华东师范大学计算机科学技术系实验报告

课程名称： 数据分析实践	**年级：** 2017 级	**实践作业成绩：**
指导教师： 兰曼	**姓名：** 英嘉豪	**作业提交日期：**
实验编号： 1	**学号：** 10175102247	**实践作业编号：** 1

1 实验名称

网络新闻数据的采集以及数据的预处理

2 实验目的

掌握本地数据的读写，网络数据的采集，应用数据分析技术，对结果进行可视化，并对统计和可视化结果进行分析和讨论并对数据分析的结果保存入磁盘。

实验要求

新闻标题数据下载及分析

1. ：下载 2 周的新闻数据（尽可能多）
2. ：进行学习过的数据预处理技术（重复记录，噪音数据清理等）
3. ：分析并观察各种板块新闻数据的分布等

3 实验内容

3.1 环境准备

3.1.1 基本环境

- Windows 10

- Python 3.6.7

3.1.2　依赖

- pandas 0.23.4
- scipy 1.2.0
- numpy 1.15.4
- bs4 0.0.1
- requests 2.23.0

3.2　获取新闻标题数据

3.2.1　从滚动栏获取新闻标题

图 1: 滚动新闻页面

```
# 加载数据
url = "http://www.chinanews.com/"
```

```
html = requests.get(url)
soup = bs(html.content, "html.parser")
url_scroll = soup.find(name='ul',attrs={"class":"nav_navcon"}).find('a')['href']
# 获取顶部导航栏 排第一的滚动页面的 url
url_scroll = "https:" + url_scroll
# 写追加的
f = open('news.csv', "a+")
for i in range(1,11):
    url_scroll = url_scroll[:-6] + str(i) + ".html"
    html_scroll = requests.get(url_scroll)
    soup_scroll = bs(html_scroll.content,"html.parser")
    news = soup_scroll.find(name="div", attrs={"class" :"content_list"})
    news = news.findAll(name='li')
    for n in news:
        news_lab = n.find(name='div', attrs={"class" :"dd_lm"})
        # 滚动栏之间有空格要特殊处理
        if news_lab == None:
            continue
        # 获取类标
        news_title = n.find(name='div', attrs={"class" :"dd_bt"}).text
        # 获取时间
        news_time = n.find(name='div', attrs={"class":"dd_time"}).text
```

这里使用 request 获取页面的内容，并用 bs4 做网页内容的解析。我们分别提取出他的类标，内容，时间等参数。但是在数据预处理的时候我发现有形如:【视频】,【图片】等新闻标题类标，这些类标对于我们后面做新闻分类是没有任何帮助的，但是单纯删掉后新闻数量又会不够，所以我分别访问滚动页面上面的类标页面，在各个页面分别再做数据爬取。

```
for nav in navbar[1:14]: # 取第一栏
    f = open(firstr,'a+',encoding="utf-8")
    if nav['href'][:5]!="http:":
        url_nav = "http:" + nav['href']  # 获取改板块的 url
    html_nav = requests.get(url_nav)
    soup_nav = bs(html_nav.content,"html.parser")
    for n in soup_nav.findAll("li") + soup_nav.findAll("em") + soup_nav.findAll("h1"):
```

```
        if n.a and n.a.string and len(n.a.string) > 7:
            f.write(str(nav.string) + "," + str(n.a.string) + "\n")
```

3.2.2　数据预处理

```
# 将新闻数据进行按标做分类处理处理重复的新闻
blockList = [' 时政"0027, ' 国际', ' 社会', ' 财经', ' 产经', ' 金融', ' 汽车',
' 港澳',' 台湾', ' 华人', ' 娱乐', ' 体育', ' 文化']
newsWordDic = dict()
for blockName in blockList:
    newsWordDic[blockName] = dict()
while line:
#print(line)
line = line.strip().split(",")
newsTitle = str(line[1:])
newsWordDic[line[0]][newsTitle] = newsWordDic[line[0]].get(newsTitle, 0) + 1
line = f.readline()
```

由于我获取新闻数据时候为了后面做聚类采用了在各个子页面读取，因为我需要将同类型进行合并并进行去重的操作。为了进一步进行各个模块新闻的分布等情况我们还需要将新闻进行进一步处理，这里我采用了分词处理并统计词的频次。

```
dirtWordList = [',', '[',']',', ','.','。','(',')','?','、','\'',
"\"",' ‘',’' ',’"',’"',' (',')',' ? ',' ',' “',’” ',': ',
'! ','—','...',' 《','》 ']
newsWorAnalise = dict() # 各个类的字典的
 for news in newsWordDic[lab]:
        newsWordList = list(jieba.cut(news))
        for word in newsWordList:
            if word not in dirtWordList and word not in stopwords:
                newsWorAnalise[lab][word] = newsWorAnalise[lab].get(word, 0) + 1
```

在初步使用jieba.cut进行分词后发现结果中有很多停用词和标点，于是我下载了哈工大的

停用词表加上自己构造的中文标题dirWordList对分词结果进行了预处理。

4 实验结果及分析

爬取了两周的新闻数据选取了共约有 11000 条新闻但是实际京经过去重后直只有 4028 条数据，数据量还是不够的，下面我可能会在爬取两周数据。我把各个板块的关键词做了统计如下：

类别	新闻数量	频繁词项
时政	415	(’ 中国’, 52) (’ 疫情’, 45) (’ 防控’, 29)
国际	551	(’ 确诊’, 145) (’ 疫情’, 126) (’ 病例’, 118)
社会	289	(’ 武汉’, 32) (’ 男子’, 18) (’ 疫情’, 17)
财经	301	(’ 中国’, 27) (’ 疫情’, 20) (’ 消费’, 19)
产经	239	(’5G’, 24) (’ 新’, 24) (’ 直播’, 14)
金融	138	(’A 股’, 13) (’ 亿元’, 12) (’ 企业’, 11)
汽车	106	(’ 汽车’, 38) (’ 中国’, 16) (’ 万元’, 13)
港澳	291	(’ 香港’, 167) (’ 确诊’, 55) (’ 澳门’, 47)
台湾	229	(’ 台湾’, 103) (’ 台’, 44) (’ 确诊’, 41)
华人	351	(’ 中国’, 121) (’ 驻’, 67) (’ 疫情’, 59)
娱乐	418	(’ 疫情’, 23) (’ 确诊’, 16) (’ 肺炎’, 15)
体育	501	(’ 疫情’, 44) (’ 中国’, 31) (’ 月’, 28)
文化	199	(’ 发现’, 14) (’ 直播’, 11) (’ 年’, 11)

表 1: 新闻数据统计

不难发现几乎所有的的板块都离不开新冠疫情的情况。

5 问题讨论

关于后面实验的猜想

对新闻数据进行标题分类第一步肯定是要构建合适的词向量，并且采用一定聚类算法，或者

分类算法对新闻数据进行分析。但是我感觉我选取的新闻的类标过多，不一定能有很好的效果，尤其是几个类标内容较为接近，而且由于所有模块都有新冠病毒的报道这样无疑会导致新闻可用性下降。

6　感想

重新复习了 repuest 库的使用，感觉还是挺好的。