南京工業大學
NANJING TECH UNIVERSITY

用户数据采集与关联分析

（结课作业）

吴志祥

18205185639

1030624832@qq.com

第二讲 词频统计

1. 基于CNKI数据库统计分析2014-2024年（近10年），“信息资源管理”或“网络营销”或其他你感兴趣的主题变化趋势。

基于 CNKI 数据库 2014-2024 年文献统计，信息资源管理与网络营销领域均呈现技术驱动、需求导向的演化特征。信息资源管理从传统图书馆、档案数字化管理起步，经大数据时代的数据治理与智慧化转型，近年聚焦 AGI 应用、数据要素市场化等主题，政策导向与技术革新共同推动其从“资源有序化”迈向“价值赋能”。网络营销则从早期的渠道拓展与流量争夺，转向私域运营、数据驱动的精细化运营，如今依托 AIGC、跨平台联动实现技术赋能，且愈发注重与实体经济融合，凸显“长期价值创造”的核心逻辑。两大领域均呈现跨学科融合趋势，前者侧重公共服务与理论构建，后者贴近市场实践与商业价值，技术迭代始终是贯穿十年的核心演进动力。

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

本程序是基于jieba分词的全文本词频统计

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

```
In [1]: import jieba
In [2]: article = open('科学家博物馆-黄旭华传记作品.txt','r',encoding = 'utf-8').read() # 打开并读取二图前10份 #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 [4]: jieba.add_word('国立交通大学') # 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 0.818 seconds.
        Prefix dict has been built successfully.
In [5]: words = list(jieba.cut(article)) #结巴分词出来的词汇
In [6]: words
Out[6]: ['在',
         '特德耶',
         '领域',
         ',',
         '我国',
         '已',
         '形成',
         '一批',
         '完整',
         '的',
         '研究',
         '、',
         '设计',
         '、',
         '试验',
         '、',
         '制造',
         '、',
         '测试',
```

字典

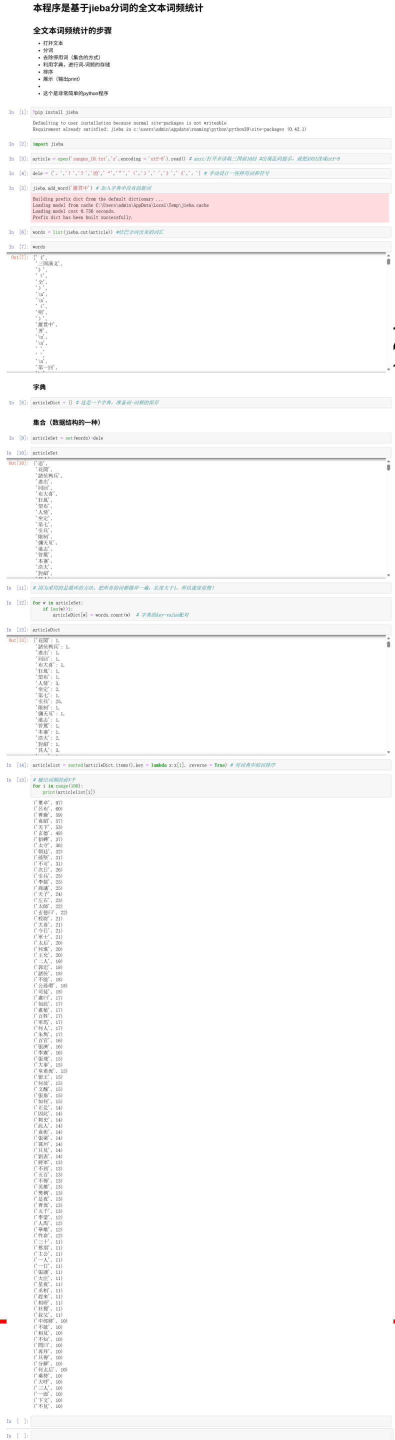
```
In [7]: articleDict = {} # 这是一个字典，准备词-词频的保存
```

集合（数据结构的一种）

```
In [8]: articleSet = set(words)-dele
In [9]: articleSet
        {'一直到',
         '一部分',
         '一项',
         七两',
         '上',
         '下',
         '下家',
         '下马',
         '不',
         '不仅',
         '不单是',
         '不得',
         '不足',
         '不过',
         '与',
         '专家',
         '专业技能',
         '专家',
         '世人',
         '两个',
In [10]: # 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！
In [11]: for w in articleSet:
             if len(w)>1:
                 articleDict[w] = words.count(w)  # 字典的key-value配对
In [12]: articleDict
Out[12]: {'首章': 1,
          '一名': 2,
          '高度': 1,
          '出版': 1,
          '1020': 1,
          '年事': 1,
          '理解': 1,
          '648': 1,
          '父母亲': 1,
          '书面': 1,
          '配合': 1,
          '齐心协力': 1,
          '以作': 2,
          '人生': 7,
          '献给': 1,
          '广泛': 1,
          '从事': 1,
          '严格遵守': 1,
          '真实': 1,
          '工作手册': 1,
          '协同': 1,
          '战术': 1,
          '大量': 1,
          '传主': 6,
          '清单': 5,
          '追踪': 2,
          '饭正': 1,
          '考虑': 1,
          '原点': 1,
          '根本原因': 1,
          '献身': 1,
In [13]: articlelist = sorted(articleDict.items(),key = lambda x:x[1], reverse = True) # 对词典中的词排序
In [14]: # 输出词频的前N个
         for i in range(100):
             print(articlelist[i])
         ('我国', 12)
         ('研制', 12)
         ('工程', 11)
         ('设计', 10)
         ('介绍', 8)
         ('主题', 8)
         ('第一代', 8)
         ('科学', 8)
         ('人生', 7)
         ('及其', 7)
         ('历史', 7)
         ('思想', 7)
         ('传记', 7)
         ('传主', 6)
         ('过程', 6)
         ('整成', 6)
         ('一生', 6)
         ('成就', 6)
         ('设计', 6)
In [ ]:
In [ ]:
In [ ]:
```

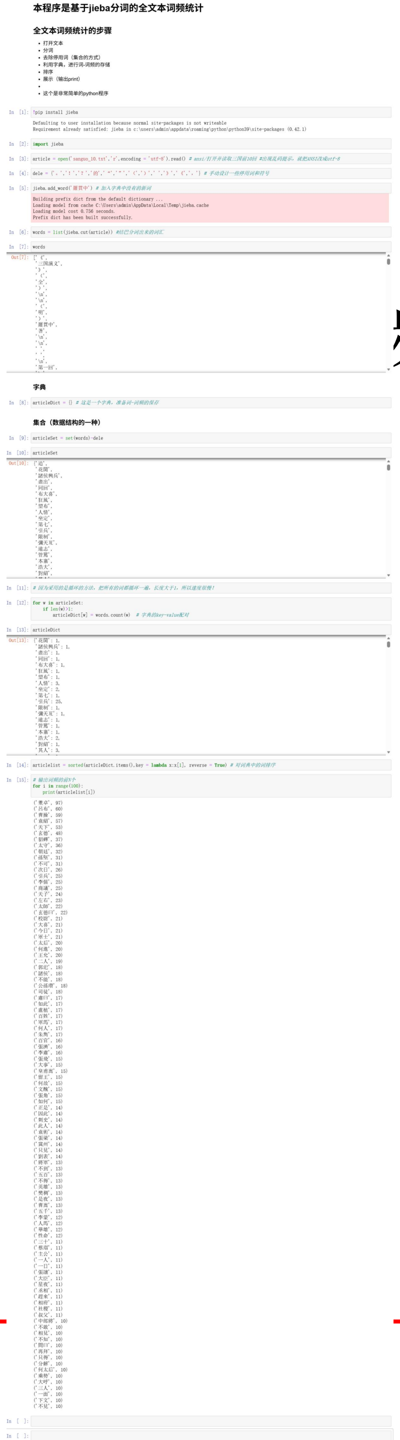
第二讲 词频统计

2. 完成ppt中的程序运行，包括全文词频统计，频统计；



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，



频统计；

第二

这部分是指定类型的频次统计

我们采用功勋科学家黄旭华院士的传记序言文本

1 词频统计

统计文本中，指定词汇的次数，步骤：

- 直接定义需要统计的词汇（term）即直接定义列表
- 打开文本
- 采用文本跟词汇匹配的方式，进行词汇统计（注：没有采用采用分词的方式）
- 用字典存储：词汇-频次
- 画图展示

2. 全文词频统计，指定类型词频统计；

```
In [1]: # 文件的打开与关闭
In [2]: #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 [3]: #f_name = open('name.txt')
In [4]: #data_name = f_name.read()
In [5]: #data_name[:70]
In [6]: #print(data_name[:50])
In [7]: #f_name.close()
In [8]: # 将文本转化为列表
In [9]: #names = data_name.split('|') # split一下names就是列表
In [10]: # 以上是课程演示里面的方法，特定的需要统计的词汇放置在txt文本中
         # 我们需要统计的词汇很少，就直接定义需要统计的词汇
In [11]: terms = ['黄旭华','核潜艇','国立交通大学']
In [12]: print(terms)
['黄旭华', '核潜艇', '国立交通大学']
In [13]: terms
Out[13]: ['黄旭华', '核潜艇', '国立交通大学']
In [14]: # 学习一种新的数据结构，字典
In [15]: terms_dict = {}
In [16]: f_txt = open('科学家博物馆-黄旭华传记序言.txt',encoding = 'utf-8')
In [17]: data_txt = f_txt.read()
In [18]: f_txt.close()
In [19]: 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、举路蓝缕、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。

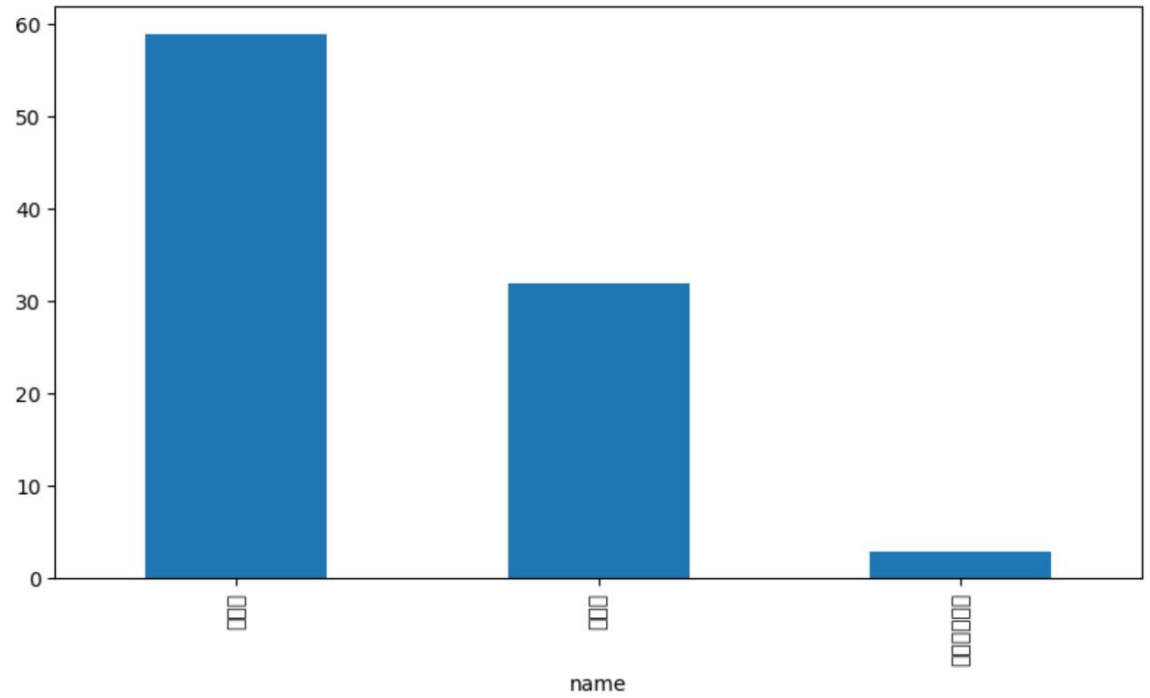
黄旭华是我国著名船舶专家、核潜艇研究设计专家、中国工程院首批院士、中国第一代核动力潜艇研制创始人之一。1924年2月24日，黄旭华出生于广东省汕尾市海丰县田墘镇，原籍广东省揭阳县。1949年，他毕业于国立交通大学造船系船舶制造专业，先后从事过民用船舶和军用舰艇的研究设计工作。1958年，黄旭华开始参与并领导我国第一代核潜艇的研究设计工作，先后出任第一代核潜艇副总设计师、第二任总设计师，历任中国船舶工业总公司及中船重工集团公司第七一九所副总工程师、副所长、所长、党委书记。黄旭华先后于1978年获全国科学大会奖、1982年获国防科工委二等奖，1986年被授予船舶工业总公司劳动模范，1989年被授予全国先进工作者，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 [20]: # 用count函数统计文本中的词汇
```

你没有发现下面的代码很神奇吗？data_txt点count一下，就可以统计词！实际上是字符串匹配的过



第二讲 词频统

2. 完成ppt中的程 词频统计；

这部分是指定类型的频次统计

1 词频统计

统计文本中，指定人名的次数，步骤：

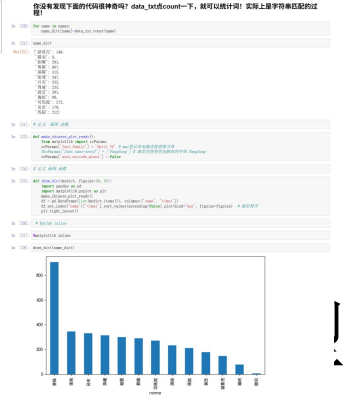
- 打开人名的文本
- 对人名进行列表化处理
- 打开文本
- 采用文本跟人名匹配的方式，进行人名统计（注：没有采用采用分词的方式）
- 用字典存储：人名-频次
- 画图展示

```
In [1]: # 文件的打开与关闭
In [2]: 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 [3]: #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',
          '呂布',
          '周瑜',
          '趙雲',
          '龐統',
          '司馬懿',
          '黃忠',
          '馬超']
In [12]: # 学习一种新的数据结构，字典
In [13]: name_dict = {}
In [14]: f_txt = open('sanguo.txt',encoding = 'GB18030')
In [15]: data_txt = f_txt.read()
In [16]: f_txt.close()
In [17]: print(data_txt[:100])
         《三国演义》（全）

         （明）羅貫中著


         第一回
         宴桃園豪傑三結義斬黃巾英雄首立功
         話說天下大勢，分久必合，合久必分：周末七國分爭，並
         入于秦；及秦滅之後，楚、漢分爭，又並入於漢；漢朝自高祖
         斬白
In [18]: # 用count函数统计文本中的词汇
```

你没有发现下面的代码很神奇吗？data_txt点count一下，就可以统计词！实际上是字符串匹配的过程！



第二讲 词频统计

3. 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”；

黄旭华：**47 次**

核潜艇：23 次

采集：22 次

小组：14 次

资料：13 次

学术：12 次

成长：10 次

研制：9 次

工作：8 次

工程：7 次

“黄旭华” 共出现了 **47 次**。

第二讲 词频统计

4. 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）。

这篇论文以谷歌图书和学术文献为数据基础，聚焦物理学家的科学声誉追踪，其研究视角与方法给人诸多启发。论文最令人印象深刻的是将“科学声誉”这一抽象概念转化为可量化的研究对象，通过分析科学家姓名在全球数字化书籍和学术文献中的出现频次，打破了传统依赖奖项、引用量等单一指标评估学术影响的局限，为科学声誉研究提供了新颖的量化路径。研究中关于“群体内偏好”的发现极具洞察力——牛顿在英式英语语境中始终保有更高声誉，而爱因斯坦在美式英语、德语等语境中更受关注，这一现象深刻揭示了科学声誉并非绝对客观，而是受语言、地域等文化因素影响，让我们看到科学传播中文化语境的重要作用。同时，论文通过共现分析明确了牛顿的万有引力定律、运动定律与爱因斯坦的相对论、量子理论是其声誉的核心支撑，却也发现部分提及源于科学家的哲学观点、生平轶事等，这说明科学声誉的构成是多元的，不仅限于学术成就本身。

更值得深思的是，论文证实了伟大科学家的声誉具有跨世纪的持久性，牛顿、爱因斯坦等先贤的名字在当代仍被广泛提及，其科学精神通过书籍传播得以延续。这一发现让我们意识到，科学研究的价值不仅在于推动学术进步，更在于形成跨越时空的文化影响力。此外，论文提出谷歌图书可作为替代计量工具，拓展了学术影响评估的边界，为衡量科学家在学术界之外的社会价值提供了新的思路，也提醒我们在大数据时代，可通过更丰富的数据源挖掘学术研究的深层意义。

整体而言，这篇论文以严谨的数据分析、创新的研究视角，让我们对科学声誉的形成、演变与影响因素有了更全面的认知，也展现了数字人文方法在科学史研究中的强大潜力，为相关领域的后续研究提供了宝贵的借鉴。