广州大学

本科毕业论文

课题名称	利用文本挖掘技术进行新闻热点关注问题分析
学　　院	计算机科学与教育软件学院
专　　业	计算机科学与技术
班级名称	计机 141
学生姓名	叶建成
学　　号	1406100037
指导教师	张艳玲
完成日期	2018 年 5 月 20 日

利用文本挖掘技术进行新闻热点关注问题分析

摘要 如今新闻泛滥，令人眼花缭乱，即使同一话题下的新闻也多得数不胜数。人们可以根据自己的职业和爱好关注专业新闻网站的不同热点要闻。因此，通过对人们关注新闻的热点问题进行分析，可以得出民众对某个领域的关切程度和社会需要解决的问题，也有利于了解当前的舆论焦点，有助于政府了解民意，便于国家对舆论进行正确引导，使我们的社会更加安定和谐。本文以财经领域为例，通过爬虫技术抓取网络上的大量财经新闻，通过对新闻内容文本进行预处理及密度聚类分析来发现热点；从发现的热点中，再进行词汇聚类分析，得出热点所涉及的人或事物，以此分析出社会对经济领域关注的问题和需要解决的问题。

关键词 新闻热点分析；DBSCAN密度聚类；k-均值；TextRank

ABSTRACT Today's news is dazzling, and even the news on the same topic is numerous. People can follow the different hot topics of professional news websites according to their own profession and hobbies. Therefore, by analyzing the hot topics of people's attention to news, it is possible to draw the public's concern about a certain area and the problems that the society needs to solve. It is also conducive to understanding the current focus of public opinion and helps the government understand public opinion and facilitate the country’s understanding of public opinion. The correct guidance of public opinion makes our society more stable and harmonious. This topic takes the financial field as an example. It crawls a large amount of financial news on the Internet through crawler technology. It finds hot spots by preprocessing and density clustering analysis of news content texts; from the hot spots found, it conducts lexical cluster analysis. Draw the people or things involved in the hotspots to analyze the social concerns of the economic field and the problems that need to be solved.

KEY WORDS news hotspots analysis; DBSCAN; k-Means; TextRank

目录

1.前 言 1
1.1 课题研究背景和意义 1
1.2 热点关注问题分析研究现状 2
1.3 本文的主要工作 3
1.4 本文的组织结构 4
2.基于新闻 API 以及 Lxml 模块的财经新闻抓取 5
2.1 引言 5
2.2 抓取网页内容的方法 5
2.3 对新浪、搜狐、新华网即时财经新闻的采集 6
2.3.1 基于新闻 API 的新闻标题、发布时间、url 的抓取 6
2.3.2 基于 lxml 模块的新闻内容的抓取 7
2.3.3 财经新闻总体数据抓取实现 7
2.3.4 多线程实现财经新闻抓取 7
2.4 实验结果 8
2.5 本章小结 9
3.基于 DBSCAN 聚类算法的新闻聚类分析 10
3.1 引言 10
3.2 文本特征提取方法 10
3.2.1 词频方法 10
3.2.2 文档频数方法 10
3.2.3 TF-IDF 11
3.3 文本聚类方法 11
3.3.1 什么是聚类 11
3.3.2 几类常见的聚类方法 12
3.4 基于密度的 DBSCAN 新闻文本聚类 13
3.4.1 新闻内容的预处理 13
3.4.2 基于 DBSCAN 聚类算法的新闻聚类 14
3.5 新闻文本聚类结果 15

3.6 本章小结 16
4.热点新闻话题排行 17
4.1 引言 17
4.2 基于 TextRank 的自动文摘算法 17
4.3 热点新闻话题排行 18
4.3.1 基于统计的总体新闻的热点排行 18
4.3.2 基于 TextRank 自动文摘算法的热点内部话题排行 18
4.4 热点新闻话题排行实验结果 18
4.4.1 总体新闻的热点排行实验结果 18
4.4.2 热点内部话题排行实验结果 19
4.5 本章小结 20
5.基于 word2vec 技术及 k-Means 聚类的词汇聚类分析 21
5.1 引言 21
5.2 词汇的词向量的获取方法 21
5.2.1 One-Hot 编码 21
5.2.2 word2vec 训练词向量 21
5.3 新闻热点内部词汇聚类 22
5.3.1 依据新闻内容训练 word2vec 词向量模型 22
5.3.2 基于 k-Means 的词汇聚类 23
5.4 新闻热点内部词汇聚类实验结果 23
5.5 本章小结 24
6.系统界面可视化及实验结果 25
6.1 系统主界面 25
6.2 系统热点详情界面 28
7.结论与展望 32
7.1 结论 32
7.2 未来工作展望 32
参考文献 34
致谢 36

利用文本挖掘技术进行新闻热点关注问题分析

1.前 言

1.1 课题研究背景和意义

近年来，互联网已经发展成为人们获取信息、关注新闻热点、了解国情乃至了解世界的重要媒介。而且，我国互联网发展迅猛，普及越来越广，网络用户大量增加，互联网逐渐融入人们的生活中，也影响并改变着人们的生活习惯和生活方式。2018 年 1 月 31 日，根据中国互联网络信息中心（CNNIC）发布的第 41 次《中国互联网络发展状况统计报告》，截止到 2017 年 12 月，我国网民规模达 7.72 亿，互联网的普及率也达到 55.8%，全年合计新增网民 4074 万人，增长率为 5.6%[1]。目前，我们已经步入了“互联网+”的生活时代，互联网对我们获取新闻信息的方式影响也越来越大，人们获取新闻资讯也逐渐向网络新闻这个方式转变。相比广播、报纸、电视等传统的新闻媒介，网络新闻具有涉及全面，实时性强，更新周期短、速度快等特点，更具优势。

另一方面，社会各方面高速发展，科技发达，信息泉涌，人们之间的交流越来越密切，生活也越来越方便。随着“大数据”时代的到来，信息量增多，信息种类繁多，涉及范围广泛，新闻五花八门，同一主题下的新闻，多得数不胜数，人们可以根据自己的职业和爱好关注专业新闻网站最近的不同热点要闻。

但是由于互联网发布新闻的条件约束小、门槛低，很多门户网站为了博人眼球、赚取流量费用，又或者是为了提升网站排名和知名度等，标题常常采用春秋笔法、以偏概全、断章取义或者低俗写法，恶意引导民众的舆论走向，进而可能引起社会动乱或者社会舆论危机等，错误的和不合时宜的新闻报道甚至会引起人民群众对政府的公信力产生质疑。因此，对新闻热点关注问题进行分析很有必要，政府可以了解民众对某个领域的关切程度和社会需要解决的问题，及时了解当前的舆论焦点和民意，从而进行正确的舆论引导和宣传，降低负面新闻对人们生活的影响。及时扼制事情的发展，增加政府在网民心中的公信度，同时也能让不法分子无法利用网络的突发性、便捷性以及难以控制的特点来达到他们不可告人的目的。

1.2 热点关注问题分析研究现状

话题检测与追踪（Topic Detection and Tracking，即 TDT）是一项信息处理技术，用于发现新闻报道中的新生消息，自动发现新闻事件，能够持续跟踪已知新闻话题等。最近几年，互联网高速发展，网络新闻也越来越成为新闻报道的主要媒介，网络新闻的爆炸式增长，也给热点话题检测和跟踪提供了良好的条件支持。

目前，有关互联网舆论热点、热点话题发现的研究越来越多，国内外不少大学和机构也都形成专门的研究课题小组，通过研究和改进网络舆情发现与分析的基本技术，也已经有很多丰硕的成果。其中源于早期面向事件的检测与跟踪（Event Detection and Tracking，简写为 EDT）的话题检测与跟踪(Topic Detection and Tracking，简写为 TDT），是由美国国际高级研究计划局发起的研究项目，其较为出名，该项目最初研究了一种算法[2]，可以识别来自于新闻数据流中的重要信息，并跟踪它们，其将话题检测与跟踪分成 5 个子任务：报道切分、话题跟踪、话题检测、首次报道检测、关联检测来实现。

国内对 TDT 领域的研究都是基于对 NLP（自然语言处理）基础上的研究，这主要是由于中英文句子构成的差异所致。因为在英文分词中，只需要按空格分割即可完成分词，而在汉语中，根据字义和词义的拆分是非常可变的，而且往往都能分出多种语义。尽管国内对 TDT 的研究相对较晚，但仍然是取得了很好的成果。王翔使用了基于句义结构模型构建了一种基于聚类的互联网热点事件发现方法，采用single-pass聚类思想和凝聚式层次聚类与K-Means聚类算法相结合的聚类算法来发现新闻热点[3]。王馨在新闻挖掘过程中使用了改进的关联规则算法，根据互信息来计算文本字符串的相似度，然后得出热点新闻的关键词集合，再进行热度计算来研究新闻热点[4]。陈龙引入了 LDA 主题模型，提出了一种多核心话题描述模型，能够识别同一话题下不同的关注核心，之后采用划分聚类与层次聚类结合的方法对新闻报道进行精确聚类[5]。赵旭剑深入研究了话题模型、时态信息处理以及话题动态演化，提出了一种面向中文网络新闻的话题信息抽取方法[6]。彭卫华使用文本分类的方法对新闻报道分类，然后使用 TDT 技术生成一系列的专题，以标题、相关词群、事件趋势图等来表示某一个专题，用话题延续的天数、前 k 天的报道数和话题延续的最长天数计算关注度对专题进行打分排序，最后将最新最热的专题呈现给用户[7]。刘林浩采用了先分类再聚类的方法来避免不同类别新闻的干扰，在分类和聚类的过程中，使用词频的平滑、向量空间的压缩以及对算法的改进来提高准确度[8]。廖君华等人采用

LDA 模型对网络热点话题主题进行提取,并利用时间标签来发现热点话题[9]。到目前为止，新闻专题研究一直是文本数据挖掘和 NLP 领域的研究热点。

但是，现今对新闻专题的研究，大多都提出新闻热点的发现，较少的对新闻内容的问题分析。本文提出对新闻内容关注的问题进行分析，实现了解当前的舆论焦点和民意的目的。

1.3 本文的主要工作

本文主要通过文本挖掘技术进行新闻热点问题分析，把从网上抓取到的财经新闻，通过对新闻内容的聚类，得到新闻热点；再对热点进行分析，通过对某一热点相关词汇的聚类，得到热点问题所涉及的人物、行业或组织等。主要涵盖的内容如图 1-1 所示：

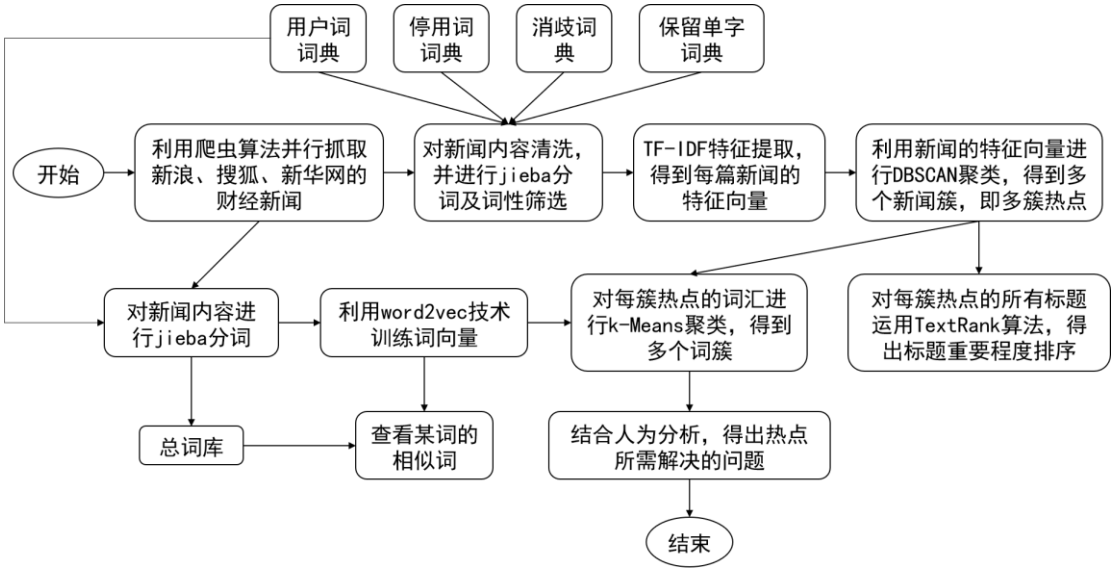


图 1-1 新闻热点关注问题分析总任务

由图 1-1 所见，本文主要研究的内容为：

1、利用新闻 API、爬虫算法、多线程并行技术，抓取三大专业财经新闻网站（新浪财经、搜狐财经、新华网财经）的大量财经新闻报道；

2、对新闻进行去重、时间段过滤，然后对新闻内容文本进行 jieba 分词并词性标注，过滤出名词、动词、简称等词性，分词前使用自定义的用户词词典增加分词的准确性，分词后使用停用词词典、消歧词典、保留单字词典过滤掉对话题无关并且影响聚类准确性的词，建立每篇新闻的词库，利用 TF-IDF 特征提取之后对新闻进行 DBSCAN 聚类，并对每个类的大小进行排序；

3、针对聚类后的每一类新闻，为了得到该处热点的话题信息，还需要提取它们

的标题，利用 TextRank 算法，对标题的重要程度进行排序，用重要性最高的标题来描述该处热点的话题；

4、对所有的新闻内容进行 jieba 分词，并训练出 word2vec 词嵌入模型，然后对聚类后的每一类新闻，提取它们的内容分词后的结果，运用 word2vec 模型得到每个词的词向量，再利用 k-Means 聚类算法进行相近词聚类。

1.4 本文的组织结构

本文总共分为七章，具体结构安排如下：

第一章，介绍课题的研究背景和意义，研究现状以及本文的总任务及结构安排。

第二章，介绍了网络爬虫的基本理论和概念，以及本文采取的爬虫算法，以及抓取新闻的方式。

第三章，详细介绍聚类算法，如 DBSCAN 等，以及介绍聚类前的预处理及特征提取过程，并对抓取的财经新闻进行聚类。

第四章，介绍自动文摘里的 TextRank 算法，本文用来计算每个标题的重要性。

第五章，介绍 word2vec 技术，word2vec 是一个可以将语言中的字词转化为向量形式的表达的模型，介绍另一种划分方法聚类 k-Means，对热点的词汇进行聚类。

第六章，设计并实现一个可视化的界面来操作以上的所有任务，获得课题研究的总结果，界面分爬虫模块、新闻聚类模块和新闻热点详情模块。

第七章，分析实验结果得出结论，总结与展望。

2.基于新闻 API 以及 Lxml 模块的财经新闻抓取

2.1 引言

随着互联网的发展，网络上的数据量也急剧增大。数据也涉及到方方面面，如财经、科技、军事等等。面对如此庞大的数据量，我们要想获取想要的资源也是一件很艰难的事情，所以网络爬虫便应运而生了。

网络爬虫是一种根据某种设定好的规则自动抓取网络信息数据的程序或者脚本，也称为网页蜘蛛或网络机器人[10]。网络爬虫所用的框架大同小异，针对不同的网站其原理都很类似。大部分爬虫的抓取框架流程是按照发送请求、获得页面、解析页面、下载内容、储存内容这些步骤一步一步进行的。

需要注意的是，网站往往有反爬虫策略来限制或者阻碍爬虫对自己网站的肆意爬取。在网站设置了反爬虫策略之后，爬虫方还想抓取相应的数据，就需要运用相应的攻克手段，即需要对原来的脚本增加防反爬虫处理。其实反爬虫策略是什么，其相应攻克反爬虫的方法就一定会存在。常见的反爬虫策略主要有：IP 限制、userAgent 限制、Cookie 限制、资源随机化存储以及动态加载技术等。而其相应的防反爬虫处理手段，如 IP 代理、用户代理、Cookie 保存与处理、自动触发技术、逆向工程或者借助类似浏览器的工具渲染动态网页等。不过对于人们经常查看的网页，其反爬虫措施一般较少，如新闻网站等。

2.2 抓取网页内容的方法

一般如果能获取到网站给出的外部 API，那么获取网站数据便轻而易举。请求 API 返回的数据往往是 json 格式的文件，只需要用 json 模块进行转化就可以获得一个有关新闻信息的字典。其键值对的形式使人们很容易通过关键字获取到其对应的值。

不过网站很多地方往往没有设置数据获取的 API，因为维护一个前端的界面代价往往小于维护好一个后台的 API。所以还是需要通过各种爬虫手段来抓取网站页面的代码，然后对网页代码进行解析来获取到想要的数据。抓取网页数据的方法一般有三种：正则表达式匹配、BeautifulSoup 模块和强大的 lxml 模块。

例如有一个网页的代码 html 有“<td>This is a text.</td>”，需要获取<td>元素中的内容。

对于正则表达式，可以运用 Python 的 re 模块，通过

“re.findall(‘<td>(.*?)</td>’, html)”正则匹配得到；

如果运用 BeautifulSoup 模块，那么可以使用

“BeautifulSoup(html).find(‘td’).text”也可获取；

如果使用 lxml 模块，也可以使用

“lxml.html.fromstring(html).cssselect(‘td’)[0]”获得数据。

不过这三者各有特点。通过对同一数据的抓取实验表明，BeautifulSoup 会比其他两种方法慢很多。但是 Lxml 的安装过程比较困难，而正则表达式方法在编写正则匹配数据的时候相对比较麻烦。

2.3 对新浪、搜狐、新华网即时财经新闻的采集

2.3.1 基于新闻 API 的新闻标题、发布时间、url 的抓取

在课题开始之时，本文曾经使用爬虫爬取过 7 个新闻网站的新闻。但是由于有些网站爬取下来的数据时间杂乱无序，也不是即时的，数据的结构也很难保持一致性。为了解决这一问题，本文最终在新浪、搜狐、新华网分别找到它们解析新闻网页内容的 API，用 API 请求数据之后，便可以得到最近新闻的标题、发布时间以及新闻链接。

从三个财经网站上，获取的新闻 API 如图 2-1 所示：

新浪、搜狐、新华网财经新闻的API分别为：

```
sina_template_url = 'http://roll.news.sina.com.cn/interface/rollnews_ch_out_interface.php' \
                    '?col=43&spec=&type=&ch=03&k=&offset_page=0&offset_num=0&num={}&asc=&page=1&r=0.{}'
sohu_template_url = 'http://v2.sohu.com/public-api/feed?scene=CHANNEL&sceneId=15&page=1&size={}'
xinhuanet_template_url = 'http://qc.wa.news.cn/nodeart/list?nid=11147664&pgnum={}&cnt={}&tp=1&orderby=1'
```

图 2-1 新浪、搜狐、新华网即时财经新闻 API

由图 2-1 可见，新浪、搜狐、新华网的财经新闻 API 分别为图 2-1 中的 sina_template_url、sohu_template_url、xinhuanet_template_url。

获得新闻 API 之后，便可以利用 Python3 的 urllib 模块，通过使用 urllib.request 的 Request 函数和 urlopen 函数分别请求和打开 url 中的内容，然后进行 read 操作并正确编码就可以获得 json 文件，通过 json 模块的解析，即可得到关于新闻的 dict 字典。通过字典获取其中的 title、time、url 的值，便可以得到新闻的标题、发布时间以及 url。

2.3.2 基于 lxml 模块的新闻内容的抓取

除了获得新闻的标题、时间、url 以外，还要获取新闻的详细内容，这便需要使用 lxml 模块的 html 解析器对 url 请求到的数据进行解析，得到包含新闻所有 html 信息的特殊的数据结构。然后可以通过 xpath 的语法来查找新闻内容的信息。如新浪财经新闻内容的 xpath 为“//*[@id="artibody"]/p”，搜狐财经新闻内容的 xpath 为 “ //*[@id="mp-editor"]/p ” ， 新 华 网 新 闻 内 容 的 xpath 为“//*[@id="p-detail"]/p”。

2.3.3 财经新闻总体数据抓取实现

财经新闻总体数据抓取过程如图 2-2 所示：

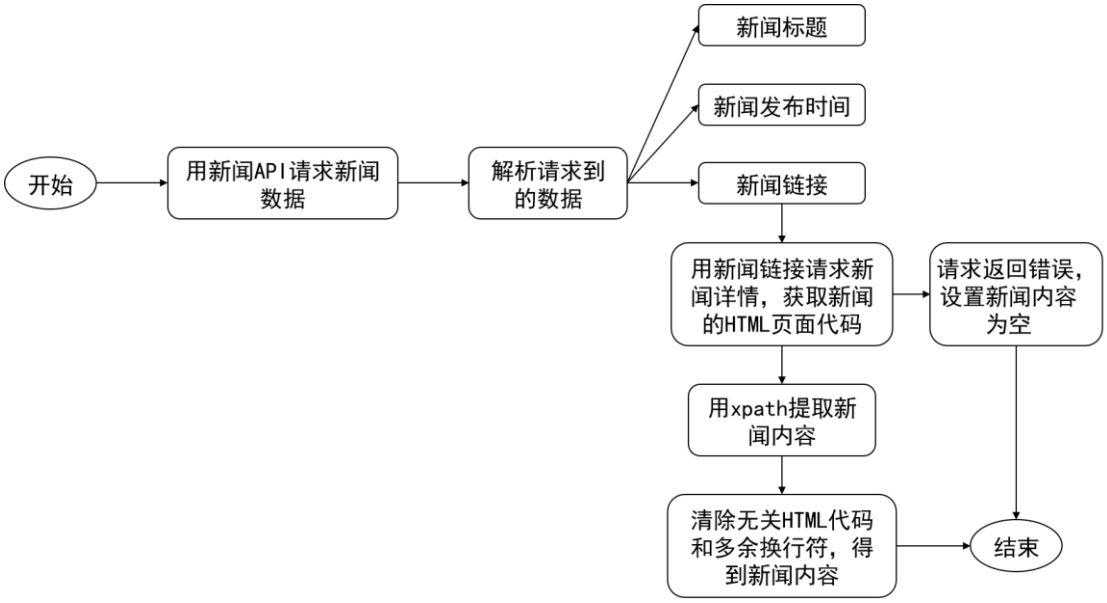


图 2-2 爬虫抓取新闻流程图

从图 2-2 中可以看出，可以先利用新闻 API 获得新闻标题、新闻发布时间和新闻 url，然后利用 url 再去抓取得到新闻内容。

2.3.4 多线程实现财经新闻抓取

经测试，财经新闻的 API 一次最多只能返回 6000 多条数据，搜狐和新华网的 API 最多都能返回 1000 条数据。如果利用上一爬取的流程串行抓取新浪财经新闻 6000 条、搜狐和新华网财经新闻各 1000 条，效率肯定非常慢。但是另外，如果抓取网站的速度过快，就会面临爬虫被封禁或者造成服务器过载的风险。为了降低这些风险但又提高抓取的效率，需要在同一域名下的两次下载之间增加了延时，但是在不同域名之间又用了多线程进行同时爬取，从而大大增加了爬虫抓取新闻的效率。多线程爬虫的抓取过程如图 2-3 所示：

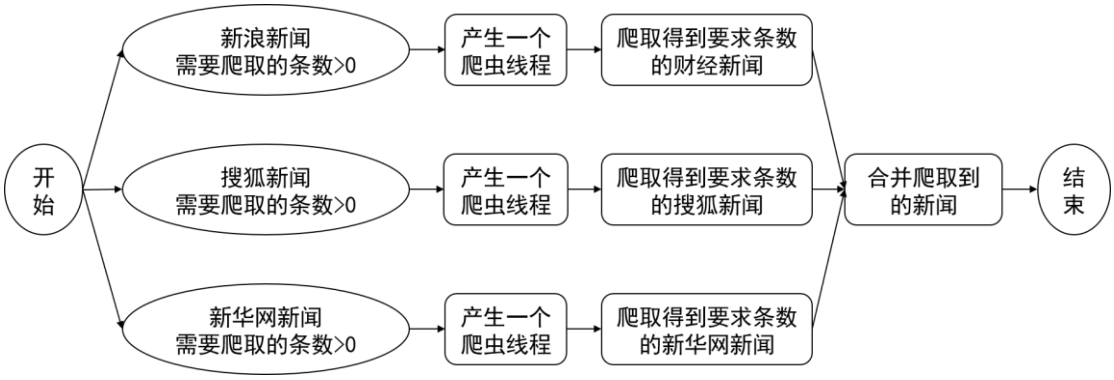


图 2-3 多线程爬虫同时抓取三个新闻网站

由图 2-3 可见，可以分别为每个新闻网站创建一个线程，这样通过多线程增加了爬虫的抓取效率，把从三个网站抓取的新闻合并在一次，达到快速抓取大约 8000 条新闻的效果。

2.4 实验结果

多线程抓取爬虫的抓取结果如图 2-4 所示：

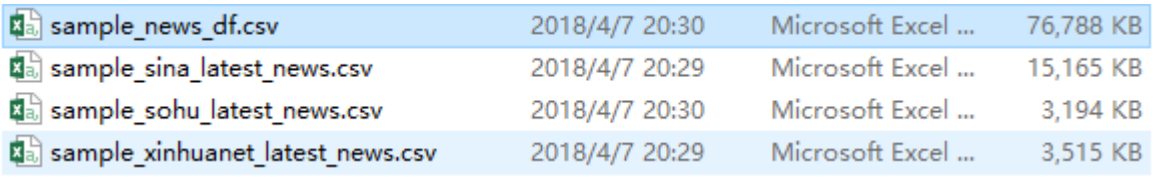
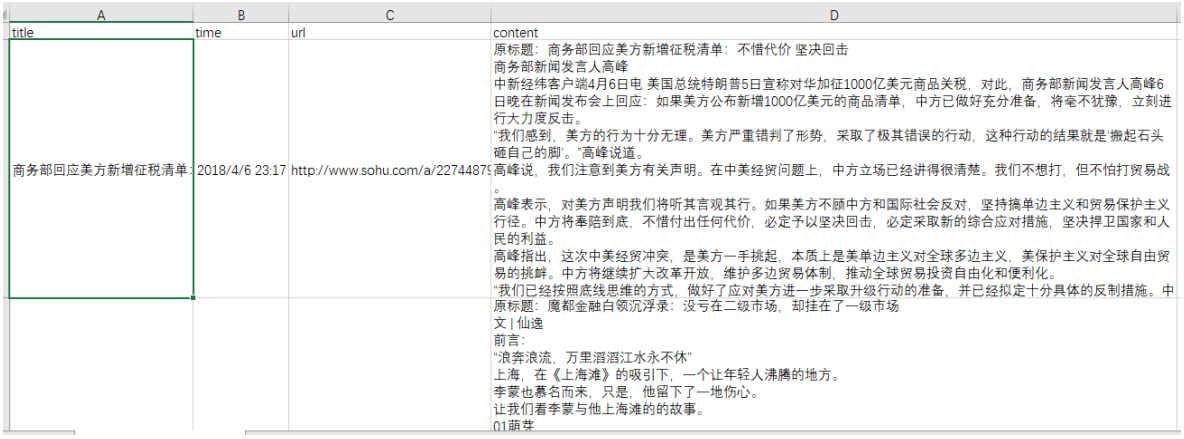


图 2-4 抓取新闻结果

本文样本数据抓取的时间是 2018 年 4 月 7 日以前的新闻。由图 2-4 可见，第一个文件是爬取下来的新闻合并之后的文件，后三个分别是新浪、搜狐、新华网抓取下来的新闻，可以看出数据量还是挺大的。

抓取的新闻格式如图 2-5 所示：



	A	B	C	D
	title	time	url	content
	商务部回应美方新增征税清单：	2018/4/6 23:17	http://www.sohu.com/a/22744875	原标题：商务部回应美方新增征税清单：不惜代价 坚决回击 商务部新闻发言人高峰 中新经纬客户端4月6日电 美国总统特朗普5日宣称对华加征1000亿美元商品关税，对此，商务部新闻发言人高峰6日晚在新闻发布会上回应：如果美方公布新增1000亿美元的商品清单，中方已做好充分准备，将毫不犹豫，立刻进行大力度反击。 “我们感到，美方的行为十分无理。美方严重错判了形势，采取了极其错误的行动，这种行动的结果就是‘搬起石头砸自己的脚’。”高峰说道。 高峰说，我们注意到美方有关声明。在中美经贸问题上，中方立场已经讲得很清楚。我们不想打，但不怕打贸易战。 高峰表示，对美方声明我们将听其言观其行。如果美方不顾中方和国际社会反对，坚持搞单边主义和贸易保护主义行径。中方将奉陪到底，不惜付出任何代价，必定予以坚决回击，必定采取新的综合应对措施，坚决捍卫国家和人民的利益。 高峰指出，这次中美经贸冲突，是美方一手挑起，本质上是美单边主义对全球多边主义，美保护主义对全球自由贸易的挑衅。中方将继续扩大改革开放，维护多边贸易体制，推动全球贸易投资自由化和便利化。 “我们已经按照底线思维的方式，做好了应对美方进一步采取升级行动的准备，并已经拟定十分具体的反制措施。中
				原标题：魔都金融白领沉浮录：没亏在二级市场，却挂在了一级市场 文	仙逸 前言： “浪奔浪流，万里滔滔江水永不休” 上海，在《上海滩》的吸引下，一个让年轻人沸腾的地方。 李蒙也慕名而来，只是，他留下了一地伤心。 让我们看李蒙与他上海滩的的故事。 01萌芽

图 2-5 新闻格式

由图 2-5 可见，文件中含有 4 列数据，分别是 title（新闻标题）、time（新闻

发布时间）、url（新闻详细内容网址）以及 content（新闻内容），新闻内容所在列的数据较多。

2.5 本章小结

本章利用爬虫算法多线程并行的抓取三大专业财经网站的即时财经新闻。其中利用新闻 API 抓取新闻的标题、发布时间和 url；利用 lxml 模块抓取出新闻的详细内容。在抓取过程中还创建了 3 个线程，分别抓取 3 个新闻网站的新闻，大大增加了抓取新闻的效率。

3.基于 DBSCAN 聚类算法的新闻聚类分析

3.1 引言

在上一章中，已经从互联网上获取到了大量的新闻报道，在本文的样例中，抓取的是 2018 年 4 月 7 日前的大约 8000 条新闻。人为的从如此多的新闻中发现热点其实还是很难的。首先必须定义热点，有些人把热点定义成新闻的评论数多的新闻事件，也有些人把热点定义为在所有新闻中报道同一事件的新闻条数多的新闻事件。本文用的是第二种定义。通过文章文本的清洗、分词以及特征提取，然后使用聚类的方式来聚集同一事件的新闻，这样便能得到新闻的热点。

3.2 文本特征提取方法

为了标识一个文本，需要获取文本的特征。一般有三种常见的提取文本特征有方法：词频方法、文档频数法和 TF-IDF。

3.2.1 词频方法

词频即一个词在文档中出现的频率。在 sklearn 中，Countvectorizer 是一个通过词汇计数来将一个文档转换为向量方法。如何计数呢？就是把所有词库中的词当成列，并初始化向量为[0,0,…,0,0]，长度由词库的词的个数决定，如果文本中出现某一词，那么某一词的数就+1，最后得到词频的向量。

另外，Countvectorizer 可以设置一下几个参数，参数 min_df 指定词汇表中的词语在文档中最少需要出现的次数，如果小于出现的次数就不将该词统计进词频向量中；参数 max_df 则指定词汇表中的词语在文档中最多出现的次数上限，如果每篇文档都出现一个词，如果设置 max_df=0.95 的话，那么该词将不会被统计进词频向量中。Countvectorizer 也可根据语料库中的词频排序选出前 max_feature 的词，只得到 max_feature 维的特征。

实际上，词频较小的词其重要程度很多时候往往高于词频较大的词，所以仅仅统计词频的方法在实际中有一定的缺陷。

3.2.2 文档频数方法

词的文档频数即指在所有的文档集中包含该词的文档的频数，不过在特征提取的时候，往往需要去除文档频数达到某一阈值或者小于某一阈值的词的文档频数。文档频数的方法速度快，但是存在缺陷。假设某一稀有的词在某一类文档中出现，但

是由于这类文档的数量较小，因此会被去除其文档频数，导致其特征丢失。

3.2.3TF-IDF

在 TF-IDF 中，单词的重要性由两个因素共同决定，它与它在文档中出现的次数成正比，但它随着语料库中出现该词的频率越多而下降[17]。

在某一文档中，词频（TF）是指词在文档中出现的次数。TF 的值往往偏向于词汇量较大的文件，即长文件。（如果用词频来决定一个词是否重要，那么长文本中的单词相同的频率往往会比短文本中的频率更高）。

词的频率的计算可由此公式算得：

$$\text{词频（TF）}=\frac{\text{词在文档中的出现次数}}{\text{文档的总词数}} \qquad \text{式 3-1}$$

TF-IDF 的另一指标，当然就是逆文档频率（IDF）。IDF 是衡量单词总体重要性的指标，其值等于文档总数除以包含该单词的文档数量的商再取其商的对数。

词的逆文档频率的计算可由此公式算得：

$$\text{逆文档频率（IDF）}=\log\left(\frac{\text{语料库的文档总数}}{\text{包含该词的文档数}}\right) \qquad \text{式 3-2}$$

为了避免某词可能从来都没出现在所有的文档中，而导致被除数为零，一般分母用（包含该词的文档数+1）代替。

最后可以计算的出每个词的 TF-IDF 值为

错误!未找到引用源。 式 3-3

某词在某文档中是高词频，而在整个文档集中，该词又是低文档频数，那么该词可以得到一个较高权重的 TF-IDF 值。因此，TF-IDF 有助于降低常见的词语特征。

3.3 文本聚类方法

3.3.1 什么是聚类

聚类分析简称聚类，是一种无监督的学习方式，与监督学习不同的是它不需要对原来的数据打上标签，不用打上标签的数据来训练一种分类的模型，它仅仅利用某种距离计算将多个数据对象划分成集合的过程，使得每个集合便是一个簇，簇中的对象距离较小，彼此相似；但与其他的簇的对象之间的距离较大，相差较大[15]。

使用聚类往往是因为数据中没有提供类标号信息，但是仍需对其进行分簇。正因为如此，聚类方法在数据分析上很常用也很好用，它可以发现数据中事先未知的群

组。

3.3.2 几类常见的聚类方法

1）划分方法

聚类的大部分划分方法是基于距离的。

首先设定一个分区数 k，划分方法会随机初始化一个划分，然后采用一种迭代的重定位技术，把划分对象从一个簇移动到另一个簇中来改进聚簇效果[11]。使用划分方法进行聚类，为了得到全局最优解，算法所需的计算量往往是巨大的。实际上往往采用一种启发式方法，如 k-均值和 k-中心点算法，它们可以渐渐的逼近局部最优解。

2）层次方法

聚类的层次方法根据层次的分解形式，可以分为两类划分方式：凝聚和分裂[12]。凝聚的方法，是一种自底向上的方法，开始初始化每个对象作为一个单独的簇，然后逐次迭代的合并相近的簇的对象成为一个大簇，直到所有的簇的对象都合并为一个簇或者满足某种条件终止。

而分裂的方法，和凝聚方法相反，是一种自顶向下的方法，开始将所有的对象作为一个簇，然后通过多次迭代，将一个个大的簇分成更小的簇，直到每个对象都在单独的一个簇中或者满足某一条件终止。

层次聚类是基于距离或者是基于密度和连通性的，因为它考虑了子空间聚类。常见的凝聚的层次聚类算法有 AGNES，分裂的层次聚类算法有 DIANA。

3）基于密度的方法

基于距离的聚类都有一个缺点，就是只能发现球状簇，而很难发现其他任意形状的簇。于是人们开发了基于密度的聚类方法，其思想是：先发现密度超过某一阈值的点，然后使这些高密度点相近的就连成一片，进而生成各种簇。

基于密度的聚类方法可以过滤噪声和离群点，使之发现任意形状的簇。常见的基于密度的聚类算法是 DBSCAN。

4）基于网格的方法

基于网络的聚类是把所有的数据对象空间量化为有限个单元，形成一个网络结构，所有的聚类操作都基于网络上的网格上进行，数据相似的网格进行合并[18]。这种方法速度快，处理时间不是由数据对象个数决定的，而仅仅依赖于量化空间的每一维度的单元数。

3.4 基于密度的 DBSCAN 新闻文本聚类

为了实现新闻内容的聚类，需要先预处理新闻的内容，再利用聚类算法进行聚类。预处理包括：内容清理、分词、特征提取。

3.4.1 新闻内容的预处理

由于抓取下来的新闻中存在抓取失败的新闻，可能也含有一些重复的新闻，而且新闻内容中也可能存在一些对反映新闻内容没有作用的字词，所以需要对其进行清洗，清洗后再分词，提取出重要的词进行特征提取。新闻内容预处理框图如图 3-1 所示：

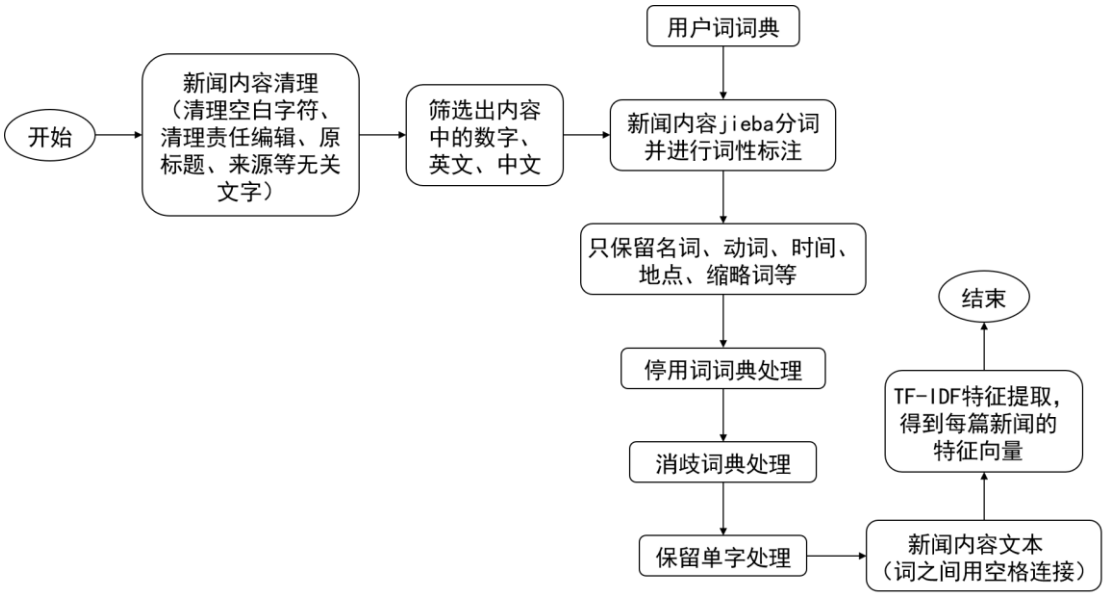


图 3-1 新闻内容数据预处理

由图 3-1 可见，预处理过程可以分为三部分：新闻内容清理、分词和词汇提取以及特征提取的过程。

1）新闻内容的清理

查看爬取下来的新闻内容，可以看到新闻中的空白符存在很多，有些地方有多个空白符连在一起，这是因为网站存在一些<p>标签里没有新闻内容的原因，因此需要清除这些多余的空白符；同时也可以看到，大多数新闻最后都有“责任编辑”、“返回搜狐，查看更多”等文字，新闻也有多处“来源”、“图片来源”、“资料来源”等文字，为了清理这些和新闻事件无关的词汇，可以利用正则表达式将其匹配，并替换为空字符。

对于新闻的聚类，多数标点符号也会影响聚类效果，所以需要将新闻内容的标点符号全部删除以便剩下的中文、英文和数字能很好的反映新闻内容。

2）基于 jieba 分词的新闻内容分词及重要词汇提取

对于清理后的新闻内容，需要对新闻进行分词操作，在分词前考虑分词工具分词效果可能不是十全十美，分词的结果可能分错误，所以本文通过观察新闻分词后的结果，自定义一个用户词词典，将分词工具的未登录词，即无法识别的词汇加入用户词词典，这样在分词前，先让分词工具对用户词词典进行分析，再使用分词工具便不会出现错误分词的现象。

分词工具分词的同时，是能给词加上词性的。这个很方便本文提取想要的相应词性的词。因为形容词、副词、助词等对新闻内容的特征贡献不大，所以在内容特征提取的过程中，只提取名词、动词、时间、地点、简称等词性的词语，以便能更好的分辨这则新闻。

不过，新闻内容提取这些词，也不一定能达到目的，为了更好的聚类效果，需要设置了停用词词典，即在分词之后，把在停用词词典中不需要的词在分词后的词中去除；同时，本文还使用消歧词典来将一个多个同一意思的词转化为同一个词；本文也创建了一个保留单字词典，只保留在词典中的单字，把其他所有的单字都去除，因为单字可能都没有特别大的意义，比如“是”、“会”等等。

3）基于 TF-IDF 的新闻内容特征提取

从上一步分词并筛选重要词之后，用剩下的这些词来进行特征提取，能更好反映新闻的特征。因为新闻中普通的词很多，但其重要性可能都比较小，本文使用 TF-IDF 来提取特征，这种特征提取能更好的提取新闻特点。

通过 TF-IDF 特征提取之后，在所有的新闻中，每篇新闻都有一个向量标识。向量上的每一个值都是一个词的 TF-IDF 值。其向量获得方式为首先统计出所有的词，把每个词当成向量的每一个维度，如果该文档中有某词，就在某词的维度上计算它的 TF-IDF 值；如果不存在某词，那么某词的维度上的值就为 0。用这种方式对所有的新闻进行特征提取，提取的结果是一个稀疏矩阵。

3.4.2 基于 DBSCAN 聚类算法的新闻聚类

即时新闻五花八门，花样繁多，而且数目参差不齐，存在大量的只有一两条只描述一个新闻事件的新闻。所以在使用 TF-IDF 进行新闻内容文本特征提取之后，使用了基于密度的 DBSCAN 算法进行新闻聚类是最为准确的做法，这种算法的特点可以发现任意形状的簇，同时可以过滤离群点，而不必把离群点分在某一个簇中，增加聚类的偏差。

DBSCAN 聚类过程如图 3-2 所示：

```
(1) 标记所有对象为unvisited;
(2) do
(3)       随机选择一个unvisited对象p;
(4)       标记p为visited;
(5)       if p的ε-邻域至少有MinPts个对象
(6)                 创建一个新簇C，并把p添加到C;
(7)                 令N为p的ε-邻域中的对象的集合;
(8)                 for N中每个点p1
(9)                           if p1是unvisited
(10)                                标记p1为visited;
(11)                                if p1的ε-邻域至少有MinPts个点，把这些点添加到N;
(12)                          if p1还不是任何簇的成员，把p1添加到C;
(13)                end for
(14)                输出C;
(15)      else 标记p为噪声;
(16) until 没有标记为unvisited的对象;
```

图 3-2 DBSCAN 聚类过程

由图 3-2 可见，聚类通过获取密度核心，由核心往密度较高的地方延展，将相近的新闻合并为同一个簇。在聚类的过程中需要输入三个参数：

1） D：一个包含 n 个对象的数据集

2） ε：半径参数

3） MinPts：邻域密度阈值

在计算过程中，使用“余弦相似度”来计算距离，所以本文在设置 ε 参数时，一般都设置为 0.3-0.5 之间，因为超过 0.5 的半径值会使不属于同类新闻聚在一起，小于 0.3 则无法识别相同事件的新闻。设置 MinPts 参数时，可以人为的假定一个阈值，假设有 10 条新闻报道同一事件就认为它就是一个热点，那么可以设置 MinPts 值为 10。

3.5 新闻文本聚类结果

通过新闻文本 DBSCAN 聚类的结果如图 3-3 所示。

由图 3-3 所见，程序运行后将聚类的产生的中间结果用 csv 文件保存下来，从保存下来的聚类结果可以看出，每条新闻都得到一个标签，标签为-1 的属于离群点，即介绍该新闻的新闻事件的其他新闻太少，无法构成热点。而其他标签的新闻则为某一热点的新闻，标签一样的新闻即为同一热点。

title	time	url	content	title_	content_	content_	label
【关注】朴	#######	http://www.sohu.com	【关注】朴	【关注】朴	韩国 中央	['韩国', '中	4
判了！ 韩国	#######	http://www.sohu.com	判了！ 韩国	判了！ 韩国	韩国 中央	['韩国', '中	4
18宗罪被判	#######	http://www.sohu.com	18宗罪被判	18宗罪被判	早晨 醒来	['早晨', '醒	4
华生宝能互	#######	http://finance.sina.com	华生宝能互	华生宝能互	华生宝 引爆	['华生宝', '	0
有一支特别	#######	http://finance.sina.com	有一支特别	有一支特别	战斗 佣军	['战斗', '佣	0
特朗普玩更	#######	http://www.sohu.com	特朗普玩更	特朗普玩更	北京 时间	['北京', '时	2
获刑24年罚	#######	http://www.sohu.com	获刑24年罚	获刑24年罚	今天下午 韩	['今天下午	4
高送转炒作	#######	http://finance.sina.com	高送转炒作	高送转炒作	导语 利用	['导语', '利	1
朴槿惠一审	#######	http://www.sohu.com	朴槿惠一审	朴槿惠一审	当地 时间	['当地', '时	4
韩国前总统	#######	http://www.sohu.com	韩国前总统	韩国前总统	日电 夏雪	['日电', '夏	4
重判24年，	#######	http://www.sohu.com	重判24年，	重判24年，	韩国 中央	['韩国', '中	4
朴槿惠一审	#######	http://finance.sina.com	朴槿惠一审	朴槿惠一审	韩国 中央	['韩国', '中	4
人民日报：	#######	http://finance.sina.com	人民日报：	人民日报：	李丽辉 陆娅	['李丽辉', '	2
中美贸易争	#######	http://www.sohu.com	中美贸易争	中美贸易争	林凛 贸易战	['林凛', '贸	2
华泰证券：	#######	http://finance.sina.com	华泰证券：	华泰证券：	来源 华泰	['来源', '华	2
变本加厉！	#######	http://www.sohu.com	变本加厉！	变本加厉！	美国 总统	['美国', '总	2
特朗普拟再	#######	http://www.sohu.com	特朗普拟再	特朗普拟再	美国 总统	['美国', '总	2
【老外谈】	#######	http://www.sohu.com	【老外谈】	【老外谈】	美国 当地	['美国', '当	2
变本加厉！	#######	http://www.sohu.com	变本加厉！	变本加厉！	当地 时间	['当地', '时	2
特朗普扬言	#######	http://www.sohu.com	特朗普扬言	特朗普扬言	美国 总统	['美国', '总	2
美对华新动	#######	http://www.sohu.com	美对华新动	美对华新动	消息 美国	['消息', '美	2
特朗普要求	#######	http://www.sohu.com	特朗普要求	特朗普要求	彭博社 报道	['彭博社', '	2
前万科独董	#######	http://www.sohu.com	前万科独董	王谦 李亮	['王谦', '李	0

图 3-3 新闻聚类结果

3.6 本章小结

新闻聚类是一个比较大的过程，需要进行数据清理、分词、特征提取、聚类等操作。本章实现了对新闻内容的聚类，在分词的时候使用了 jieba 分词工具，而特征提取的时候则通过 TF-IDF 提取新闻内容特征，最后使用基于密度的 DBSCAN 聚类算法进行聚类，并最终把聚类的标签在文件中用 label 列指出。

4.热点新闻话题排行

4.1 引言

上一章实现了对新闻的聚类，但是要想直观的观察新闻聚类的效果，还需要进行统计分析，为了清楚的了解每处热点的话题信息，可以把标题作为新闻的话题，对每条新闻的标题进行排行，得出话题的排名。

4.2 基于 TextRank 的自动文摘算法

TextRank 是一种基于图的用于文本的排序算法，基本思想来自于 Google 的 PageRank 算法[13]。类似于网页的排名，对于词语可得到词语的排行，对于句子也可得到句子的排名，所以 TextRank 可以进行关键词提取，也可以进行自动文摘。其用于自动文摘时的思想是：将每个句子看成 PageRank 图中的一个节点，若两个句子之间的相似度大于设定的阈值，则认为这两个句子之间有相似联系，对应的这两个节点之间便有一条无向有权边，边的权值是相似度，接着利用 PageRank 算法即可得到句子的得分，把得分较高的句子作为文章的摘要。

TextRank 算法的主要步骤如下：

（1）预处理：分割原文本中的句子得到一个句子集合，然后对句子进行分词以及去停用词处理，筛选出候选关键词集。

（2）计算句子间的相似度：在原论文中采用如下公式进行计算句子 1 和句子 2 的相似度：

$$\text{句子的相似度}=\frac{\text{两个句子都出现的词的数目}}{\log(\text{句子1中的词的数目})+\log(\text{句子2中的词的数目})} \quad \text{式 4-1}$$

对于两个句子之间的相似度大于设定的阈值的两个句子节点用边连接起来，设置其边的权重为两个句子的相似度。

（3）计算句子权重：

$$\text{句子1的权重}=(1-\text{阻尼系数})+\text{阻尼系数}\times\sum_{\text{与句子1相连的所有句子，如句子2}}\frac{\text{句子1和句子2的相似度}\times\text{句子2的权重}}{\text{所有与句子2相连的句子的边的权重和}} \quad \text{式 4-2}$$

由公式 4-2 可多次迭代计算直至收敛稳定之后可得各句子的权重得分。

（4）形成文摘：将句子按照句子得分进行倒序排序，抽取得分排序最前的几个

句子作为候选文摘句，再依据字数或句子数量要求筛选出符合条件的句子组成文摘。

4.3 热点新闻话题排行

4.3.1 基于统计的总体新闻的热点排行

因为先前已经获得每条新闻所在的类别，可以按照类别进行数量统计，从而获得各个类别新闻数量的排行，将数量最多的新闻所在的簇当成最热热点，第二次之，以此类推。

4.3.2 基于 TextRank 自动文摘算法的热点内部话题排行

利用 TextRank 算法进行热点内部话题排行的过程如图 4-1 所示：

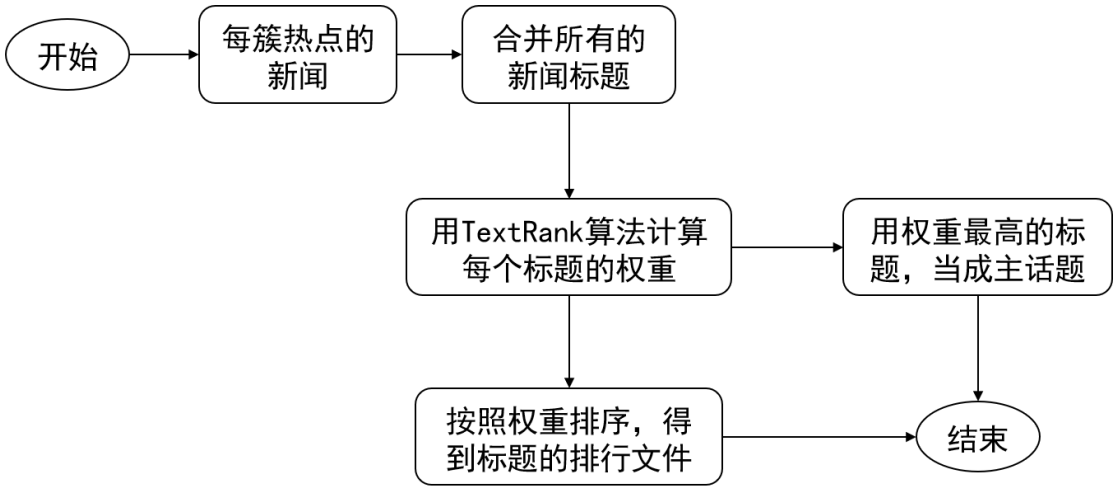


图 4-1 标题重要性排序

为了能使用 TextRank 算法，在本文中得将标题合并成类似于文章的样子。然后调用 TextRank 算法进行计算，得到达到字数要求的句子的排序。对于权重最高的句子，可以认为就是该处热点的主话题。

4.4 热点新闻话题排行实验结果

4.4.1 总体新闻的热点排行实验结果

总体新闻热点的排行结果如图 4-2 所示：

title	time	url	title_	content_	content_cu	label	rank
【关注】朴	#######	http://wwv	【关注】朴	韩国 中央	['韩国', '中	4	7
判了！ 韩国	#######	http://wwv	判了！ 韩国	韩国 中央	['韩国', '中	4	7
18宗罪被判	#######	http://wwv	18宗罪被判	早晨 醒来	['早晨', '醒	4	7
华生宝能互	#######	http://finar	华生宝能互	华生宝 引爆	['华生宝', '	0	2
有一支特别	#######	http://finar	有一支特别	战斗 佣军	['战斗', '佣	0	2
特朗普玩更	#######	http://wwv	特朗普玩更	北京 时间	['北京', '时	2	1
获刑24年罚	#######	http://wwv	获刑24年罚	今天下午 朴	['今天下午'	4	7
高送转炒作	#######	http://finar	高送转炒作	导语 利用	['导语', '利	1	6
朴槿惠一审	#######	http://wwv	朴槿惠一审	当地 时间	['当地', '时	4	7
韩国前总统	#######	http://wwv	韩国前总统	日电 夏雪	['日电', '夏	4	7

图 4-2 热点新闻排行结果图

由图 4-2 可见，每条新闻都通过类别大小进行排序，在新闻列表中，通过增加一列“rank”来代表新闻的热度，热度从 1 开始由高到低增大。不是热点的新闻 rank 标志为-1。

4.4.2 热点内部话题排行实验结果

热点内部的话题排行结果如图 4-3 所示：

```
万科A：钜盛华拟清算处置9个资管计划所持万科股份
宝能拟清算万科资管计划，曾考虑自己筹钱接盘
持股万科3年浮盈480亿，宝能拟清盘资管计划
钜盛华拟清算持股万科资管计划，稳定股价成首要考虑
钜盛华：拟清算处置9大资管计划所持万科A股份
钜盛华拟清算9个资管计划所持股份，早盘万科A大跌4%
刘姝威炮轰的宝能资管计划要清盘，万科股价一度下跌3.24%
宝能9大资管计划清仓万科11.42亿股，380亿谁当接
将清算9个资管计划，宝能系勾勒退出万科路径
宝能开始出货了，钜盛华减持11.42亿股万科股权赚百亿
宝能9资管将清仓万科:浮盈482亿落袋几多
万科股权再生变:宝能首提处置资管计划所持万科股份
钜盛华拟处置万科持股，宝能浮盈或接近500亿
宝能系拟处置大宗万科股权，前提是避免股价大幅波动
万科A股开盘跌3.98%，钜盛华拟处置万科持股
华生宝能互怼引爆500亿盈利走向：宝万大戏如何演绎
前海人寿2017年净利润下滑65%，全年未新售万能险
(附宝万大事记)
9大资管计划清仓万科11.42亿股
宝能行将处置的万科股权，或成为机构的香饽饽
皮海洲：宝能要溜走说法不确切，对万科股价影响不大
宝能回应华生：投资万科合法合规，与项俊波案没有关联
华生：项俊波被立案因猫鼠错位卷入宝能收购万科案
万科:宝能系拟通过大宗或协议转让减持万科A逾11亿股
前万科独董华生：宝能涉虚假增资，收购万科违规使用险资，宝能:投资合法合规
钜盛华资管清算万科股票，分析称对股价无直接影响
姚振华又刷屏：清仓11亿股万科，举牌1000天爆赚50
6大细节看宝能卖万科:赚多少谁接盘
宝能系“迈步”撤离，万科A股价小幅震荡
浮盈500亿+分红47亿+100%质押融资，宝能大赚后突然“引退”万科背后，一场更大的布局已经展开
有一支特别能战斗的佣军，叫万科独董
浮盈近500亿宝能向万科说分手，传宝能系将彻底退出
宝能终于要减持了
```

图 4-3 第 2 热点标题重要性排序结果

由图 4-3 的第 2 处热点的话题排序结果可以看出，这种方法确实能很好的得到一个热点的主话题，也能很好的从标题的排序中得出话题对该处热点的贡献。在第 2 处热点中，“钜盛华清算万科资管计划”这一话题较为热火。

4.5 本章小结

本章通过统计方法对总体新闻进行热点排行；利用 TextRank 算法进行热点内部话题的排序。结合这两种方法，可以发现各处热点并得到各处热点所涉及的话题。

5.基于 word2vec 技术及 k-Means 聚类的词汇聚类分析

5.1 引言

利用报道的同类新闻的数量大小获得了新闻的热点，然而为了分析新闻内容，还需要对新闻内容中的词汇进行分析。本文首先运用 word2vec 词嵌入的技术训练词向量模型，以此来获得新闻中筛选出来的重要的词的词向量，用词向量来对词汇进行聚类分析。

5.2 词汇的词向量的获取方法

5.2.1One-Hot 编码

one-hot 是一种能将词转换为向量的编码方式，也叫独热编码。

例如，有一句子：今天天气很晴朗。

如果词库一共出现就只有四个词“今天”、“天气”、“很”、“晴朗”。那么可以用词向量(1,0,0,0)来唯一表示“今天”这个词，用词向量(0,1,0,0)来唯一表示“天气”，用词向量(0,0,1,0)来唯一表示“很”这个词，用词向量(0,0,0,1)来唯一表示“晴朗”，以这种方式来唯一标识一个词语在有些地方确实很简单也很实用，但是这种方法有两个缺点，因为这种方法在表示词的时候，所使用的向量有太多的零值，如果用来表示一个句子，这样合成的矩阵显得极为稀疏。另一方面，用 one-hot 编码来标识词，忽略了句子中词的位置信息，如果还有其他句子如“你很美”，那么“很”因为之前已经编码过了，所以为了保证唯一性，“很”的词向量已经不能改变了还是原来的(0,0,1,0)，这种方式使得句子之间的词的关系不复存在了，也不能很好的计算词之间的相似度。

5.2.2word2vec 训练词向量

传统的 one-hot 编码已经不能很好的计算词的相似度，即词的词向量之间的距离。word2vec 技术恰恰可以解决这样的问题。word2vec 是 Google 于 2013 年开源的一个用于训练获取词向量的工具包[16]，它简单、高效，因此备受欢迎和关注。

word2vec 主要分为 CBOW（Continuous Bag-of-Words Model，连续词袋模型）和 Skip-Gram（Continuous Skip-gram Model，跳字模型）两种训练模式[14]。

CBOW 模型：用 context(w)去预测 w，目标是最大化概率 p(w|context(w))

Skip-gram 模型：用 w 去预测 context(w)，目标是最大化概率 p(context(w)|w)

context(w)是指词汇 w 的上下文，如果设置阈值为 N，那么 context(w)指的就是句子中 w 的前 N 个词和 w 之后的 N 个词。

word2vec 模型的两种训练模式如图 5-1 所示：

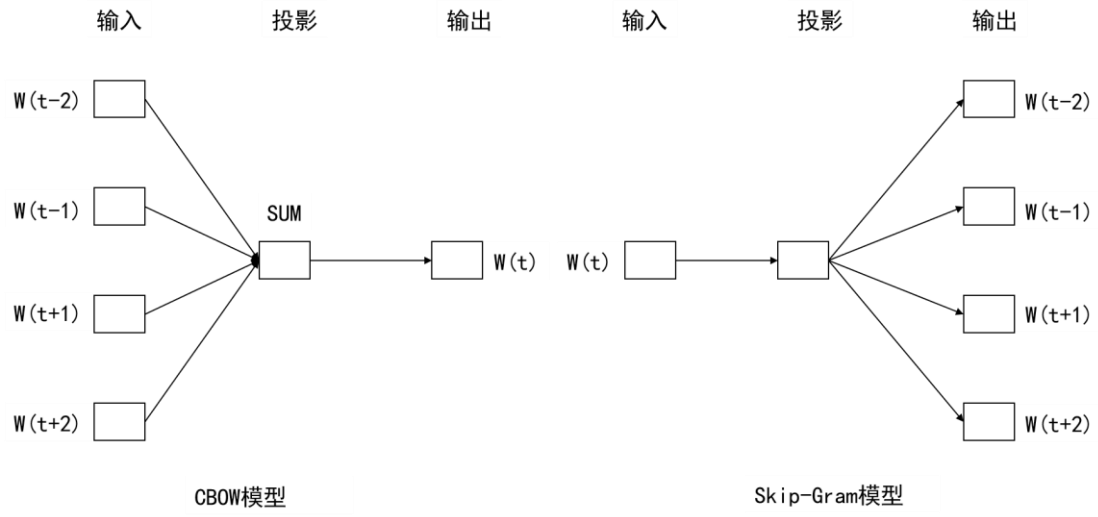


图 5-1 word2vec 两种训练模式

由图 5-1 可见，word2vec 的两种训练模式其实是很类似的，CBOW 模型是将单词 W(t)的上下文作为输入，从而预测单词 W(t)；而 Skip-Gram 模型是由单词 W(t)作为输入，从而来预测出单词 W(t)的上下文。

5.3 新闻热点内部词汇聚类

5.3.1 依据新闻内容训练 word2vec 词向量模型

由新闻内容训练 word2vec 词向量模型的过程如图 5-2 所示：

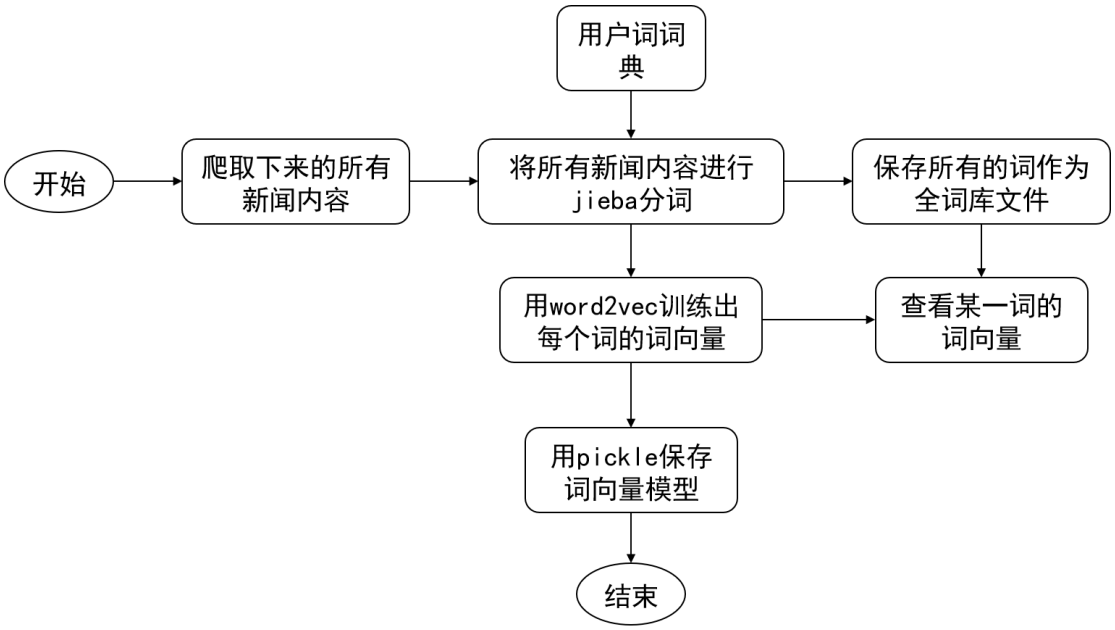


图 5-2 用所有的新闻内容训练词向量模型

由图 5-2 可见，本文运用了爬取下来的大约 8000 条新闻进行分词，接着进行训练，得到词向量模型，并保存下来。保存模型的运用 pickle 模块，该模块实现了基本的数据序列和反序列化。pickle 的序列化操作常常用来保存程序运行中的对象到文件中，永久存储；通过 pickle 的反序列化模块，又可以在程序中创建之前保存成文件的对象。

5.3.2 基于 k-Means 的词汇聚类

词汇的聚类过程如图 5-3 所示：

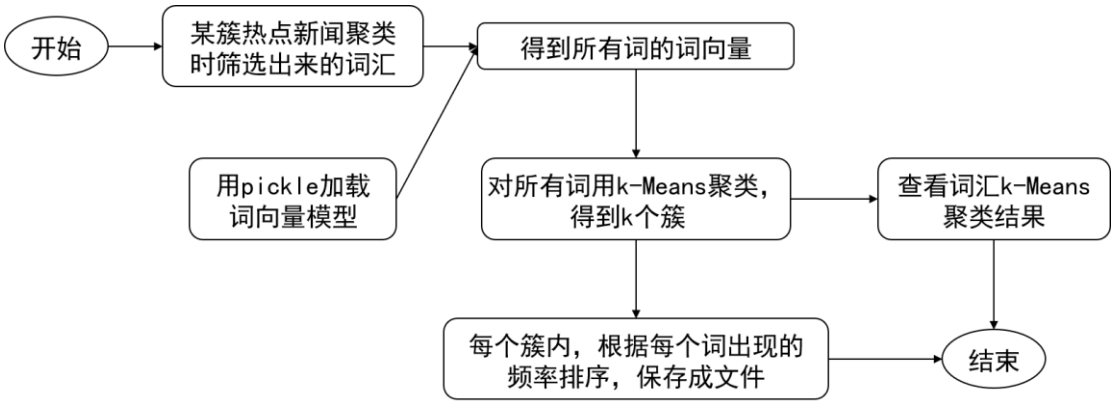


图 5-3 词汇聚类

由图 5-3 可见，本文对一个热点的所有新闻的词汇通过 word2vec 模型得到各个词的词向量，然后对其进行 k-Means 聚类。

k-Means 聚类的方法可以简单的如图 5-4 所示：

(1) 从D中任意选择k个对象作为初始簇中心;
(2) repeat
(3) 根据簇中的对象的均值，将每个对象分配到最相似的簇;
(4) 更新簇均值，即重新计算每个簇中的对象的均值;
(5) until不再发生变化;

图 5-4 k-Means 聚类过程

由图 5-4 可见，k-Means 聚类过程是一个迭代更新数据簇的中心的过程。其中输入是两个参数，一个是 k，即聚类设置的簇的数目；另一个是 D，即包含 n 个对象的数据集。输出则是 k 个簇的集合。

5.4 新闻热点内部词汇聚类实验结果

新闻热点内部词汇聚类的结果如图 5-5 所示：

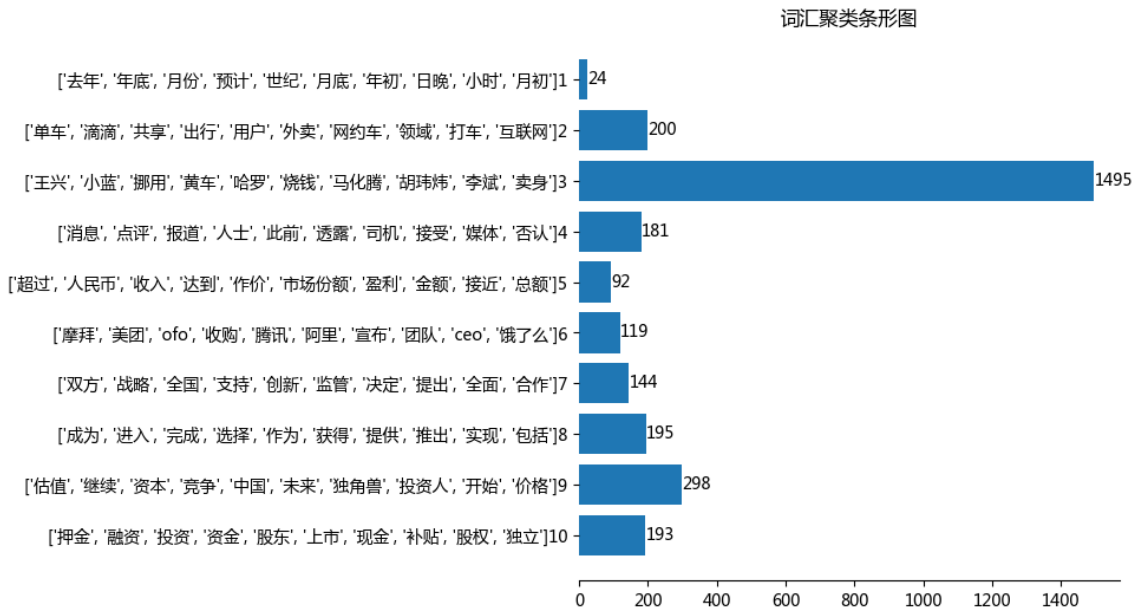


图 5-5 第 3 处热点词汇聚类条形图

由图 5-5 可见，如果对第 3 处热点“美团收购摩拜”事件的词汇进行聚类分析，由聚类生成的条形图的内容可以看出，聚类算法将词汇分成了 10 个类，每个类中的词汇彼此相似，最为明显的是第 6 类，基本都是代表机构的词汇，如“摩拜”、“美团”、“ofo”、“腾讯”等。说明这些机构对“美团收购摩拜”事件很有影响。

5.5 本章小结

本章通过 word2vec 技术训练了爬取下来的所有新闻的文本内容，从而训练出每个词的词向量，而且越相近的词，词向量之间的距离也越小，距离可以通过词向量之间的余弦相似度、欧式距离等计算得到。随后可以使用训练出来的词向量模型转化新闻文本中的词汇，对每一处热点中的新闻内容文本的词汇进行聚类。通过这种词汇聚类的方式，可以获得与热点有联系的相关人或事物的集合。

6.系统界面可视化及实验结果

本文主要是利用文本挖掘技术进行新闻热点关注问题分析，得出民众对某个领域的关切程度和社会需要解决的问题，了解当前的舆论焦点和民意，便于国家对舆论进行正确引导，使我们的社会更加安定和谐。

6.1 系统主界面

为了方便操作及理解，本文使用 Python 的 tkinter 模块设计了一个系统操作界面，其主界面如图 6-1 所示：

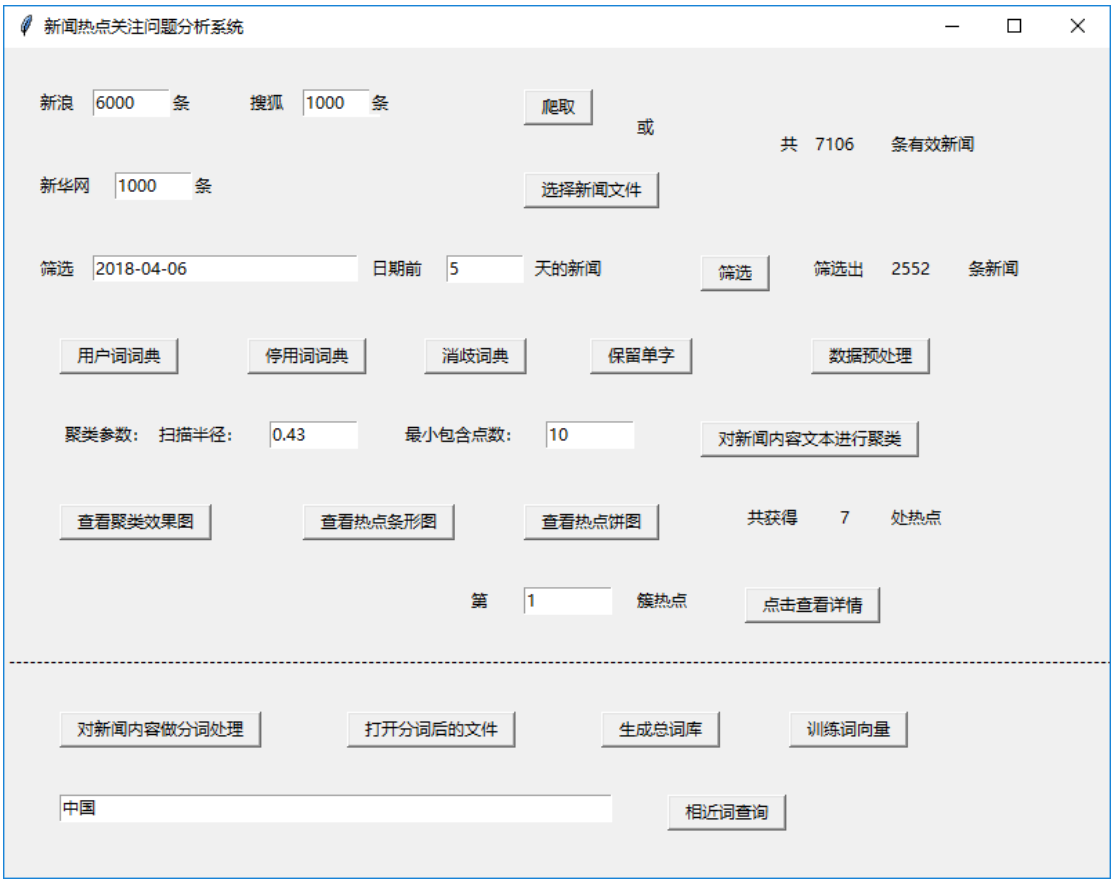


图 6-1 新闻热点关注问题分析系统界面

由图 6-1 可见，本系统主界面可以实现本文的各项任务：

1、在界面中，可以设置从新浪、搜狐、新华网等网站抓取的新闻条数，在本文测试样例中分别抓取这三个新闻网站的新闻为 6000、1000、1000 条；如果当前处于离线状态可以通过“选择新闻文件”按钮选择抓取后存储的离线新闻 csv 文件。

2、在界面中，可以根据新闻时间来筛选新闻，通过设置“筛选”按钮所在行的时间参数来实现。在本文中筛选 2018 年 4 月 1 日到 6 日的新闻。

3、在界面中，可以设置词典，通过用户词词典、停用词词典、消歧词典、保留

单字词典四个按钮来设置词典中的词。

4、可以进行新闻内容数据预处理，通过“数据预处理”按钮来实现内容清理、分词、特征提取等数据预处理操作。

5、可以设置 DBSCAN 聚类参数，通过设置“扫描半径”、“最小包含数”等参数来实现对新闻样例的聚类操作，本文设置的扫描半径为 0.43、最小包含数为 10。

6、可以查看聚类效果，通过“查看聚类效果图”、“查看热点条形图”、“查看热点饼图”按钮来实现。本文通过聚类得出了 7 个热点新闻簇，其聚类效果图、热点条形图、热点饼图分别如图 6-2、如图 6-3、如图 6-4 所示。

由图 6-2、条形图 6-3、饼图 6-4 可以看出，7 个新闻热点的热度不同，说明人们的关注度不同。在图中条形图只展示了每处热点频率最高的前 10 个词，而在饼图中只展示了每处热点频率最高的前 5 个词。可以看出在 2018 年 4 月 1 日到 6 日期间，人们对于“中美贸易战”的关注度最高，另外“钜盛华拟清算持股万科资管计划”、“美团收购摩拜”、“中财委首提结构性去杠杆”、“CDR 试点”、“高送转炒作被终结”、“朴槿惠干政门事件”等新闻事件也是这段时间的热点。

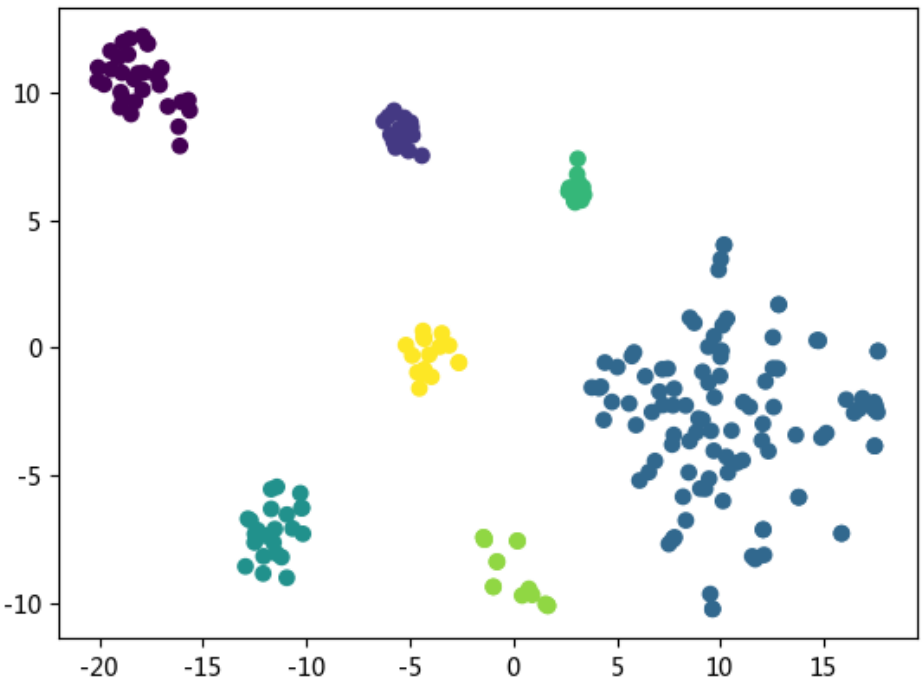


图 6-2 新闻聚类效果散点图

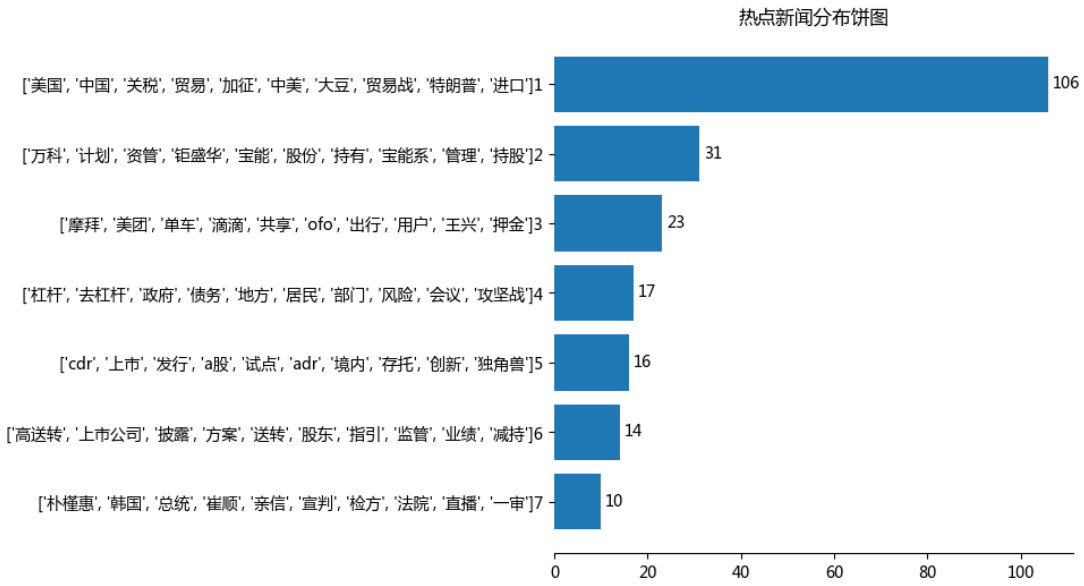


图 6-3 热点新闻排行条形图

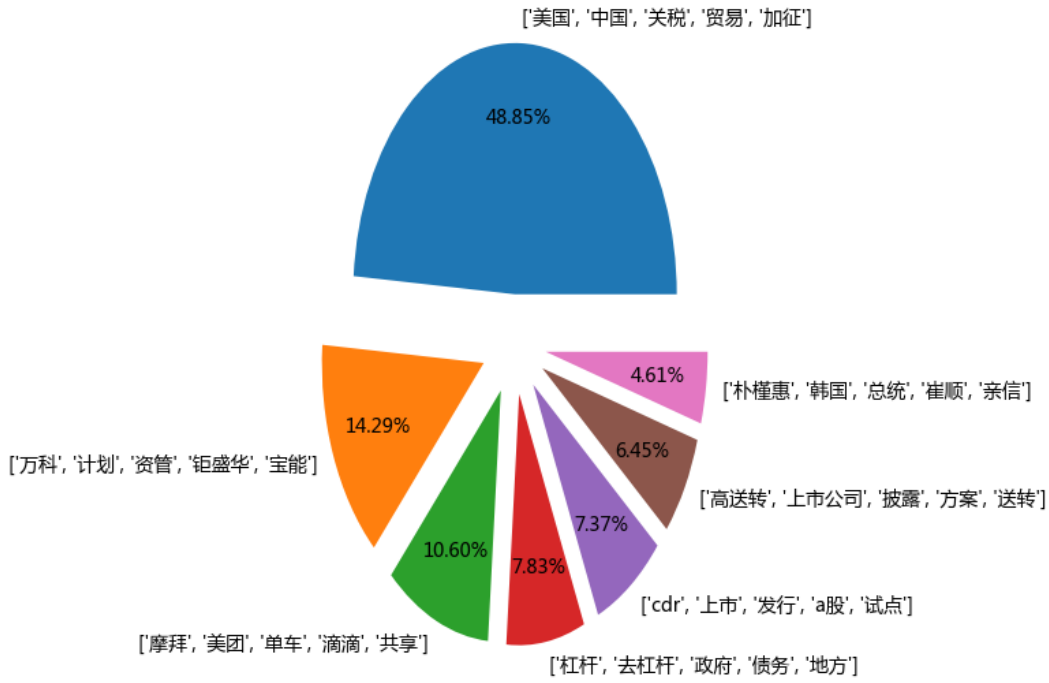


图 6-4 新闻热点分布饼图

7、在主界面底部，可以进行词向量模型的训练，通过对新闻的分词，训练词向量，可以获得每个词的词向量，可以通过“相近词”按钮查看相近的词。如“中国”可以查看得到与“我国”、“美国”、“日本”等与国家相关的名词相近。

6.2 系统热点详情界面

另外，为了更清楚的了解每处热点的新闻话题详情，本系统还设置了热点详情界面如图 6-5 所示：

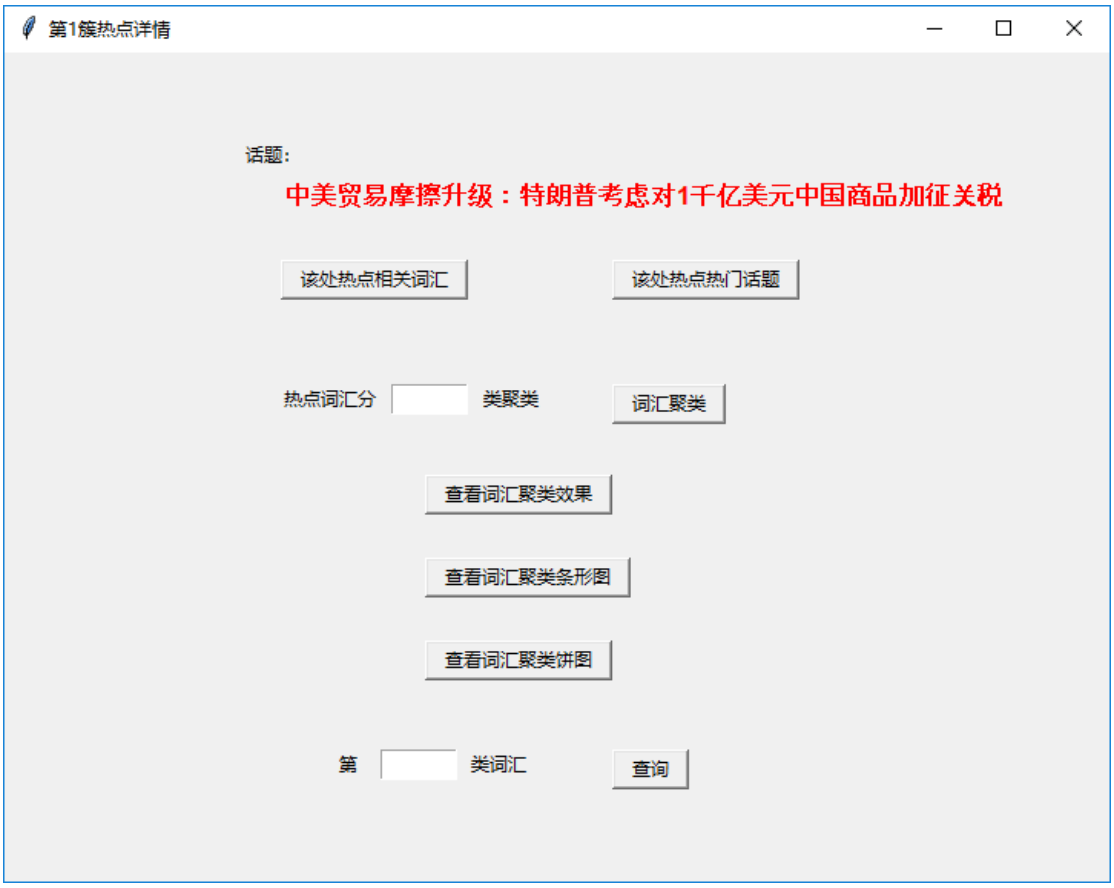


图 6-5 查看热点详情界面

由图 6-5 可见，在热点详情界面中，可以实现本文的各项任务：

1、可以查看热点的主话题，也可得到话题的排行、热点相关词汇。

2、在界面中，可以通过设置 k-Means 聚类的参数 k 值，即聚类的类别数，利用“词汇聚类”按钮对词汇进行聚类。

3、聚类之后可进行聚类结果的查看，同样可以查看聚类效果散点图、词汇聚类条形图、饼图。

4、在界面中，也可以详细查看某一簇词汇的所有词。

举最热的话题“中美贸易战”为例，通过“查看热点热门话题”按钮可以查看该热点话题排行结果如图 6-6 所示。

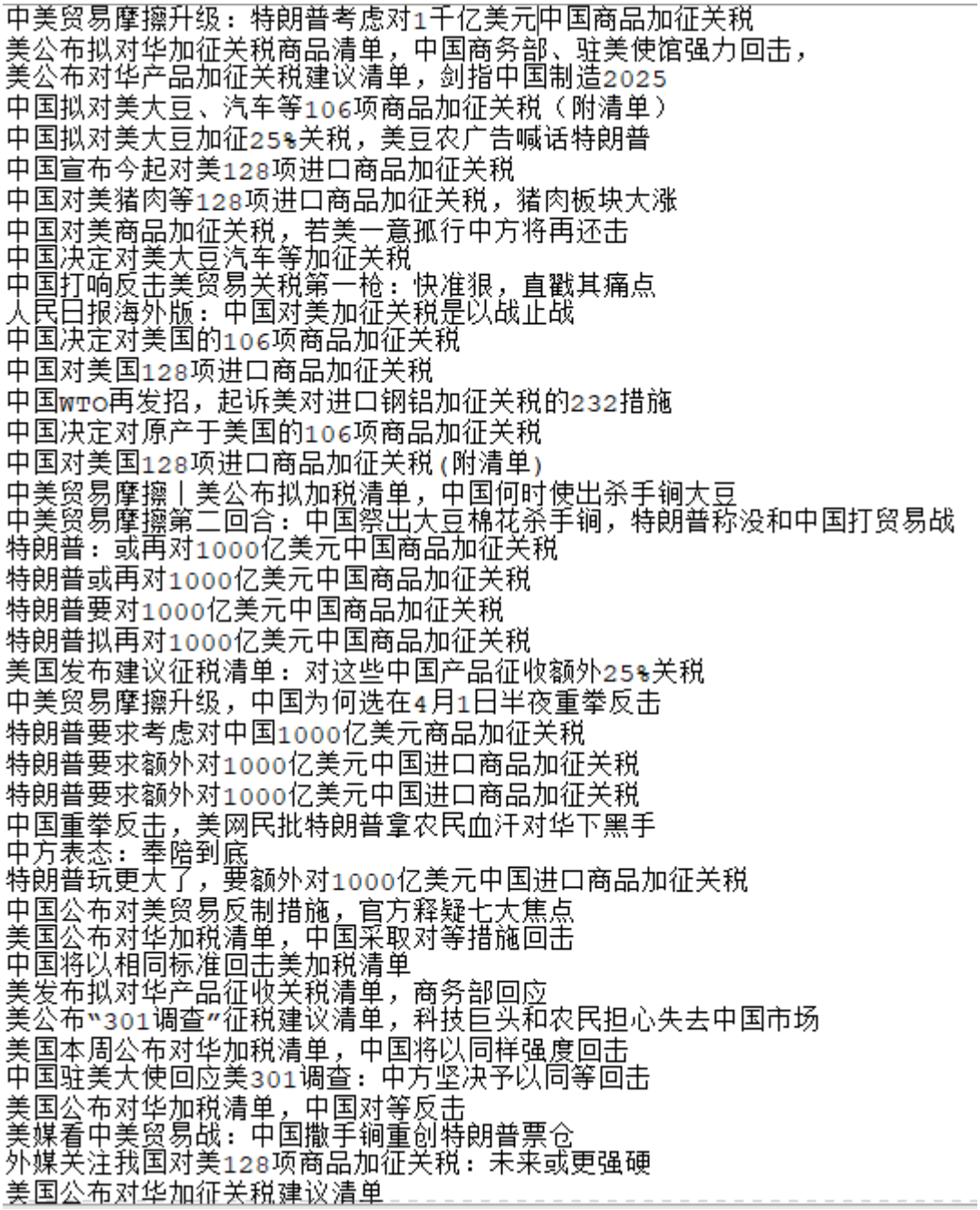


图 6-6 中美贸易战话题排序结果

由图 6-6 的排序结果可以看出各个子话题对“中美贸易战”这个热点的贡献。其中“特朗普对中国商品加征关税”在“中美贸易战”事件中最为热火，“中国对美国大豆、汽车等加征关税以反击”热度也很高。

另外可以通过“词汇聚类”按钮将词汇聚成 k 类，本文设置的类别数 k 为 17。可以查看词汇聚类效果，通过“查看词汇聚类效果图”、“查看词汇聚类条形图”、“查看词汇聚类饼图”按钮来实现，其聚类效果图、热点条形图、热点饼图分别如图 6-7、如图 6-8、如图 6-9 所示。

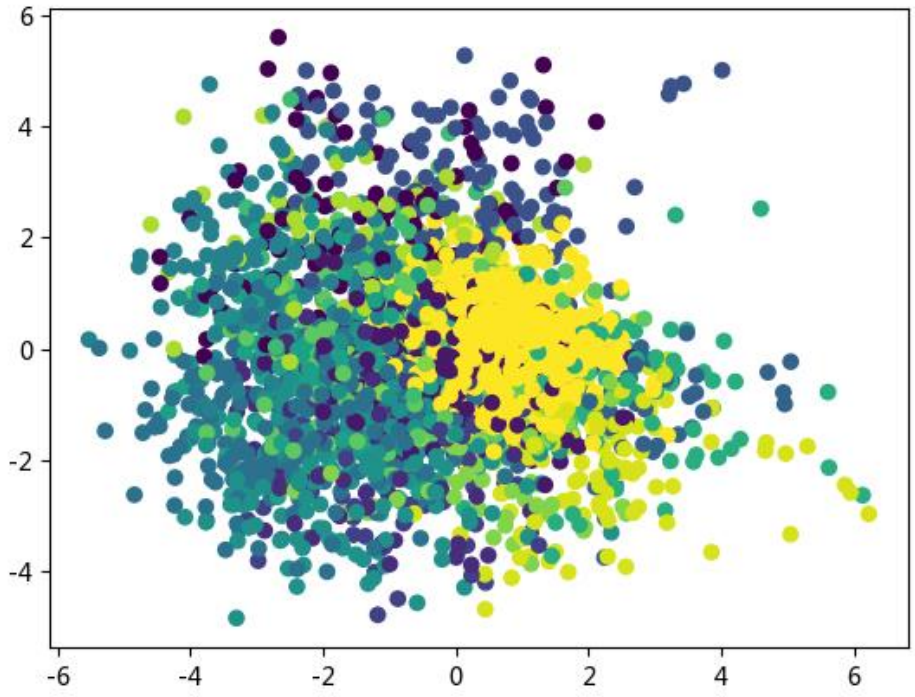


图 6-7 词汇聚类效果

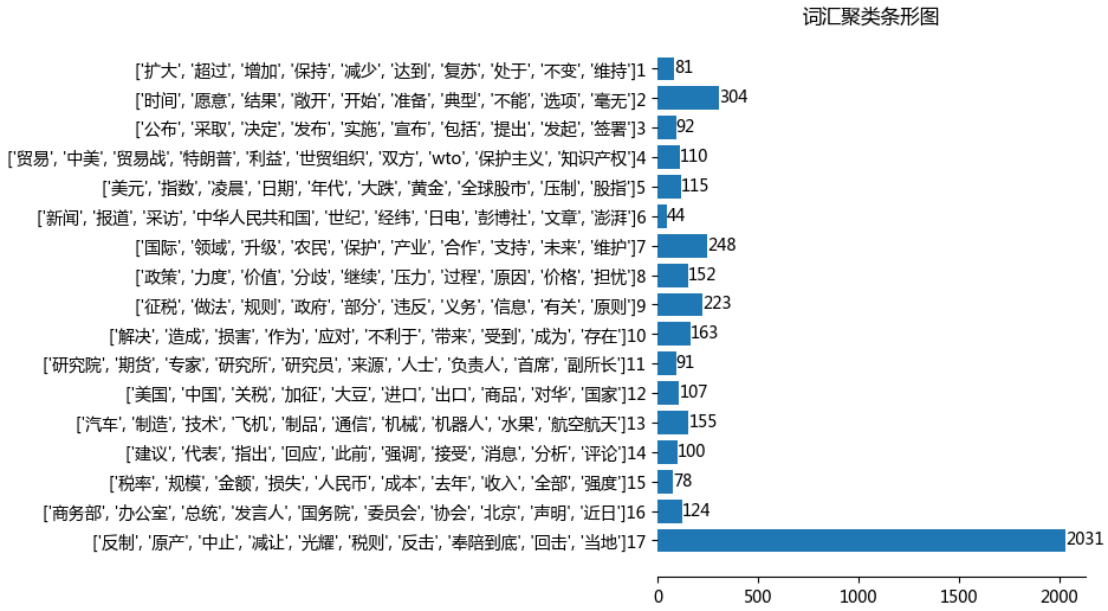


图 6-8 词汇聚类条形图

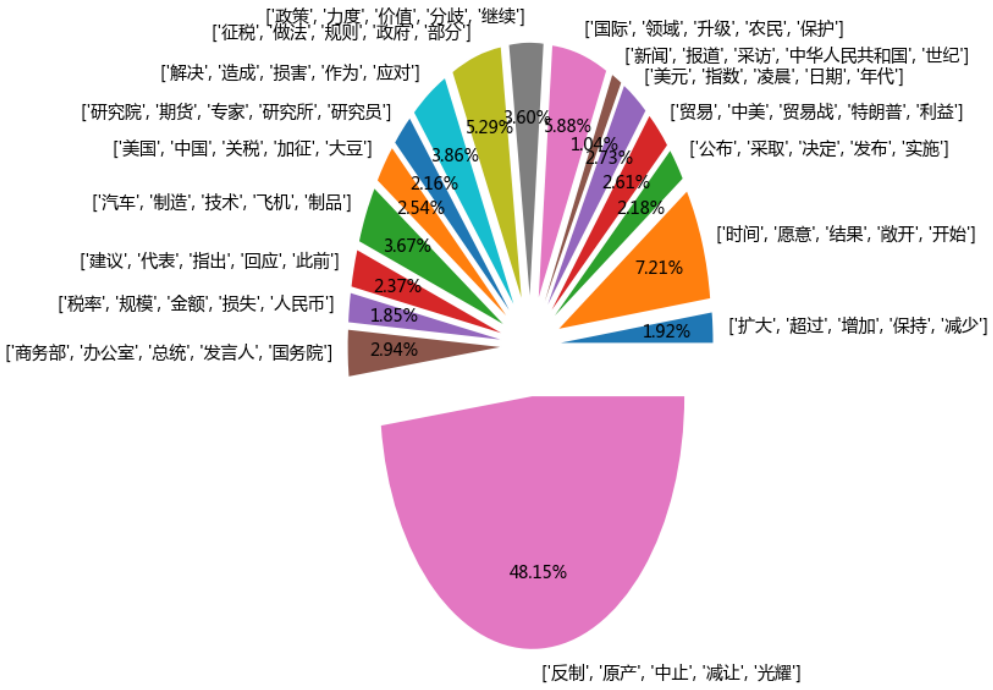
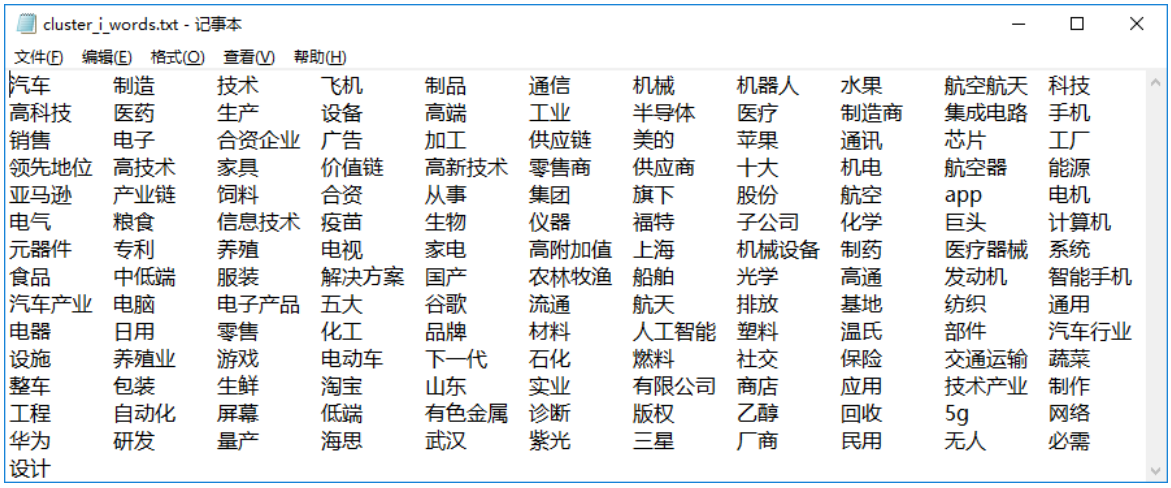


图 6-9 词汇聚类饼图

由图 6-7、图 6-8、图 6-9 可见，因为词汇量和词汇聚类的类别较大，所以聚类的效果一般。我们可以从聚类出的某些类别中可以得到，如类别 13，可见“中美贸易战”与汽车、制造、飞机、通信、机械等紧密相关。可查看所有 13 类的词汇如图 6-10 所示。


cluster_i_words.txt - 记事本

文件(F) 编辑(E) 格式(O) 查看(V) 帮助(H)

汽车	制造	技术	飞机	制品	通信	机械	机器人	水果	航空航天	科技
高科技	医药	生产	设备	高端	工业	半导体	医疗	制造商	集成电路	手机
销售	电子	合资企业	广告	加工	供应链	美的	苹果	通讯	芯片	工厂
领先地位	高技术	家具	价值链	高新技术	零售商	供应商	十大	机电	航空器	能源
亚马逊	产业链	饲料	合资	从事	集团	旗下	股份	航空	app	电机
电气	粮食	信息技术	疫苗	生物	仪器	福特	子公司	化学	巨头	计算机
元器件	专利	养殖	电视	家电	高附加值	上海	机械设备	制药	医疗器械	系统
食品	中低端	服装	解决方案	国产	农林牧渔	船舶	光学	高通	发动机	智能手机
汽车产业	电脑	电子产品	五大	谷歌	流通	航天	排放	基地	纺织	通用
电器	日用	零售	化工	品牌	材料	人工智能	塑料	温氏	部件	汽车行业
设施	养殖业	游戏	电动车	下一代	石化	燃料	社交	保险	交通运输	蔬菜
整车	包装	生鲜	淘宝	山东	实业	有限公司	商店	应用	技术产业	制作
工程	自动化	屏幕	低端	有色金属	诊断	版权	乙醇	回收	5g	网络
华为	研发	量产	海思	武汉	紫光	三星	厂商	民用	无人	必需
设计										

图 6-10 第 13 类词汇

由图 6-10 可见，在中美贸易这一热点问题上，人们对中国在飞机、制造业、通信、机械、机器人、航空航天科技等高新技术和高级制造方面的关切程度更加迫切，政府可以在这方面给予更多正确引导和支持。

7.结论与展望

7.1 结论

本文主要进行新闻热点关注问题分析，实现了热点新闻的发现和新闻热点的关联事物分类获取。主要实现了如下任务：

本文主要完成的工作包括以下几个方面：

第一、利用多线程技术和爬虫算法实现了对三大专业财经新闻网站，新浪财经、搜狐财经和新华网财经的新闻的并行爬取。

第二、利用 jieba 分词、TF-IDF 算法、DBSCAN 聚类算法分别对抓取的 2018 年 4 月 1 日到 6 日期间的财经新闻进行数据预处理、特征提取以及聚类分析。

第三、利用 TextRank 算法对每处热点进行标题重要程度的排行，结合聚类结果和标题排行可获得了财经新闻热度最高的是“中美贸易战”事件，之后分别是“钜盛华拟清算持股万科资管计划”、“美团收购摩拜”、“中财委首提结构性去杠杆”、“CDR 试点”、“高送转炒作被终结”、“朴槿惠干政门事件”。

第四、利用 word2vec 技术训练词向量模型，然后对某处热点的词汇进行 k-Means 聚类，获得热点的关联事物。例如“中美贸易战”热点涉及的领域有钢铁、汽车、农产品、制造、飞机、猪肉、制造业、通信、机械、机器人、水果、航空航天、科技等等。

分析结果说明，民众的关注热点是中美贸易，但在中美贸易这一热点问题上，人们对中国在飞机、制造业、通信、机械、机器人、航空航天科技等高新技术和高级制造方面的关切程度更加迫切，说明老百姓对国家科技强盛的热切期盼，政府可以在这方面给予更多正确引导和支持。

通过上述功能的实现，本文完成了毕业设计的预期任务，实现了预定目标。

7.2 未来工作展望

新闻热点关注问题分析是一个挺有难度，同时也挺有趣的研究，还有很多可以进行进一步研究和探索的地方：

1、本文的新闻热点问题分析最终分析出的热点词汇的分类中，相近词汇之间的关系由词汇常常出现的位置决定，“喜欢”和“讨厌”往往很接近，可以改进计算词汇的词向量，使之距离增大；可以设计一种计算词向量的方法使同一行业的常用

词的向量距离较小，不同行业之间的词之间的向量距离较大。通过这样的词向量模型分类词汇，可以很明显的知道该处热点所影响的行业，以及通过词汇的多少来测量影响的强度。

2、本文还可以往难度更高的语义分析方面进行研究，聚类得出热点之后，通过语义分析获取热点的话题短语，同样也可以得出热点的相关事件或者问题短语。

参考文献

[1].第41次《中国互联网络发展状况统计报告》发布[J].中国广播,2018(03):96.

[2] Allan, James, ed. Topic detection and tracking: event-based information organization[M]. Springer Science & Business Media, 2012.

[3] 王翔. 基于数据挖掘的热点新闻发现及系统方法研究[D].湖北工业大学,2017.

[4] 王馨. 网络新闻热点发现研究[D].河北大学,2015.

[5] 陈龙. 新闻热点话题发现及演化分析研究与应用[D].南京理工大学,2017.

[6] 赵旭剑. 中文新闻话题动态演化及其关键技术研究[D].中国科学技术大学,2012.

[7] 彭卫华. 互联网新闻热点挖掘系统的研究与实现[D].哈尔滨工业大学,2010.

[8] 刘林浩. 网络热点新闻事件挖掘和跟踪分析方法的研究与实现[D].中南大学,2010.

[9] 廖君华,孙克迎,钟丽霞.一种基于时序主题模型的网络热点话题演化分析系统[J].图书情报工作,2013,57(09):96-102+118.

[10] Lawson R. Web scraping with Python[M]. Packt Publishing Ltd, 2015.

[11] 王千,王成,冯振元,叶金凤.K-means 聚类算法研究综述[J].电子设计工程,2012,20(07):21-24.

[12] 王兰. 基于层次聚类的簇集成方法研究[D].河北大学,2010.

[13] 蒲梅,周枫,周晶晶,严馨,周兰江.基于加权TextRank的新闻关键事件主题句提取[J].计算机工程,2017,43(08):219-224.

[14] 周练.Word2vec 的工作原理及应用探究[J].科技情报开发与经济,2015,25(02):145-148.

[15] Jiawei Han, Micheline Kamber. Data Mining Concepts and Techniques[M]. 范明，孟晓峰译，机械工业出版，2001.

[16] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.

[17] Salton, G. and McGill, M. J. 1983 Introduction to modern information retrieval. McGraw-Hill, ISBN 0-07-054484-0.

[18] 蔡元萃，陈立潮.聚类算法研究综述[J].科技情报开发与经

济,2007(01):145-146.

致谢

在论文完稿之际，我谨向所有在工作、学习和生活中给予我关心和帮助的人们致以衷心的感谢！

在过去的四年里，我得到了许多老师、同学、亲人和朋友的帮助。首先，我要对我的毕设导师张艳玲副教授致以衷心的感谢。感谢张艳玲老师在毕业设计期间给予我耐心而又细致的指导。同时也是你的开拓创新精神，孜孜不倦的教学态度，始终激励着我努力学习、刻苦专研。从选题开始，每一步都是在张老师的指导下完成的，花费了张老师大量的心血，在此向张老师致以深切的谢意和由衷的祝福！

也要感谢我的公司里的朋友郭嘉贤、陈嘉盛、吴建军给我毕业设计的意见，教会我很多知识和技能，也带领着我踏入自然语言处理的行业中。同时也感谢我的同学兼同事郑奕讯在同我一起研究深度学习，感谢你对我学习上的帮助。

感谢我的父亲、母亲能容忍我的一切，是你们的关心和支持，支撑着我前行，给我动力，你们始终是我最强大的靠背。感谢我的女朋友，因为你的陪伴给我无限的温暖。

最后衷心感谢各位评委老师，感谢你们为审阅本文而付出辛勤的汗水。祝你们身体健康，万事如意！