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

南京工業大學
NANJING TECH UNIVERSITY

01 文本数据分析

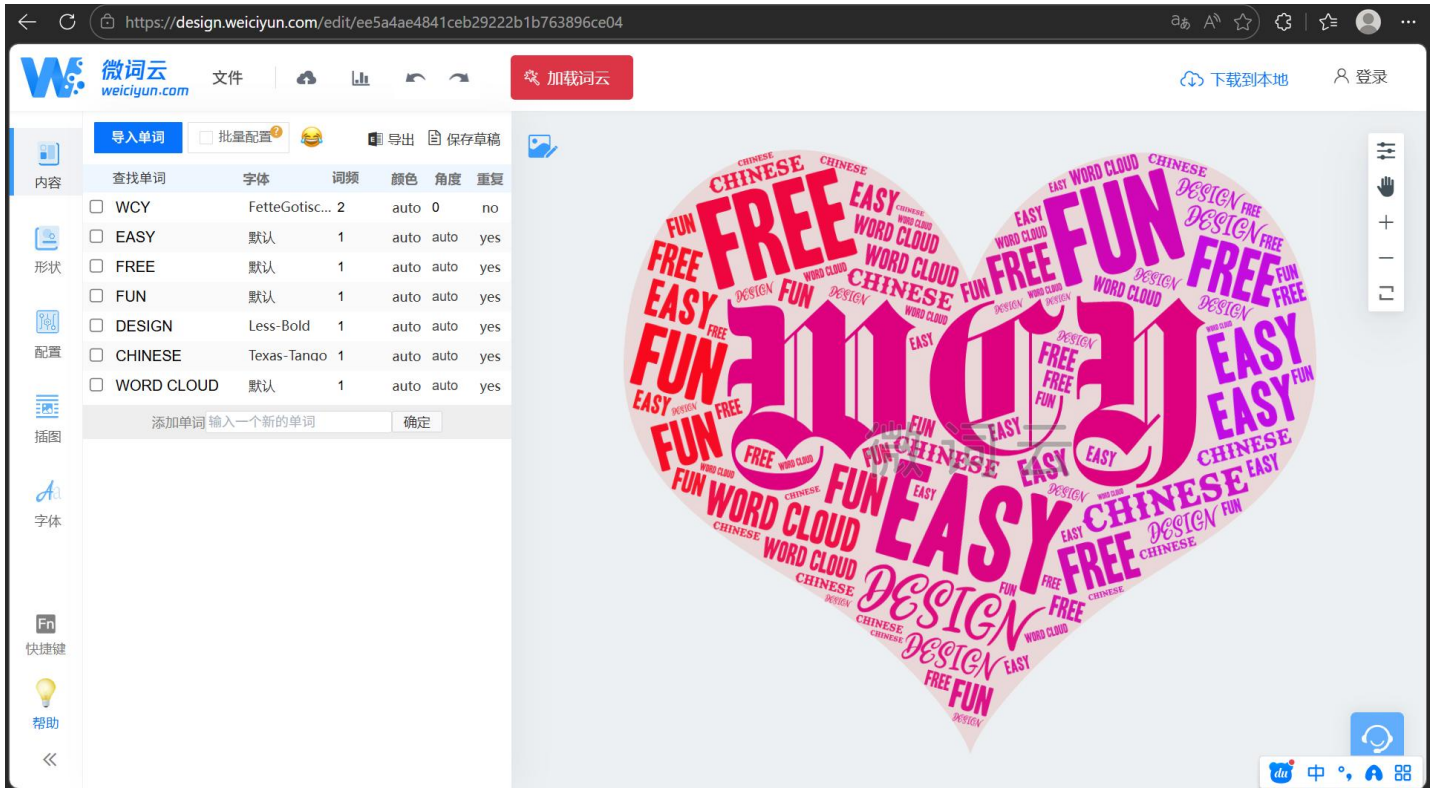
吴志祥

18205185639

cnwzx2012@njtech.edu.cn

第一讲 课程导言与分词

1. 学习使用在线NLPIR分词系统或微词云分词或清华大学分词演示系统（案例演示截图）；



第一讲 课程导言与分词

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

第一讲 课程导言与分词

3. 完成课后作业（001-004，4份代码的运行）。

1.基本分词

In [7]:
```
import jieba
```

In [8]:
```
seg_list1 = jieba.cut("曾经有一份真诚的爱情摆在我的面前，我没有珍惜，等到失去的时候才追悔莫及，人世间最痛苦的事情莫过于此。如果上天能够给我一个
```

In [9]:
```
print('$'.join(seg_list1))
```

```
曾经$有$一份$真诚$的$爱情$摆在$我$的$面前$，$我$没有$珍惜$，$等到$失去$的$时候$才$追悔莫及$，$人世间$最$痛苦$的$事情$莫过于此$。$如果$上天
$能够$给$我$一个$重新$来$过$的$机会$，$我会$对$那个$女孩子$说$三个$字$：$‘$我爱你$’$。$如果$非要$给$这份$爱$加上$一个$期限$，$我$希望$是
$，$一万年
```

In [10]:
```
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
```

In [11]:
```
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@。
```

第一讲 课程导言与分词

3. 完成课后作业（001-004，4份代码的运行）。

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

In [12]:
```
jieba.load_userdict('dict.txt')
```

In [13]:
```
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
```

In [14]:
```
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/。
```

第一讲 课程导言与分词

3. 完成课后作业（001-004，4份代码的运行）。

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

In [15]:
```
stopwords = [line.strip() for line in open('stop_words.txt', 'r', encoding='utf-8').readlines()]
```

In [16]:
```
seg_list_stopw = jieba.cut("曾经有一份真诚的爱情摆在我的面前，我没有珍惜，等到失去的时候才追悔莫及，人世间最痛苦的事情莫过于此。如果上天能够给
```

In [17]:
```
final = ''
```

In [18]:
```
#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
```

In [19]:
```
print(final)
```

```
曾经/有/一份/真诚/爱情/摆在/我/面前/我/没有/珍惜/等到/失去/时候/才/追悔莫及/人世间/最/痛苦/事情/莫过于此/如果/上天/能够/给/我/一个/重新/来/
过/机会/我会/对/那个/女孩子/说/三个/字/：/‘/我爱你/’/如果/非要/给/这份/爱/加上/一个/期限/我/希望/一万年/
```

第一讲 课程导言与分词

3. 完成课后作业（001-004，4份代码的运行）。

第一讲 课程导言与分词

3. 完成课后作业（001-004，4

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。

In [18]:
```
# 简单分词
```

In [19]:
```
import jieba
```

In [20]:
```
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
```

In [21]:
```
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/。
```

In [22]:
```
# 加入用户词典
```

In [23]:
```
jieba.load_userdict('dict.txt')
```

In [24]:
```
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
```

In [25]:
```
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/。
```

In [26]:
```
# 加入词典之后，哪些词汇被分出来了呢？
```

In [27]:
```
# 使用停用词表
```

In [28]:
```
# stopwords = [line.strip() for line in open('stop_words.txt','r', encoding='utf-8').readlines()]
```

In [29]:
```
stopwords = open('stop_words.txt','r', encoding='utf-8').read()
stopwords = stopwords.split('\n')
```

In [30]:
```
stopwords
```

Out[30]:
```
['的', '了', '是', '啊', '、', '，', '。', '停用']
```

In [31]:
```
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
```

In [32]:
```
final = ''
```

In [33]:
```
for seg in seg_list_huang:
    if seg not in stopwords:
        final+= seg+'/'
```

In [34]:
```
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

第一讲 课程导言与分词

3. 完成课后作业（001-004，4份代码的运行）

In [1]:
```
# 1 第一步，我们把文本人工提取出来
```

In [2]:
```
# 落实“企业管理年”主题，加强QEHS三体系建设，通过自动化、数字化、智能化升级改造加快新一代信息技术与企业生产经营融合，打造精益制造能力，提升精
```

In [3]:
```
# 2 第二步，安装需要的python库，也就是工具包
# 我们的步骤是：1）使用jieba分词，对文本进行分词；2）统计给定词汇的频次，比如：自动化、数字化、安全。
```

In [4]:
```
import jieba
from collections import Counter
```

In [5]:
```
# 文本
text = """
落实“企业管理年”主题，加强QEHS三体系建设，
通过自动化、数字化、智能化升级改造加快新一代信息技术与企业生产经营融合，
打造精益制造能力，提升精细化管理水平，助力公司从“制造”升级为“智造”，
从而提高经营效率和效益;通过集团一体化智数管理平台建设与运营，
丰富企业供应链管理、生产工艺控制等管理工具，不断增强生产经营过程数据获取与分析能力，
强化全过程一体化管理，提高自动化、数字化、智能化的供应链管理能力，
为体系安全稳定运行与管理水平提升保驾护航，致力打造安全智能化工厂；
打造助剂互联网技术合作和商务合作平台，构建具有国际竞争力的供应链体系。
"""
```

In [6]:
```
# 1. 分词处理
words = jieba.lcut(text)
```

```
Building prefix dict from the default dictionary ...
Loading model from cache C:\Users\admin\AppData\Local\Temp\jieba.cache
Loading model cost 0.746 seconds.
Prefix dict has been built successfully.
```

In [7]:
```
words
```

Out[7]:
```
['\n',
 '落实',
 '“',
 '企业',
 '管理',
 '年',
 '”',
 '',
 '主题',
 '，',
 '',
 '加强',
 'QEHS',
 '三',
 '体系',
 '建设',
 '，',
 '\n',
 '通过',
 '自动化',
 '、',
```

In [8]:
```
# 2. 定义要统计的特殊词汇
target_words = ['数字化', '智能化', '安全']
```

In [9]:
```
# 统计词频
word_counts = Counter(words)
```

In [10]:
```
# 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 [11]:
```
# 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
```

第一讲 课程导言与分词

3. 完成课后作业（001-004，4份代码的运行）。

提取到的实体和专业术语：

```json
{
  "理论": [
    "肿瘤免疫微环境",
    "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": [
    "T细胞耗竭",
    "scRNA-seq",
    "非小细胞肺癌",
    "免疫抑制信号通路",
    "PD-1/PD-L1",
    "TGF-β路径",
    "个体化免疫治疗"
  ]
}
```

第一讲 课程导言与分词

4. 谈一谈在营销学科/领域，文本、文本分词以及实体的内涵。例如：客户关系管理中，文本分析的价值。（仅营销）

阅读这篇关于学术论文研究方法自动分类的论文，最直观的感受是其对“全文价值”的深度挖掘与实践突破。在以往的相关研究中，受限于文献获取条件和技术成本，大多依赖摘要、标题等短文本信息开展分类，却忽略了全文中更丰富的研究方法上下文——这篇论文恰恰抓住了这一核心痛点，以 820 篇图书情报领域的论文全文为基础，通过专家标注构建语料库，为分类任务提供了更扎实的数据源，这种“回归文本本身”的研究思路，让分类结果更具可信度。

论文在技术设计上的严谨性也令人印象深刻。面对学术论文可能存在多种研究方法的现实情况，其采用多标签分类的思路，兼顾问题转换法与算法自适应法，构建了 7 种不同的分类模型进行对比，既考虑了标签间的关联性，又覆盖了不同底层分类器的特性。尤其是朴素贝叶斯算法在分类器链策略中取得 0.705 的 F1 值，证明了传统机器学习算法在特定场景下依然具备强大的适配性，而实验法、计量法因特征表征能力强而获得更高分类准确率的结果，也揭示了“方法本身的特性”与“文本特征提取”之间的密切关联，为后续研究提供了明确的优化方向。

更值得深思的是，论文不仅关注技术层面的优化，更兼顾了研究的实际应用价值。人工分类学术论文研究方法存在成本高、主观性强的弊端，而该研究构建的自动分类模型，既降低了人工依赖，又能为科研人员选择研究方法、分析学科方法演变趋势提供数据支撑，真正实现了技术服务于学术研究的初衷。同时，论文也坦诚指出了研究的局限性，如训练集规模对泛化效果的影响、样本不均衡的问题等，这种客观的自我审视，让研究更具延续性和拓展空间。

整体而言，这篇论文以清晰的逻辑、扎实的实验和务实的价值导向，为文本分类在学术研究领域的应用提供了优秀范例。它让我意识到，学术研究中“看似基础的调整”（如从摘要到全文的数据源转换），往往能带来突破性的成果，而技术创新的核心，始终是围绕“解决实际问题”展开的精准发力。