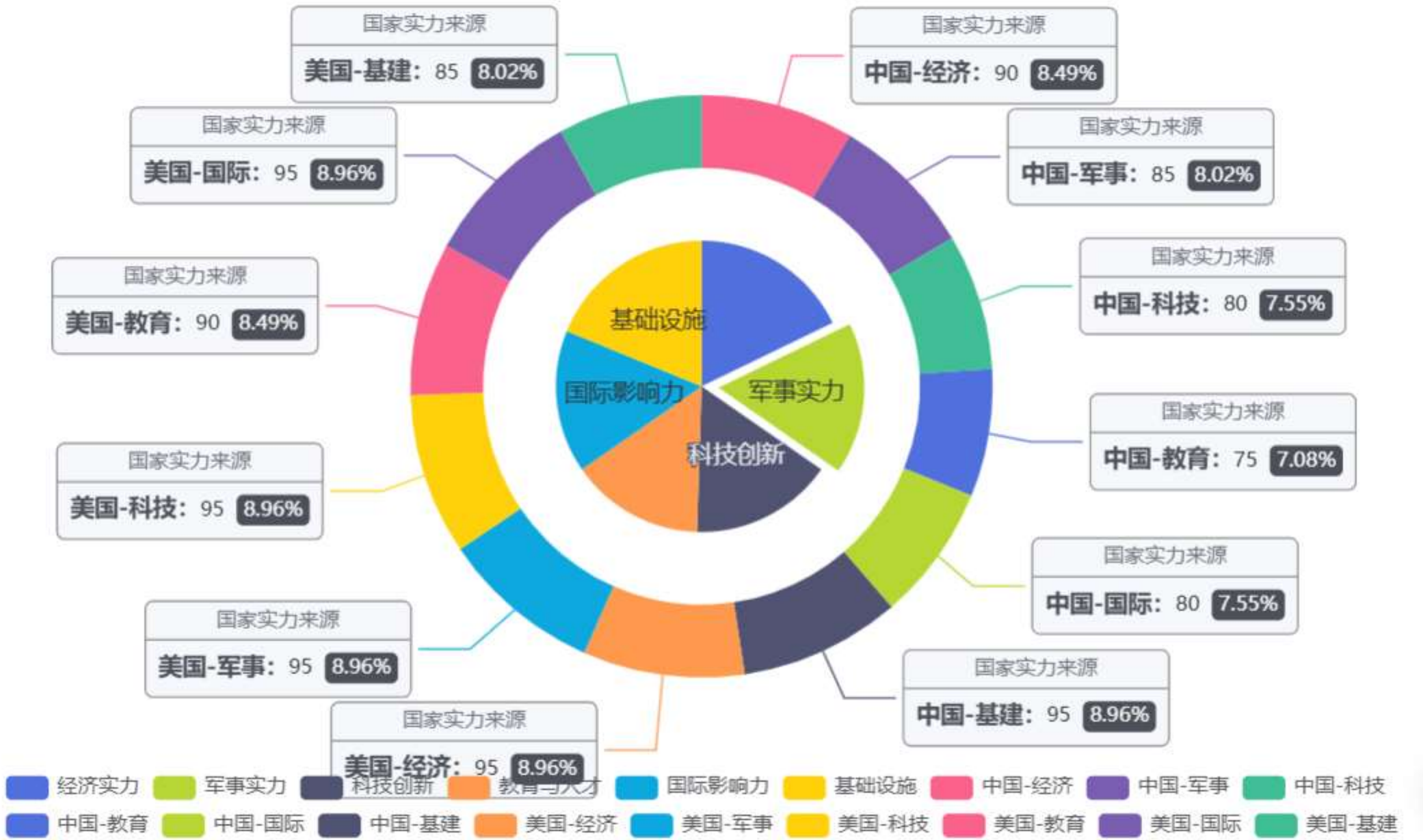


这张词云图以球形呈现，通过“成长”“坚持”“奋斗”“梦想”“希望”“努力”“创新”等高频关键词，生动传达出积极向上、脚踏实地、勇敢追梦的人生态度，强调在挑战中坚守初心、在合作与积累中实现突破，整体洋溢着励志与信念的力量。

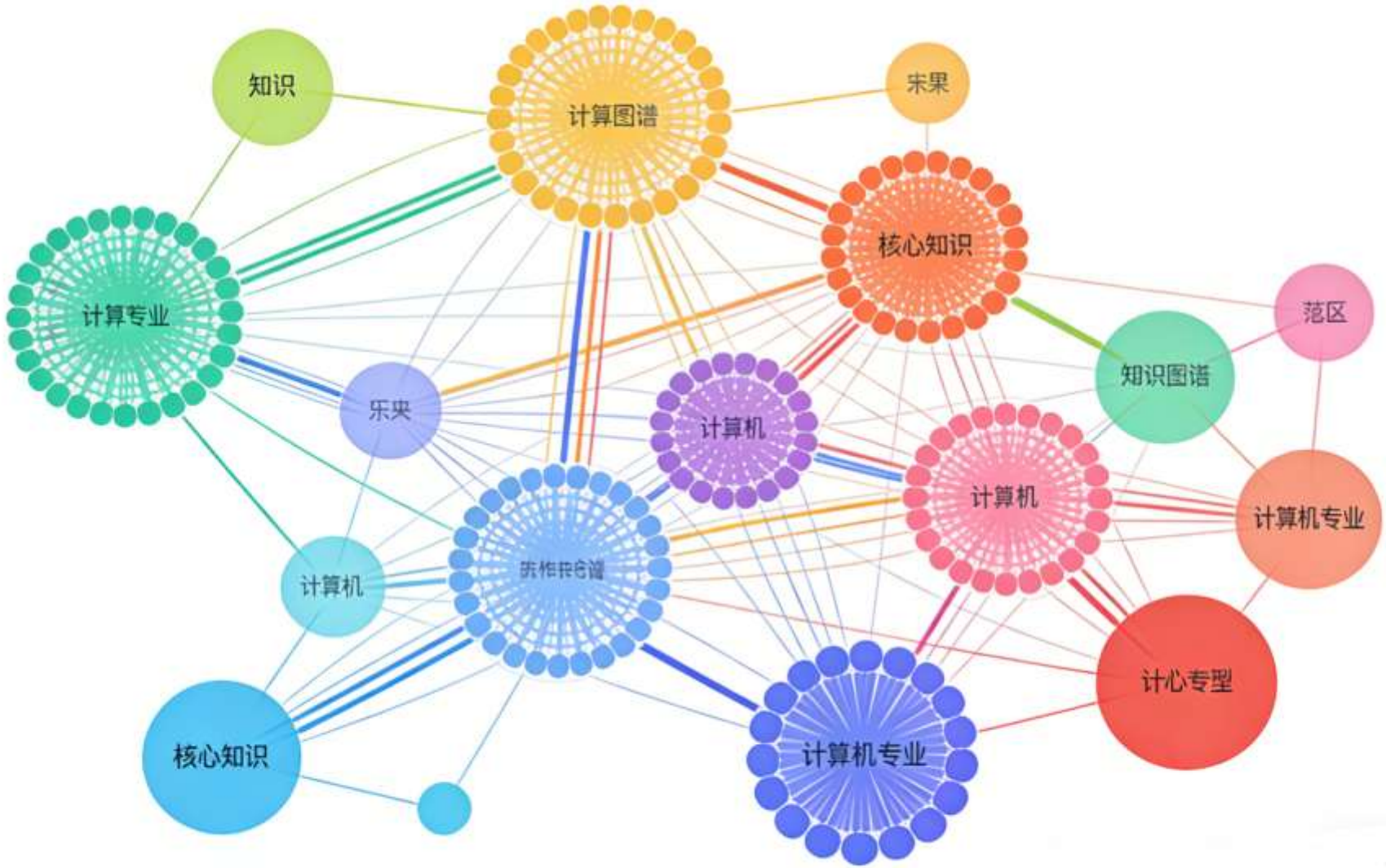
2.


《红楼梦》主要人物关系图
家族与情感网络
薛宝钗
林黛玉
贾宝玉
贾政
贾母
王熙凤
贾家
林家
薛家
王家
史家
僧道/隐士
丫鬟


中美综合国力对比（示意）
数据为相对评分（0-100）
经济实力
基础设施
国际影响力
教育与人才
科技创新
军事实力
95
85
95
80
95
75
90
95
85
80
中国
美国


国家实力来源
美国-基建：85 8.02%
国家实力来源
中国-经济：90 8.49%
国家实力来源
美国-国际：95 8.96%
国家实力来源
中国-军事：85 8.02%
国家实力来源
美国-教育：90 8.49%
国家实力来源
中国-科技：80 7.55%
国家实力来源
美国-科技：95 8.96%
国家实力来源
中国-教育：75 7.08%
国家实力来源
美国-军事：95 8.96%
国家实力来源
中国-国际：80 7.55%
国家实力来源
美国-经济：95 8.96%
国家实力来源
中国-基建：95 8.96%
基础设施
国际影响力
军事实力
科技创新
经济实力
军事实力
科技创新
教育与人才
国际影响力
基础设施
中国-经济
中国-军事
中国-科技
中国-教育
中国-国际
中国-基建
美国-经济
美国-军事
美国-科技
美国-教育
美国-国际
美国-基建

3


知识
计算图谱
宋果
计算专业
核心知识
范区
知识图谱
乐央
计算机
计算机
计算机专业
计算机
核心知识
计心专型
计算机专业


黄旭华
采集
资料
工作
核潜艇
专业
进行
学术
成长
我国
年
所
工程
核潜艇
研制
曲折
交通
大学
重点
要求
特点
都
院士
轨迹
传记
研究
设计
介绍
按照
直接
照片
采集
小组
成就
中国
精神
求学
过程
历史
第一代
核潜艇
完整
完成
撰写
贡献
思想
他
为
依据
代表
提升
处理
成长
访谈
人生
主要
技术
一生
传主
还原
黄旭
国立
交通
船舶
成熟
展示
实物
时间
包括
黄旭华
院士
档案
实施
实现
系统
关于
及其
其中
等
重要
任务
分析
李世英
使
地下党
自己
形成
本章
获
负责
人员
并
以
叙述
学习
有关
逐一
分为

第四讲 情感分析

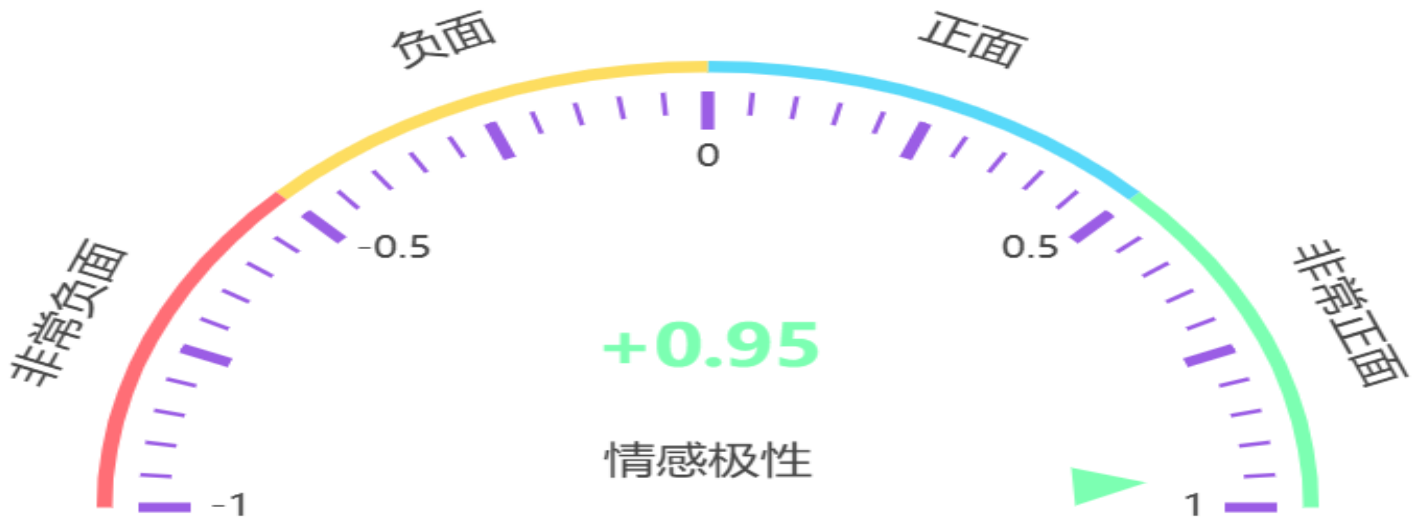
1.

“这是一部男人必看的电影。”人人都这么说。但单纯从性别区分，就会让这电影变狭隘。《肖申克的救赎》突破了男人电影的局限，通篇几乎充满令人难以置信的温馨基调，而电影里最伟大的主题是“希望”。当我们无奈地遇到了如同肖申克一般囚禁了心灵自由的那种囹圄，我们是无奈的老布鲁克，灰心的瑞德，还是智慧的安迪？运用智慧，信任希望，并且勇敢面对恐惧心理，去打败它？ 经典的电影之所以经典，因为他们都在做同一件事——让你从不同的角度来欣赏希望的美好。

215/1000

情感分析

情感极性



sentiment_analysis_1_chuji

```
In [14]:
text = "I am happy today. I feel sad today."
from textblob import TextBlob
blob = TextBlob(text)

In [15]: # 果然不动的打印出来了？
         # 实际上已经把文本分成了句子了，看一看
         blob.sentences
Out[15]: [Sentence("I am happy today."), Sentence("I feel sad today.")]

In [16]: blob
Out[16]: TextBlob("I am happy today. I feel sad today.")

In [ ]:

In [17]: blob.sentences[0].sentiment
Out[17]: Sentiment(polarity=0.8, subjectivity=1.0)

In [18]: # 上面的结果什么意思呢？
         # 情感极性0.8，主观性1.0，说明一下，情感极性的变化范围是[-1, 1]，-1代表完全负面，1代表完全正面，
         # 我表达的是我很高兴，那么这个结果是对的

In [19]: blob.sentences[1].sentiment
Out[19]: Sentiment(polarity=-0.5, subjectivity=1.0)

In [20]: # 整段文本的情感呢？
         blob.sentiment
Out[20]: Sentiment(polarity=0.1500000000000002, subjectivity=1.0)
```

```
In [21]: test_cn = u"我今天很快乐，我今天很愤怒。"

In [22]: #注意这引号都是我们加了一个字母u，它很重要，因为它提示Python，“这一段我们输入的文本编码格式是Unicode，别搞错了哦”，关于文本编码格式的细节，有

In [23]: from snownlp import SnowNLP

In [24]: senti_cn = SnowNLP(test_cn)

In [25]: # 看看snownlp包的分句能力
         for sentence in senti_cn.sentences:
             print(sentence)
我今天很快乐
我今天很愤怒

In [26]: senti_cn_1 = SnowNLP(senti_cn.sentences[0])

In [27]: # 一个细节上的问题，英文是s.sentiment，中文是s.sentiments，多了一个s
         # 另外，这句话上和英文的处理有不同，比如建造新语句：senti_cn.sentences[0].sentiments是会报错的
         senti_cn_1.sentiments
Out[27]: 0.971889316039116

In [28]: senti_cn_2 = SnowNLP(senti_cn.sentences[1])

In [29]: senti_cn_2.sentiments
Out[29]: 0.07769313772213482

这里你肯定发现了问题——“愤怒”这个词表达了如此强烈的负面情感，为何得分依然是正的？

这是因为SnowNLP和textblob的计分方法不同。SnowNLP的情感分析取值，表达的是“这句话代表正面情感的概率”。也就是说，对“我今天很愤怒”一句，SnowNLP
认为，它表达正面情感的概率很低很低。

这样解释就是ok了

In [30]: senti_cn.sentiments
Out[30]: 0.7237619924203908
```

该 Notebook 实现了基于词典的情感分析初级方法：首先定义了包含正面和负面情感词的自建词典，然后通过计算文本中正负向词汇的加权得分（考虑否定词和程度副词的修饰作用）来判断整体情感倾向，并对示例评论进行测试，最终根据得分阈值输出“正面”、“负面”或“中性”的情感标签。

sentiment_analysis_2_timeline

In [37]:
```
plt.savefig('timeline.png') # 看不到？改一改？
```

```
<Figure size 432x288 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：

In [38]:
```
df.sort_values(['sentiments'])[:1]
```

Out[38]:

	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。

In [39]:
```
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，所以大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
```

该代码实现了一个完整的中文评论情感分析流程：首先读取包含评论与时间戳的CSV数据，利用SnowNLP对每条评论进行情感打分（0为负面，1为正面），然后绘制情感得分随时间变化的趋势图，并通过排序精准定位情感最负面的评论内容，从而结合自动化分析与人工解读，识别服务问题并提出改进建议。

sentiment_analysis_3_大模型_健康文本细粒度情感抽取

In [2]:
```python
import requests
import json

# DeepSeek API 端点
url = "https://api.deepseek.com/v1/chat/completions"

# 替换为您的 DeepSeek API 密钥
API_KEY = "sk-5a102bf40b204935afdf202dd12f7658" # 直接复制过来

# 请求头，包含 API 密钥和内容类型
headers = {
    "Authorization": f"Bearer {API_KEY}",
    "Content-Type": "application/json"
}
```

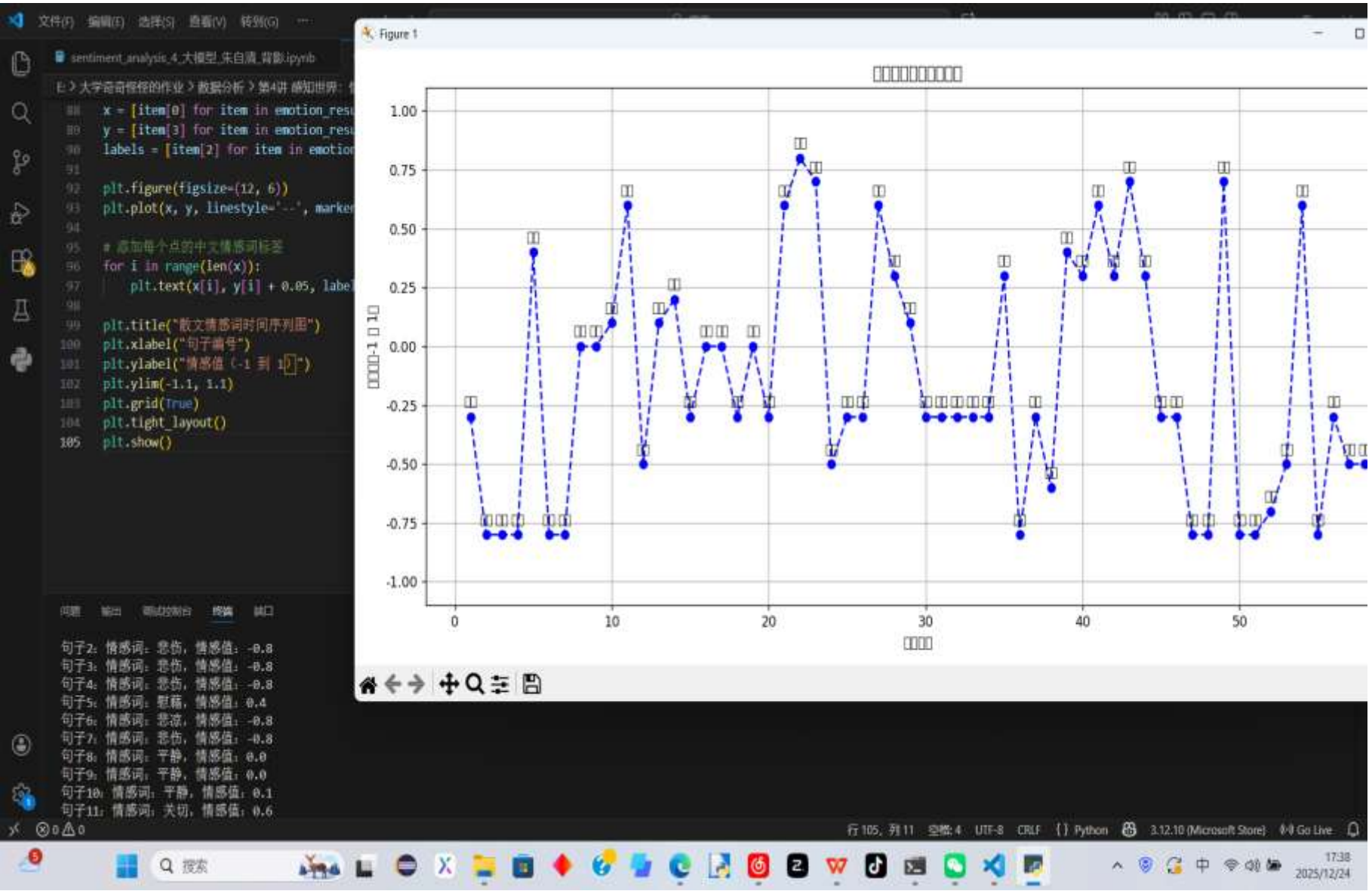
In [3]:
```python
# 患者描述文本
text = (
    "我今年58岁，退休后感到生活失去了重心，开始出现失眠、头痛和疲乏无力的症状。"
    "后来，皮肤变得异常敏感，连衣服的触碰都像针扎一样疼痛。"
    "我常常感到心慌、胸闷，背部沉重得像压了一块石头。"
    "对光线和声音变得极度敏感，电话铃声都会让我惊恐。"
    "多次到医院检查，结果都显示没有器质性疾病。"
    "我变得不愿出门，不想与人交流，整天把自己关在屋里，拉紧窗帘，感觉生活毫无意义。"
)

# 构建提示词，要求模型提取细粒度情感实体
prompt = (
    "请从以下患者描述中提取出具体的身体部位、症状以及对应的情感状态，"
    "并以 JSON 格式返回，格式如下："
    "{'实体': [{'部位': '...', '症状': '...', '情感': '...'}, ...]}\n\n"
    f"患者描述：{text}"
)
```

In [5]:
```python
# 请求体，包含模型参数和提示词
```

对健康领域的用户文本进行细粒度情感分析：首先定义了严格的JSON格式输出要求（包含情感极性、强度及原因），然后逐条调用模型对示例健康评论进行推理，解析返回结果，并最终将结构化的分析结果保存为JSON文件，实现了从原始文本到可量化、可解释情感洞察的自动化抽取流程。

sentiment_analysis_4_大模型_朱自清_背影



该代码读取朱自清《背影》全文，按句号和感叹号分句后，逐句调用 DeepSeek 大模型 API 进行情感分析，提取每句的情感词与 -1 到 1 的情感值，并通过正则表达式解析结果；最后利用 Matplotlib 绘制情感值随句子顺序变化的时间序列图，标注对应情感词，直观呈现散文中情感的细腻起伏。

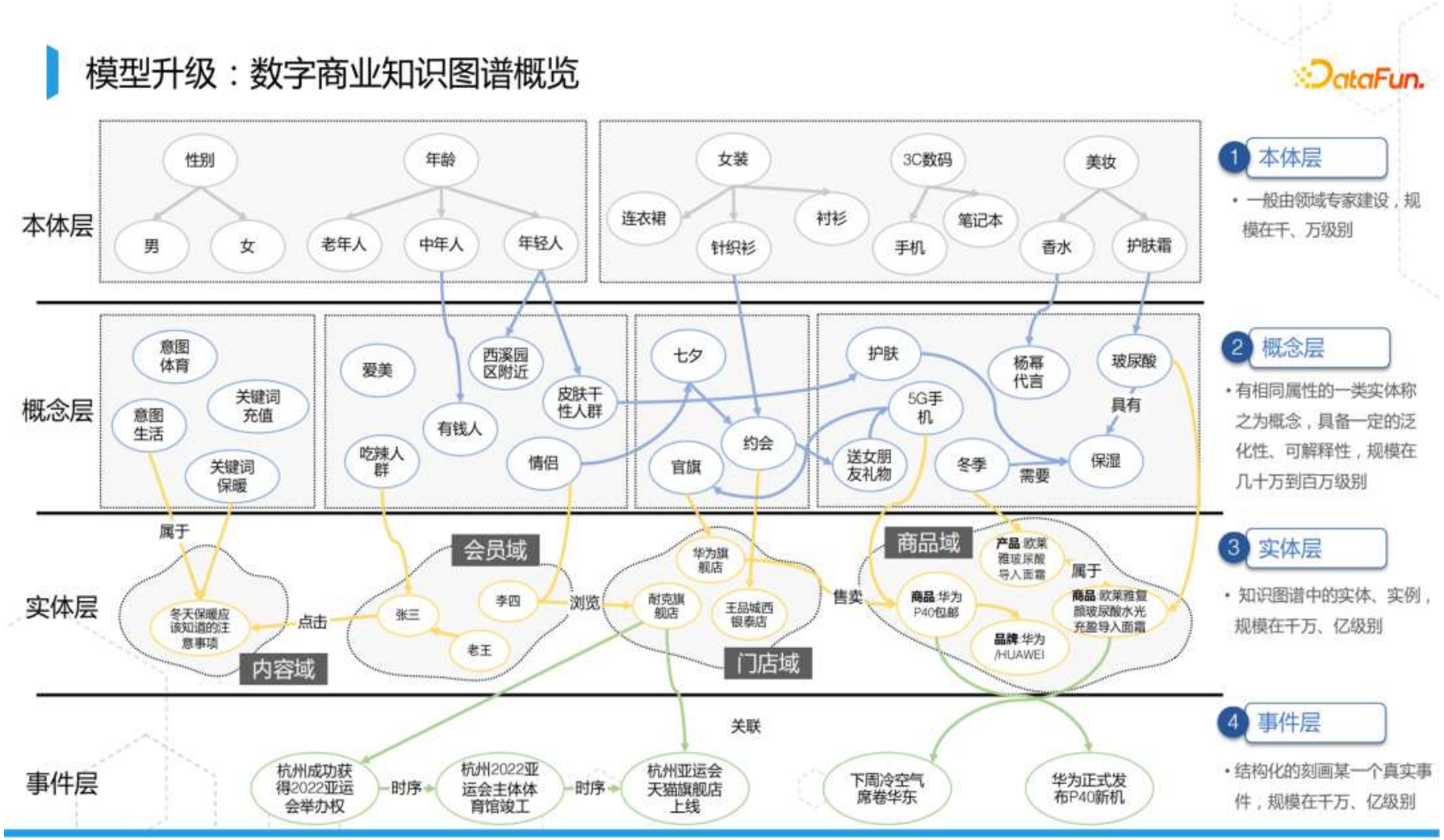
第六讲 知识图谱理念

阿里商品大脑

截至 2025年12月，阿里“商品大脑”已构建起一个以数字商业知识图谱为核心、多主体协同、全链路闭环的智能生态体系。这一生态不仅支撑平台治理与消费者体验，更深度赋能供给侧升级与全球贸易数字化。

一、知识图谱架构：四层融合体系

阿里商品大脑的知识图谱采用“本体–概念–实体–事件”四层融合架构



二、生态协同：开放共建 + 场景闭环

1. 产学研联合构建：
由阿里藏经阁知识引擎团队与浙江大学知识图谱实验室共同维护；
通过天池平台举办知识图谱竞赛（如“电商商品同款识别挑战赛”），吸引高校与开发者参与算法优化。

2. 产业侧深度嵌入
1688平台：商家上传商品时，系统自动基于图谱建议合规属性、热门关键词、竞品对标；
跨境出口：在义乌“数智贸易大模型”中，商品图谱自动映射海外合规标准（如FDA、CE、RoHS）；
品牌合作：联合欧莱雅、华为等品牌共建品牌专属子图谱，精准管控授权链路与防伪信息。

3. 消费者–商品–服务闭环
用户搜索“敏感肌可用的防晒”，图谱联动：
商品库（含“无酒精”“物理防晒”标签）；
内容库（小红书/逛逛评测）；
服务库（过敏包退、皮肤科咨询入口）。

三、技术底座：大模型 + 图谱 + 芯片三位一体

提升非结构化文本理解能力（如直播话术、评论情感）；支持零样本属性抽取

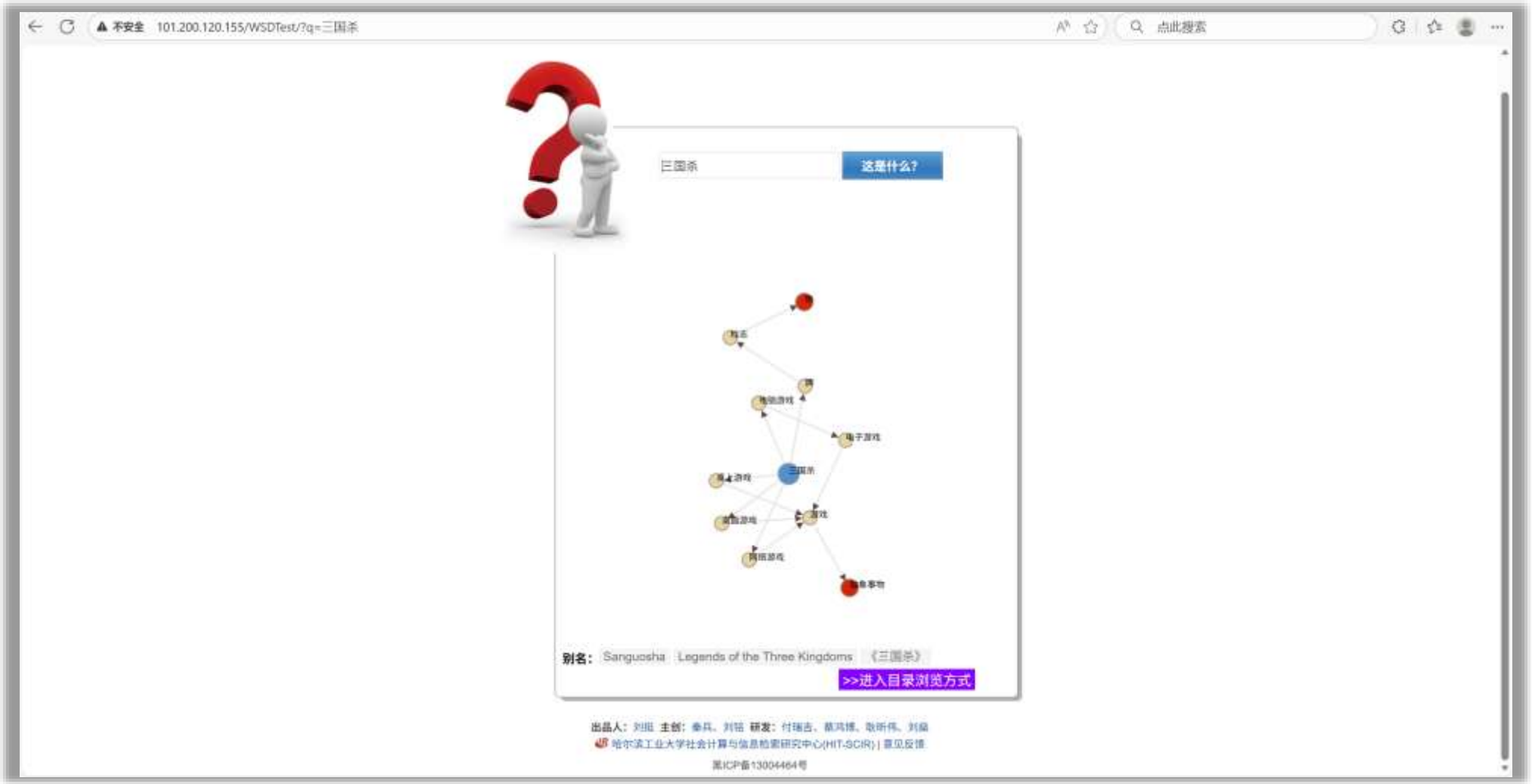
实现图文、视频、语音与商品实体的跨模态关联（如“视频说“抗老”→图谱打标“抗衰老”）

加速图谱推理与图像审核，单日处理超2亿张商品图

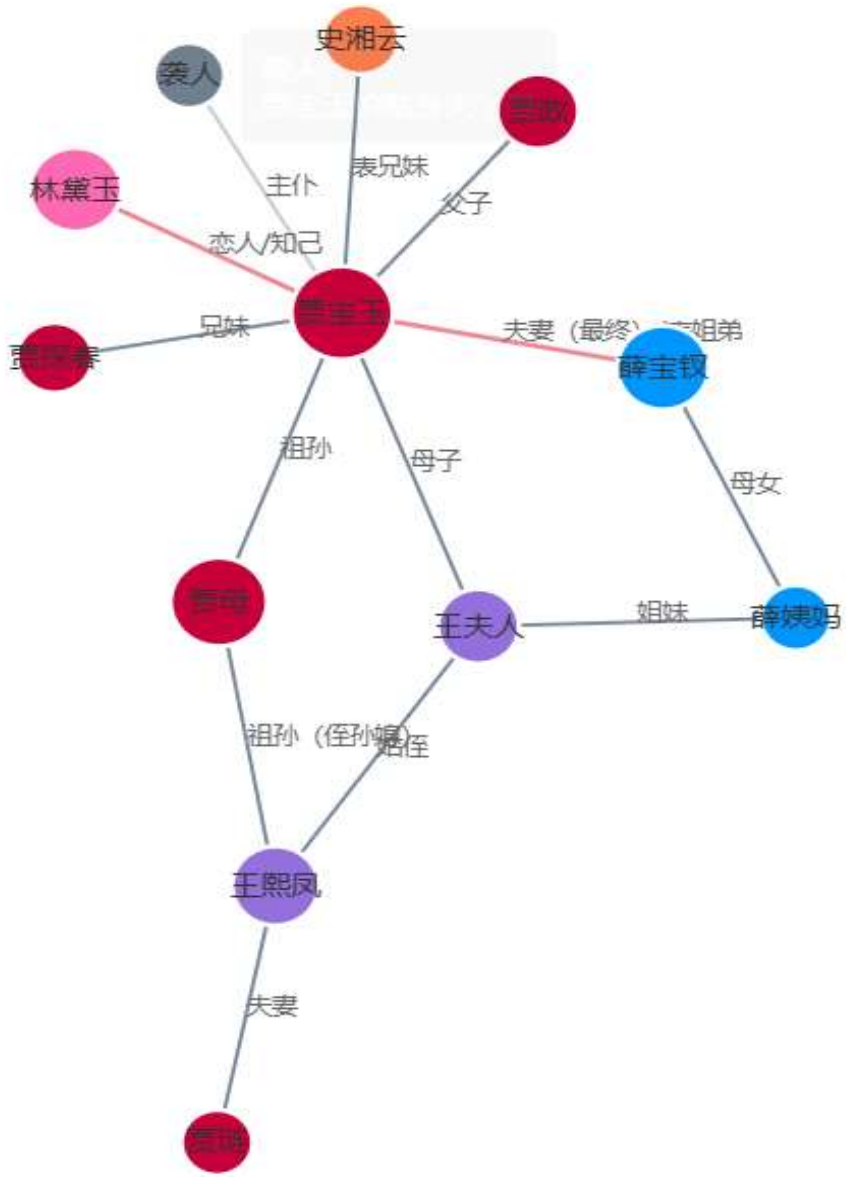
基于用户行为、舆情、政策实时更新图谱节点与关系权重

第六讲（2） 知识图谱工具

1

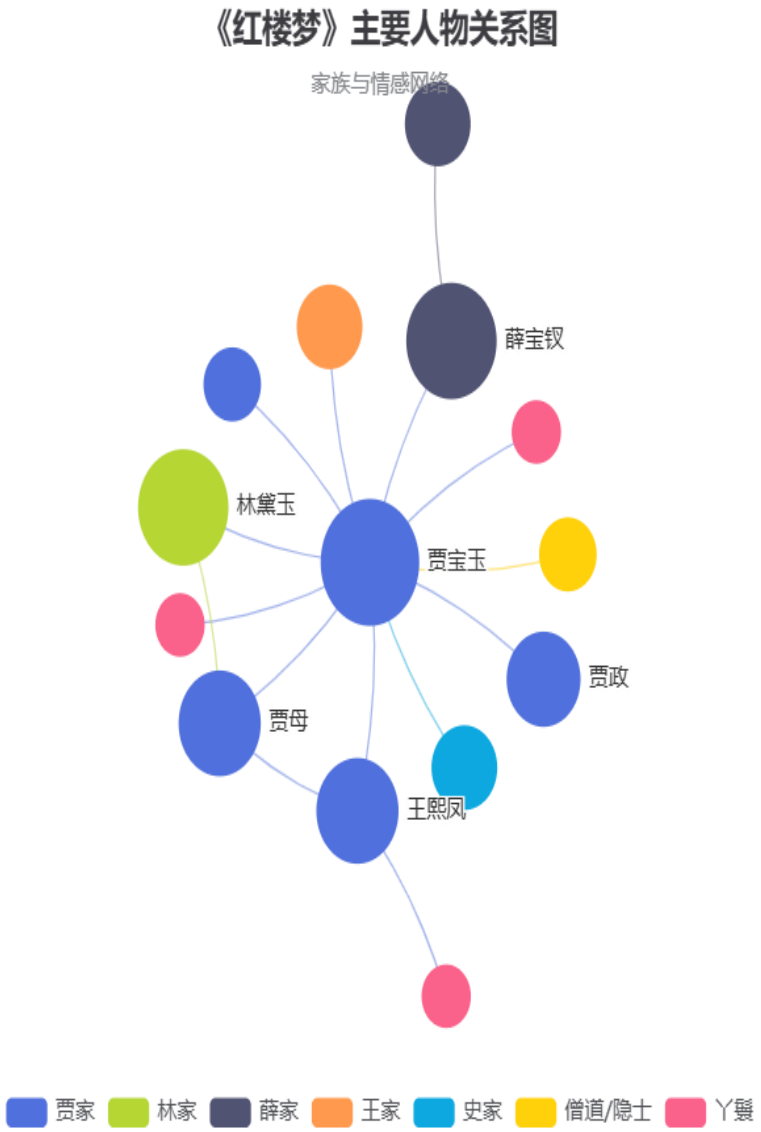

不安全 101.200.120.155/WSDTest/?q=三国杀
点此搜索
三国杀
这是什么?
别名: Sanguosha Legends of the Three Kingdoms 《三国杀》
>>进入目录浏览方式
出品人: 刘挺 主创: 秦兵、刘铭 研发: 付瑞吉、蔡鸿博、耿昕伟、刘扬
哈尔滨工业大学社会计算与信息检索研究中心(HIT-SCIR) | 意见反馈
黑ICP备13004464号

2

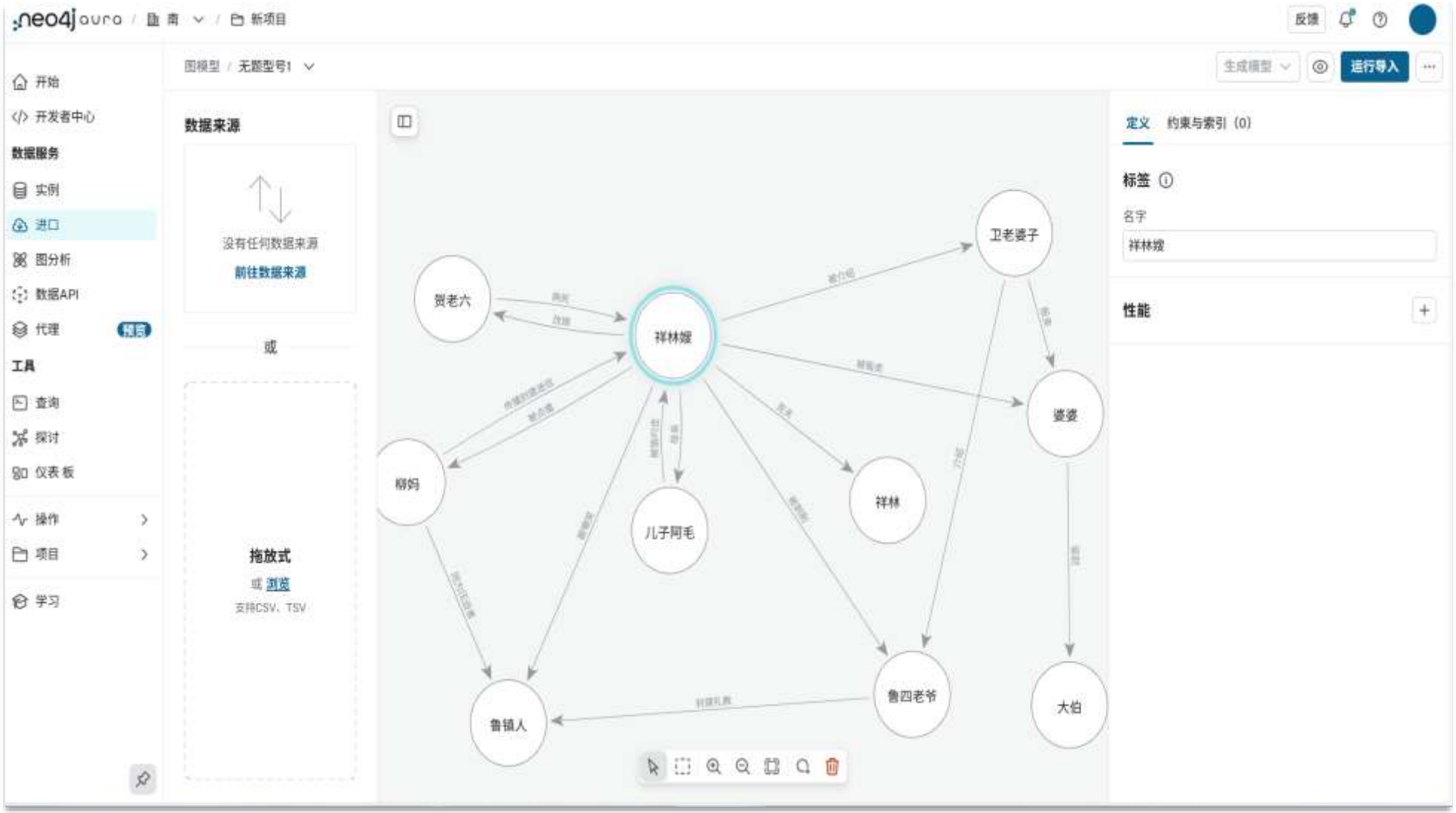


该图谱是基于 HTML 结合 D3.js 实现的《红楼梦》核心人物交互式力导向知识图谱，以节点对应贾宝玉、林黛玉、薛宝钗等十二位红楼关键人物，节点颜色按贾、林、薛、王、史氏家族及仆役身份清晰分类，节点大小对应人物在书中的核心重要程度，鼠标悬浮节点可查看人物核心身份简介；以边连接各节点呈现人物间关联，按情感、亲属、主仆三类核心关系差异化设计边的样式，红色粗线凸显贾宝玉与林、薛二人的情感关系，深蓝色线标注家族亲属关系，浅灰色半透明线区分主仆关系，且每条边均标注具体关系描述，同时采用力导向布局让图谱自动规整排布，支持手动拖拽节点调整位置避免重叠，直观清晰地展现了荣国府核心人物的网络关联，助力快速梳理红楼核心人物间的亲属羁绊、情感联结与身份依附关系。

3



4


neo4j aura
新项目
反馈
开始
开发者中心
数据服务
实例
进口
图分析
数据API
代理
工具
查询
探讨
仪表板
操作
项目
学习
图模型 / 无题型号1
生成模型
运行导入
数据来源
没有任何数据来源
前往数据来源
或
拖放式
或 浏览
支持CSV、TSV
定义
约束与索引 (0)
标签
名字
祥林嫂
性能
贺老六
祥林嫂
卫老婆子
婆婆
柳妈
儿子阿毛
祥林
鲁镇人
鲁四老爷
大伯