文本复制检测报告单(全文标明引文)

№:ADBD2019R_2019060509254920190605093952442444570823　　检测时间:2019-06-05 09:39:52

检测文献:　蔡同波_15213526_新闻热点爬取与可视化的研究与实现
作者:　蔡同波
检测范围:　中国学术期刊网络出版总库
中国博士学位论文全文数据库/中国优秀硕士学位论文全文数据库
中国重要会议论文全文数据库
中国重要报纸全文数据库
中国专利全文数据库
图书资源
优先出版文献库
大学生论文联合比对库
互联网资源(包含贴吧等论坛资源)
英文数据库(涵盖期刊、博硕、会议的英文数据以及德国Springer、英国Taylor&Francis 期刊数据库等)
港澳台学术文献库
互联网文档资源
CNKI大成编客-原创作品库
个人比对库
时间范围:　1900-01-01至2019-06-05

检测结果

去除本人已发表文献复制比：0.6%　　跨语言检测结果：0%

去除引用文献复制比：0.6%　　总文字复制比：0.6%

单篇最大文字复制比：0.3%（报废磷酸铁锂动力电池破碎产物的风选特性分析及实验研究）

重复字数：[80]	总字数：[12696]	单篇最大重复字数：[42]
总段落数：[1]	前部重合字数：[38]	疑似段落最大重合字数：[80]
疑似段落数：[1]	后部重合字数：[42]	疑似段落最小重合字数：[80]

指　标：☐疑似剽窃观点　☑疑似剽窃文字表述　☐疑似自我剽窃　☐疑似整体剽窃　☐过度引用

表　格：0　　公　式：0　　疑似文字的图片：0　　脚注与尾注：0

（注释：无问题部分　文字复制部分　引用部分）

1. 蔡同波_15213526_新闻热点爬取与可视化的研究与实现　　总字数：12696

相似文献列表

去除本人已发表文献复制比：0.6%(80)　　文字复制比：0.6%(80)　　疑似剽窃观点：(0)

1	报废磷酸铁锂动力电池破碎产物的风选特性分析及实验研究	0.3%（42）
	何双华(导师：朱华炳) -《合肥工业大学博士论文》- 2018-04-01	是否引证：否
2	农商行政策对晋中市农业机械化发展影响研究	0.3%（36）
	杨志东(导师：李志伟) -《山西农业大学博士论文》- 2017-12-01	是否引证：否
3	初探高职院校网络德育环境的构建	0.3%（36）
	钮群; -《蚌埠学院学报》- 2012-08-20	是否引证：否

原文内容

新闻热点爬取与可视化的研究与实现

摘要

随着互联网和智能硬件的普及，各种信息数量正不断增加，每天从不同的话题之中都会产生大量的新闻数据，如何从众多的新闻数据中快速地了解信息的要点，有效提高获取重要信息的效率和减少在阅读时间上的消耗正变得越来越重要。通过对新闻数据进行分析，可以得出人们对某个领域的关切程度和社会需要解决的问题，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。本文以中国新闻网网站为例，通过爬虫技术抓取社会新闻版块数据，再对采集到的新闻内容进行处理及词频词性分析来计算关键字，最后借助数据可视化技术提供直观的呈现，得出热点所涉及的相关人或事物，以此分析出社会领域关注问题和需要解决的问题情况。

【关键词】网络爬虫；新闻；热点；Python；可视化

Research and implementation of news hotspot crawling and visualization

Abstract

With the popularity of the Internet and intelligent hardware, the amount of information is increasing. Every day, a lot of news data will be generated from different topics. How to quickly understand the main points of information from numerous news data, effectively improve the efficiency of obtaining important information and reduce the consumption of reading time is becoming more and more important. Through the analysis of news data, we can get the degree of people's concern in a certain field and the problems that society needs to solve, which is conducive to understanding the current focus of public opinion, helping the government to understand public opinion, and facilitating the correct guidance of public opinion by the state. Taking CNN website as an example, this paper grabs the data of social news section by crawler technology, then processes the collected news content and calculates the keywords by word frequency and part of speech analysis. Finally, it provides an intuitive presentation with data visualization technology, and obtains the relevant people or things involved in hot spots, so as to analyze the problems concerned and need to be solved in the social field.

[Key words] web crawler; news; hotspot; Python; visualization

目录

引言

随着5G、AI智能时代的到来，数据以爆炸的方式增长。各大新闻媒体产生的新闻数量正变得越来越多，一方面是用户数量不断增加，另一方面是因为各种物联网设备的出现给人们生活方式带来了很大改变，产生了许多与社会生活相关数据信息。2019年2月末，中国互联网络信息中心（CNNIC）发布了第43期《中国互联网发展统计报告》，报告显示截至到去年12月我国互联网普及率已经达到了59.6%，用户数量达到8.29亿，中国的互联网发展正迎来前所未有的发展变革，5G网络技术的普及将更彻底地改变人们的生活和生产方式，产业互联将触及经济和社会的各个方面。与传统的报纸、电视和广播相比，网络新闻在传播、互动和渗透力方面拥有更多的优势，使人们获取新闻信息更及时和便捷。

但是，随着信息量的不断增加，在同一新闻主题下拥有各种各样的新闻报道，大量的新闻出现在人们的生活中。由于在互联网上发布新闻对个人用户没有太多的限制，一些不法分子常常通过恶意“刷榜”、编造谎言恶意引导舆论方向，这可能会给社会带来不少的不良影响和危机。因此，很有必要检测分析当前新闻中的热门信息情况，为热点聚焦和防控提供一定的参考。本文通过可视化技术可以直观的计算统计出当前新闻的热点情况，为新闻决策和异常监控提供依据，以减小负面新闻对人们生活的影响，及时掌握事物的发展情况。

1 研究现状、总体任务和结构安排

1.1 新闻热点可视化研究现状

在大数据时代，面对复杂繁多的新闻信息，人们很希望能够快速从新闻报道中提取出关键的信息。数据可视化分析通过大数据的技术手段，针对数据的分布区域和在各行业的应用进行分析。长久以来，新闻数据可视化分析一直存在不少的问题，不仅需要使用先进的技术，还需要能够获得海量的数据，媒体组织应该共享彼此之间的信息，有效降低获取数据的成本，提高新闻数据的价值。另外，目前的新闻媒体发展仍受到关注度不高、新闻质量参差不齐的限制，应该努力提供创造性和个性化的新闻数据体验，采用多种模式来满足不同的用户，这也是未来数据可视化的发展目标和方向。

1.2 本文的主要工作

本文主要通过对新闻文本数据采集、分词和词频词性统计得到新闻热点词，并对热点数据提供可视化信息和分析，以帮

助公众了解重要事件关注情况。

本文主要研究的内容为：

1、利用多线程技术和爬虫算法实现了对中国新闻网的新闻的并行爬取。

2、利用中科院的ICTCLAS分词、交叉信息熵算法分别对抓取的2019年3月1日到31日期间的社会新闻进行数据处理、词频词性统计以及JSON化处理。

3、利用PyEcharts技术对数据结果进行分类统计并绘制可视化视图。

1.3 本文的组织结构

本文的结构共分有五个部分，具体安排如下：

第一部分，引言。介绍了本课题的研究背景和意义，研究现状、总体任务和结构安排。

第二部分，基于新闻API以及BeautifulSoup模块的数据抓取。介绍了网络爬虫的概念、采取的爬虫算法和抓取新闻的方式。

第三部分，基于交叉信息熵算法的关键字提取。详细介绍介绍分词前的各种处理、分词算法以及词频词性统计过程，并对抓取的新闻数据进行JSON化处理。

第四部分，实验结果及数据可视化。设计并实现数据的可视化呈现数据背后的深层含义，得出研究的最终结果。

第五部分，总结。通过对数据的可视化结果分析、得出研究的结论。

2 基于新闻API以及BeautifulSoup模块的数据抓取

新闻API作为新闻数据的入口，是后台向前端传输数据的通道。而对应此规则的数据信息则是爬虫按照某一特定算法抓取的目标新闻内容，新闻爬虫依次通过发送API请求、获取和解析新闻页面数据、最后将格式内容存储进数据库中，直到页面数据遍历采集完毕才停止执行。

2.1 基于新闻API的新闻链接抓取

图 2 SEQ 图 * ARABIC 1 中国新闻网社会新闻API从中国新闻网网站上，获取的新闻API如图2-1所示：

图 2 SEQ 图 * ARABIC 2 新闻URL抓取

使用python的requests提供的get()方法我们可以非常简单的获取的指定网页的内容，使用BeautifulSoup的select方法得到该页面下所有标签中的所有URL，并将URL存入到url_list的表中。如图2-2所示：

2.2 基于BeautifulSoup模块的新闻内容的抓取

本文的爬虫系统主要由三个部分构成，分别是抓取，分析和存储。程序首先通过访问API接口请求解析新闻数据，然后根据定义的格式信息分析存储，新闻的整体数据抓取过程如图2-3所示：

图2-3 新闻整体数据抓取流程图

抓取到网页的内容后，要做的就是提取出想要的内容。首先导入BeautifulSoup库，使用BeautifulSoup可以非常简单的提取网页的特定内容。如图2-4所示：

图2-4 新闻内容抓取

数据库部分的存储集合包括两个部分，分别是新闻链接url_lists、新闻内容item_info集合，两个集合的详细设计如下表：

表2-1 新闻内容item_info集合

属性名含义类型说明

_id 唯一标识 String

Link 新闻链接 String URL

Content 新闻内容 String

Year 年份 String

Month 月份 String

Day 日期 String

表2-2新闻链接url_lists集合

属性名含义类型说明

_id 唯一标识 String

Links 新闻链接 String URL

2.3 实验结果

本文样本数据抓取的时间是2019年3月份的所有新闻数据。由图2-5可见，url_lists中共有433.6KB的链接数据，item_info中共有17.6MB的新闻内容数据。

图2-5 新闻数据大小

按照时间顺序抓取的新闻链接集合如图2-6所示：

图2-6新闻链接url_lists集合

获得新闻链接后再根据链接进行深层次的内容解析获得新闻内容，如图2-7所示：

图2-7新闻内容item_info集合

由图2-6、2-7可见，url_lists文件中含有2列数据，分别是id（标识符）、links（链接），item_info中含有6列数据，分别是id（标识符）、link（链接）、content（新闻内容）、year（年）、month（月）、day（日），新闻内容所在列的数据较多。

2.4 小结

本章利用爬虫算法多线程抓取中国新闻网网站的社会新闻。主要实现页面链接和新闻的爬取。通过某一天的主页面URL得到该页面下的所有URL，并将URL存入到url_list的集合中去，使用BeautifulSoup模块从单个新闻的链接中得到这个链接中的新闻内容存入到item_info集合中。

3 基于交叉信息熵算法的关键字提取

在上面第三章中，已经通过爬虫程序获取到了2019年整个3月份大约5161条新闻数据，通过新闻文本数据的分词和词频

词性统计分析，便能很容易获得全部数据中的热点情况。

3.1 交叉信息熵算法

在信息理论中，熵是不确定事件的度量，是信息中拥有的的平均信息量。一个词叫做关键词，原因在于这个词能搭配很多很丰富的其他词语表达更重要的含义，于是便产生了对于一个词的信息熵定义，数学表达如下：H(w)=plogp

W代表单词，p代表单词周围出现的不同单词的数量。例如，在一篇文章中，现在有两个XWY，一个ZWS，那么W的左信息熵是：-23log23-13log13

其中，23表示关键句X在3个句子中出现两次，而Y只出现一次，因此Y的信息熵为23，对于一个ZWS来说，他们两个的信息熵则是相同的。如果是以XWS，YWS例，那么W的右边信息熵的值是0，因为是-log1。

对文档中所有的词语计算上下文信息熵的情况，如果一个词的上下文信息熵都比较大，那么这个词就很容易被判断为关键词。

3.2 分词和词频词性统计

新闻数据在实现分词和词频词性统计时，首先需要读取数据并写入文本文件中，如图3-1所示：

图3-1 新闻文本数据

由于中科院计算所的NLPIR分词系统只能分割500K以内的数据(实际只有490K),所以不得不对大容量的txt文件进行分割成500K以下的文件。输出文件500K大小的文件，在GBK编码中一个汉字占2个字节，故定义文本大小lensize=256000，表示读取256000个字符，这些字符在最后存在GBK文件里的时候文件大小为：256000*2/1024 = 500K，实际情况下可能里面有一些英文字符所以实际大小为490k左右，切分后的文本文件如下图3-2所示：

图3-2 文本分割

另外，由于NLPIR分词系统只支持GBK的编码，还需要将UTF-8格式文本转换成GBK格式。相关代码如图3-3所示：

图3-3 文本编码转换

3.3 实验结果

使用中国科学院的ICTCLAS分词系统对新闻文本数据分词计算操作的结果如图3-4所示：

图3-4分词原始数据

得到原始数据后我们并不能立即进行可视化的显示，还需要进一步借助Excel进行处理，这里主要是将分割后的关键字内容进行合并，如图3-5所示：

图3-5 Excel分解后的数据

但是经Excel处理后的新闻数据依然不能直接使用，需要转成JSON文本文件才行，通过BeJSON在线转换，可以将Excel表格数据转至JSON文本文件，结果如图3-6所示：

图3-6 JSON化的分词数据

3.4 小结

新闻关键字提取是一个比较大的过程，需要对新闻文本数据进行文本分割，本章实现了对新闻内容的抽取，在分词的时候使用了中国科学院的ICTCLAS分词系统，并做了新闻内容的词频和词性结果统计，最后将分词文本转成JSON格式的数据文件，为下一步的可视化实现提供数据支撑。

4 实验结果及数据可视化

本文主要是利用文本词频统计技术进行新闻热点问题可视化分析，希望可以通过可视化技术准确直观的显示当前媒体报道中的各类热点情况，以此为个人、社会和国家提供新闻热点的关注情况，为社会舆论更好发展提供一定的指导。

4.1 实验结果

结果显示2019年整个3月份，产生了大约17 M的新闻数据（如图4-1所示），共抓取了约5161条新闻信息，总计大约608万字（如图4-2所示），平均每天产生166多条新闻信息。

图4-1 新闻文本大小

图4-2 新闻数据总量

然后，通过中国科学院的ICTCLAS分词系统对这608万字的新闻数据文档处理，得到了28万3千多的关键词数据，如图4-3所示：

图4-3 分词数据总量

4.2 数据可视化

为了可视化的显示提取后的数据，需要借助PyEcharts技术动态的将JSON化后的新闻数据显示在网页上，如图4-4所示，这是3月份在社会新闻版块重要的热点名词统计情况。

图4-4 三月份热点名词

对应的热点名词数据视图表如下：

表 4-1 三月份热点名词

{'words': '检察机关', 'nums': 462.19}
{'words': '医护人员', 'nums': 428.08}
{'words': '纪委监委', 'nums': 391.02}
{'words': '脱贫攻坚', 'nums': 384.82}
{'words': '微博', 'nums': 362.07}
{'words': '乡村振兴', 'nums': 357.79}
{'words': '官方微博', 'nums': 356.35}
{'words': '搜救', 'nums': 350.42}
{'words': '野生动物', 'nums': 349.31}
{'words': '食品安全', 'nums': 345.85}
{'words': '引领', 'nums': 335.61}

{'words': '主管部门', 'nums': 308.41}
{'words': '客服', 'nums': 305.03}
{'words': '垃圾分类', 'nums': 282.68}
{'words': '志愿服务', 'nums': 273.25}
{'words': '公示', 'nums': 264.36}
{'words': '履职', 'nums': 263.86}
{'words': '法律法规', 'nums': 255.11}
{'words': '告诉记者', 'nums': 246.86}
{'words': '空气质量', 'nums': 243.44}
{'words': '澎湃新闻', 'nums': 235.78}
{'words': '民营企业', 'nums': 235.5}
{'words': '赏樱', 'nums': 234.84}
{'words': '新闻发布会', 'nums': 232.34}
{'words': '食材', 'nums': 228.88}
{'words': '生态环境', 'nums': 225.82}

根据以上信息我们可以清晰地看到占据了新闻名词榜的榜首的是“检察机关”，其中“医护人员”和“纪委监委”分别占据第2、3名。 在过去的3月份，正是我们国家举行第十三届全国人民代表大会会议的时候，作为国家的权力机关与重要政府部门，出现频率之高反应其在国家政治生活中的重要作用，与此同时“纪委监委”作为重要的监察部门同样受到广泛的关注。对于关键词医护人员的出现也不奇怪，2019年国家卫健委对医护人员提出了更多更具体的要求，医护人员的薪资待遇方面也即将迎来改革，在各项突发事件中，医护人员首当其冲救死扶伤，在社会生活中发挥了巨大的作用。

那么，如何发现某一时间新闻是发生在哪些人身上，或者说某一事件的参与者受到的关注量是怎样的，可以很清晰的从图4-5获得想要的结果。

图4-5 出现最多的人名

对应的人名数据视图表如下：

表4-2 出现最多的人名

{'words': '曹园', 'nums': 439.66}
{'words': '王磊', 'nums': 323.06}
{'words': '王某', 'nums': 213.76}
{'words': '陈某', 'nums': 177.56}
{'words': '孙强', 'nums': 149.13}
{'words': '刘某', 'nums': 136.96}
{'words': '王乃生', 'nums': 104.22}
{'words': '张英', 'nums': 94.12}
{'words': '沈巍', 'nums': 93.98}
{'words': '刘瑞强', 'nums': 87.5}
{'words': '王华州', 'nums': 84.76}
{'words': '张立', 'nums': 84.47}
{'words': '胡瑞娟', 'nums': 75.49}
{'words': '胡尹萍', 'nums': 74.34}
{'words': '刘瑞芹', 'nums': 73.16}
{'words': '胡某', 'nums': 73.02}
{'words': '李建明', 'nums': 71.1}
{'words': '陈伟起', 'nums': 70.01}
{'words': '罗应玖', 'nums': 68.94}
{'words': '陈培新', 'nums': 68.33}
{'words': '刘海蛟', 'nums': 65.61}
{'words': '张玉环', 'nums': 65.26}
{'words': '陈宗祥', 'nums': 63.79}
{'words': '黄志发', 'nums': 62.93}
{'words': '方彦格', 'nums': 62.28}
{'words': '唐超', 'nums': 61.11}
{'words': '李海涛', 'nums': 60.55}
{'words': '李云峰', 'nums': 60.31}
{'words': '潘志平', 'nums': 59.0}
{'words': '刁继龙', 'nums': 55.96}
{'words': '陈静瑜', 'nums': 54.18}
{'words': '李某', 'nums': 52.47}
{'words': '郑检凤', 'nums': 51.39}
{'words': '黄淑芬', 'nums': 50.71}
{'words': '袁长生', 'nums': 49.44}
{'words': '蔡徐坤', 'nums': 47.49}

{'words': '张静芬', 'nums': 45.79}
{'words': '尹月娥', 'nums': 44.93}
{'words': '费敏秀', 'nums': 44.91}
{'words': '李雷', 'nums': 38.77}

通过上面的词云图我们发现王某、刘某、胡某在本月份出现的频率最高，同时这也是新闻媒体在新闻报道中最常用的指代性词语，继续探究其本后的情况，其实也间接表明中国的姓氏分布状况，因为王姓、刘姓的在全国姓氏人口中占据更大的一部分，所以相关联的姓氏人名新闻出现的概率便会更大一些。透过图中具体的人名信息，可以很清楚的了解到这些人大多是3月份重大热门事件的新闻人物，媒体报道的数量自然也就更大一些。

如图4-6是新闻在全国各省市的统计分布图：

图4-6 各省市新闻出现情况

通过不同的颜色，我们可以看到在不同的地方发生的热门事件。红色表示该地区的新闻频率非常高。北京，上海，重庆和西藏位居前列，而安徽，黑龙江和吉林的频率远远不够，继续看看每个城市的情况，如图4-7所示：

图4-7全国各城市出现频率

对应的全国省市数据视图表如下：

表4-3全国各城市出现频率

{'words': '北京', 'nums': 854.58}
{'words': '上海', 'nums': 276.16}
{'words': '西藏', 'nums': 223.39}
{'words': '广州', 'nums': 88.38}
{'words': '重庆', 'nums': 78.85}
{'words': '海林', 'nums': 64.55}
{'words': '北京市', 'nums': 62.56}
{'words': '成都', 'nums': 55.4}
{'words': '杭州', 'nums': 45.52}
{'words': '武汉', 'nums': 40.05}
{'words': '广东', 'nums': 39.0}
{'words': '江苏', 'nums': 34.26}
{'words': '浙江', 'nums': 33.65}
{'words': '四川', 'nums': 30.42}
{'words': '新疆', 'nums': 23.36}
{'words': '深圳', 'nums': 22.98}
{'words': '湖南', 'nums': 22.43}
{'words': '南京', 'nums': 19.71}
{'words': '西安', 'nums': 19.51}
{'words': '海南', 'nums': 13.37}
{'words': '山东', 'nums': 12.83}
{'words': '盐城', 'nums': 11.14}
{'words': '甘肃', 'nums': 10.36}
{'words': '山西', 'nums': 10.25}
{'words': '香港', 'nums': 9.38}
{'words': '云南', 'nums': 9.14}
{'words': '广西', 'nums': 8.66}
{'words': '福州', 'nums': 7.57}
{'words': '贵州', 'nums': 6.83}
{'words': '武都区', 'nums': 6.18}
{'words': '拉萨', 'nums': 5.57}
{'words': '兰州', 'nums': 4.78}
{'words': '响水县', 'nums': 4.68}
{'words': '台湾', 'nums': 4.1}
{'words': '陕西', 'nums': 3.64}
{'words': '郑州', 'nums': 3.6}
{'words': '宁夏', 'nums': 3.42}
{'words': '英德', 'nums': 2.83}
{'words': '内蒙古', 'nums': 2.68}
{'words': '顺德', 'nums': 2.6}
{'words': '柳州', 'nums': 2.27}
{'words': '福建', 'nums': 2.27}
{'words': '河南', 'nums': 2.24}
{'words': '南海', 'nums': 2.22}
{'words': '重庆市', 'nums': 2.14}
{'words': '个旧', 'nums': 2.03}

{'words': '牡丹江市', 'nums': 1.98}
{'words': '易县', 'nums': 1.83}
{'words': '中山', 'nums': 1.66}
{'words': '广州市', 'nums': 1.65}
{'words': '黄山', 'nums': 1.63}
{'words': '云浮', 'nums': 1.55}
{'words': '天津', 'nums': 1.48}
{'words': '南宁', 'nums': 1.46}
{'words': '大连', 'nums': 1.43}
{'words': '石家庄', 'nums': 1.31}
{'words': '昆山', 'nums': 1.29}
{'words': '澳门', 'nums': 1.25}
{'words': '吉林市', 'nums': 1.11}
{'words': '沈阳', 'nums': 1.09}
{'words': '河北', 'nums': 1.08}
{'words': '盐城市', 'nums': 1.04}
{'words': '石家庄市', 'nums': 1.03}
{'words': '湖北', 'nums': 1.03}
{'words': '常州', 'nums': 1.02}

过去三月哪些机构更频繁地出现在新闻中？可以通过此圆环饼图查看这些机构的情况，如图4-8所示：

图4-8高频机构

对应的高频机构数据视图表如下：

表4-4高频机构

{'words': '农产品', 'nums': 72.25}
{'words': '教育部', 'nums': 66.46}
{'words': '中新社', 'nums': 62.45}
{'words': '公安部', 'nums': 59.92}
{'words': '老百姓', 'nums': 52.63}
{'words': '机器人', 'nums': 37.1}
{'words': '中国共产党', 'nums': 30.94}
{'words': '清华大学', 'nums': 28.77}
{'words': '中国科学院', 'nums': 22.89}
{'words': '北京大学', 'nums': 21.04}
{'words': '淘宝', 'nums': 20.08}
{'words': '广发证券', 'nums': 18.47}
{'words': '党中央', 'nums': 17.01}

通过上面的圆饼图可以清楚地看出，在过去3月里，关键词农产品所占比例最高，出现频率最高。我国作为农业大国，粮食生产一直是政府工作关注的重点，随着乡村振兴战略全面实施，农业信息化市场化的水平将得到显著的提升。教育部紧随其后的可能原因是研究生院校的复审阶段，同时这也是今年发布教育工作点的时候。除清华大学，中国科学院和北京大学外，其他大多数部门都是政府部门,大多数政府部门都能反映政府在我国的工作以及人们生活的重视和重要性。同时，召开相关会议也是这些热点词出现的重要原因。

5 总结

本文主要进行新闻热点问题分析，实现了新闻热点的发现和可视化分析。主要实现了如下任务：第一、利用多线程技术和爬虫算法实现了对中国新闻网的新闻的并行爬取。第二、利用中科院的ICTCLAS分词、信息熵算法分别对抓取的2019年3月1日到31日期间的社会新闻进行数据处理、词频词性统计以及JSON化处理。第三、利用PyEcharts技术对数据结果进行分类统计并绘制可视化视图。分析结果说明，2019年3月份社会民生是民众的关注热点，在社会民生问题上，人们对中国在社会保障、教育、食品、舆论、机器人、互联网等方面的关切程度更加迫切，说明老百姓对国家更快更好发展的热切期盼，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。

新闻热点可视化分析是一种基于大量网络新闻数据的图表分析方式，本文以中国新闻网的新闻数据为来源，计算统计了某一领域的热点信息情况，借助PyEcharts强大的数据呈现技术，让普通的数据信息焕发出崭新的魅力。在研究方法上，发现还有如下能够改善的地方：可以采取多方向爬取不同源网站的各种新闻数据进行统计，通过聚合分析，可以获得更准确的统计数据；本文的分词和词频词性分析借助的是中国科学院的ICTCLAS分词，但是由于系统限制，需要对分词文本做切分为单个大小小于500K的文件，过程中难免会出现部分的文件中存在相同但有不同统计量的关键字，需要借助其他工具比如Excel进行相同词合并，操作实现上还比较繁琐；在数据可视化方面实现了对当前处理后静态数据的呈现，如何直接与程序的API接口连接，动态的根据传输的数据及时变化有更深层次的研究意义。

参考文献

[1] .第43次CNNIC中国互联网报告发布[J].中国广播,2019(04):48.
[2] 荣晗.基于分布式的网络爬虫系统的研究与实现[D].电子科技大学,2017.
[3] 尚楚涵.互联网舆情信息挖掘技术研究与实现[D].华南理工大学,2013.
[4] 孙健康.大数据背景下数据新闻的可视化研究[D].南昌大学,2017.
[5] 王宝龙. 面向新闻领域的文本数据获取系统的设计与实现[D].北京邮电大学,2010.

[6] 巫函.数据挖掘与可视化技术对新闻阅读体验的改善——以腾讯网在巴西世界杯期间的报道为例[J].西部学刊(新闻与传播),2016(07):53-55.
[7] 高云.大数据时代数据新闻的发展现状探析[J].视听,2019(04):184-185.
[8] 张进浩.数理统计在数据分析中的应用[J].金融经济,2017(12):130-131.
[9] 杨凯利,山美娟.基于Python的数据可视化[J].现代信息科技,2019,3(05):30-31+34.
[10] Thelwall, M. A web crawler design for data mining[J]. Journal of Information Science, 2001, 27(5):319-325.
[11] Fatemeh Ahmadi-Abkenari,Ali Selamat. An architecture for a focused trend parallel Web crawler with the application of clickstream analysis[J]. Information Sciences,2011,184(1).
[12] Lishan Deng. Innovative Application of Python in Data Crawling —Chinese Version of Movie Recommendation Platform[J]. Journal of Physics: Conference Series,2019,1168(3).
[13] 甘凌博,员婕.数据新闻的交互可视化策略探析——以2018年世界杯报道为例[J].新闻传播,2019(03):37-38+41.
[14] 邓若蕾.大数据时代数据新闻的可视化传播效果探究[J].传播力研究,2019,3(04):32.
[15] 娄立原,周选州.可视化数据新闻实践探析[J].青年记者,2019(02):33-34.
[16] 林文涛,陈伟强,刘杭燕,叶楠.面向热点新闻的爬虫系统设计与实现[J].数字通信世界,2019(01):261-262+132.
[17] Ali Mesbah,Arie van Deursen,Stefan Lenselink. Crawling Ajax-Based Web Applications through Dynamic Analysis of User Interface State Changes[J]. ACM Transactions on the Web (TWEB),2012,6(1).
[18] Shaojun Zhong,Zhijuan Deng. A Web Crawler System Design Based on Distributed Technology[J]. Journal of Networks,2011,6(12).
[19] 鲁义轩.整合与分析:发挥大数据价值的两大关键[J].通信世界,2013(20):48.

指　标
疑似剽窃文字表述

1. 结构共分有五个部分，具体安排如下：
第一部分，引言。介绍了本课题的研究背景和意义，

说明：1.总文字复制比：被检测论文总重合字数在总字数中所占的比例

2.去除引用文献复制比：去除系统识别为引用的文献后，计算出来的重合字数在总字数中所占的比例

3.去除本人已发表文献复制比：去除作者本人已发表文献后，计算出来的重合字数在总字数中所占的比例

4.单篇最大文字复制比：被检测文献与所有相似文献比对后，重合字数占总字数的比例最大的那一篇文献的文字复制比

5.指标是由系统根据《学术论文不端行为的界定标准》自动生成的

6.红色文字表示文字复制部分;绿色文字表示引用部分;棕灰色文字表示作者本人已发表文献部分

7.本报告单仅对您所选择比对资源范围内检测结果负责