**《Python语言》**

**课程实验报告**

|  |  |
| --- | --- |
| 实验名称： | 疫情期间网民情绪分析 |

|  |  |  |  |
| --- | --- | --- | --- |
| 姓名： | 张芝林 | 学号： | 20211704494 |
| 专业： | 金融工程2101 | 所属学院： | 金融学院 |

**目录**

[**一、实验目的和内容** 4](#_Toc114304649)

[**（一）实验目的** 4](#_Toc114304650)

[**（二）实验内容** 4](#_Toc114304651)

[**二、算法** 4](#_Toc114304652)

[**（一）算法模型介绍** 5](#_Toc114304653)

[**（二）程序实现介绍** 5](#_Toc114304654)

[**三、输入数据集分析** 6](#_Toc114304655)

[**四、数据预处理** 7](#_Toc114304656)

[**五、源码流程分析** 7](#_Toc114304657)

[**六、实验结果** 9](#_Toc114304658)

[**七、总结与思考** 9](#_Toc114304659)

**一、实验目的和内容**

**（一）实验目的**

自2019年12月底新冠肺炎疫情爆发以来，以微博为代表的社交媒体在传播疫情信息、宣传防控措施等方面发挥了重要作用，并成为央视新闻、人民日报等官方媒体发布疫情信息的重要媒介。随着疫情的不断发展，相关微博新闻和评论的数量持续增长，众多网民对于疫情的情绪反应强烈，导致正常的生活受到明显影响。因此，以新冠肺炎疫情相关的微博新闻及其评论作为研究对象，在建立面向疫情的疫情事件以及面向网民的情绪的基础上，深入分析疫情期间出现的重点事件对网民情绪的影响，为各级政府准确掌握网络舆论情况，科学高效地做好防控宣传和舆情引导工作提供延绵之力的支撑，针对疫情相关话题开展网民情绪识别的任务，为探索微博这一媒体平台面向疫情事件的网络情绪呈现现状及进行相应的舆情治理提供参考。

**（二）实验内容**

新冠疫情这一波及全球的突发公共事件引起了极大关注和广泛的情绪反应。并对人们生活生产的方方面面产生了重要影响，并引发国内舆论的广泛关注，众多网民参与疫情相关话题的讨论。

爬取网络微博ID和微博内容，数据集依据与“新冠肺炎”相关的230个主题关键词进行数据采集，抓取了2020年1月1日—2020年2月20日期间共计1万条微博数据（给定）。微博作为一种重要的信息传播载体，在疫情信息发布与传播中发挥着重要作用。通过Excel（Office2019）表格办公软件清洗与筛选爬取的所，简单取样所抽样的微博内容数据，处理有效微博内容构成研究样本，设计Python语言算法对微博内容进行情绪识别，利用Python开发环境通过分析微博内容所发布相关文字性内容文本，借鉴参考与设立情感标签词库，分析内容情绪偏向，判断微博内容，以积极的、消极与中性作三个指标，做情绪指标的计算等相关的分析。

**二、算法**

**（一）算法模型介绍**

|  |  |
| --- | --- |
| 输入 | 微博爬取数据，数据以csv格式进行储存。 |
| 处理 | 通过Python开发环境（作者所用Spyder，安装Anaconda自带的高级IDE），利用第三方库snowNLP针对爬取微博中文的自然语言处理库，利用其主要功能情感分析用来分析每个序号的文本内容的情绪 |
| 输出 | 通过Python开发环境（作者所用Spyder，安装Anaconda自带的高级IDE），利用第三方库matplotlib.pyplot实现数据处理的可视化，导出柱状性图表，得出更加直观的实验结论。 |

**（二）程序实现介绍**

|  |  |  |
| --- | --- | --- |
| 序号 | 材料 | 备注 |
| 1 | 微博爬取数据 | 数据集依据与“新冠肺炎”相关的230个主题关键词进行数据采集，抓取了2020年1月1日—2020年2月20日期间共计1万条微博数据（给定），数据以csv格式进行储存。 |
| 2 | Python开发环境（作者所用Spyder，安装Anaconda自带的高级IDE） | 可以利用上的第三方库：  (1)jieba第三方库;jieba模块是一个python第三方中文分词模块,可以用于将语句中的中文词语分离出来。  (2)numpy库:numpy是Python用来进行矩阵运算、高维数组运算的数学计算库。  (3)matplotlib.pyplot库：matplotlib是 Python下著名的绘图库，matplotlib通过pyplot模块所提供的函数就可快速绘图以及设置图表的各种细节。  (4) snowNLP是一个Python写的类库，而且是针对中文的自然语言处理库。其主要功能便有情感分析。 |
| 3 | Excel（Office2019）表格办公软件 | 用于爬取数据的初步清洗与筛选。 |

**三、输入数据集分析**

数据集依据与“新冠肺炎”相关的230个主题关键词进行数据采集，抓取了2020年1月1日—2020年2月20日期间共计1万条微博数据（给定），其中格式如下: 微博id，格式为整型。微博发布时间，格式为xx月xx日 xx:xx。发布人账号，格式为字符串。微博文本内容，格式为字符串。微博图片，格式为url超链接，[]代表不含图片。微博视频，格式为url超链接，[]代表不含视频。

通过Excel（Office2019）表格办公软件进行用于爬取数据的初步清洗与筛选，得出格式为：序号；微博文本内容，格式为字符串。（样例如下）

|  |  |
| --- | --- |
| 序号 | 微博文本内容 |
| 1 | 患难见真情??????????????????????????????????[爱心]??“不管您是泰国人还是中国人抑或是其他任何国籍民众，如果在泰国确诊新冠肺炎2019-nCoV，那么费用全由泰国政府承担??”#PrayForChina#我们在一起，中国加油[拥抱][拥抱][拥抱] |
| 2 | 阴霾下的光亮……湖北人在石家庄#武汉加油#2020年的春节，是灰色的，每天醒来，多希望一切是场噩梦！这一年，武汉，成了举国的痛；这一年，湖北人在很多地方，几乎成了过街老鼠，人们避之唯恐不及，一切，只因为型冠状病毒性肺炎（2019-nCOV）！这是不能抹去的人类与动物不能和谐共处的浓重的一笔。 |

**四、数据预处理**

通过Python开发环境（作者所用Spyder，安装Anaconda自带的高级IDE）中第三方库snowNLP，一个Python写的类库，可以针对爬取微博中文的自然语言处理库，利用其主要功能情感分析用来分析每个序号的文本内容的情绪。snowNLP文本分类功能将原始评论分为正负情感两个数据集，返回值为情绪的概率。情感评分在[0,1]之间，越接近1，情感表现越积极，越接近0，情感表现越消极。（举例说明如下）

|  |  |
| --- | --- |
| 示例代码 | 1. **from** snownlp **import** SnowNLP 3. a = SnowNLP('好好玩啊') 4. **print**(a.sentiments) |
| 执行效果 | 0.8654879391484723 |

考虑到微博媒体平台用户所发布文本内容的有效性程度，对于判断文本内容的情绪价值是积极的、消极的还是中性的，就需要设定一个合理的情绪指标区间供划分判断进行情绪识别。

|  |  |
| --- | --- |
| 内容(comments) | 情感评分(sentiment) |
| 好 | 0.6558628208940429 |
| 不错 | 0.8612132352941176 |
| 一般 | 0.526232781807808 |
| 坏 | 0.33423913043478237 |

故此以0.5为分界点，区间设定为1.0，大于0.6为积极，小于0.4为消极，其中[0.4,0.6]的区间判为中性，进行两面的评论占比计算。

**五、源码流程分析**

介绍代码实现的详细过程情况。

|  |
| --- |
| 源码流程 |
| 1. **from** snownlp **import** SnowNLP 2. **import** jieba 3. **import** matplotlib.pyplot as plt 5. #用GB18030 因为py在打开csv的时候utf-8可能会报错： 6. file=open(r"path","rt",encoding='gb18030',errors='ignore') 7. #打开： 8. text=file.readlines()  11. #设立数据存放的字典： 12. counts={} 13. #逐行分析： 14. **for** line **in** text: 15. words=jieba.lcut(line) 16. index = SnowNLP(line) 17. values = index.sentiments 18. **if** 1 > values > 0.6: 19. values = 1 20. counts[1]=counts.get(1,0)+1 21. **if** 0 < values < 0.4: 22. values = -1 23. counts[-1]=counts.get(-1,0)+1 24. **if** 0.4 <= values <= 0.6: 25. values = 0 26. counts[0]=counts.get(0,0)+1 27. **print**(words[0],"情绪分析值为",values) 29. #数据统计分析: 30. a=counts[1] 31. b=counts[0] 32. c=counts[-1] 33. total=a+b+c 34. **print**("\n分析统计:",counts,"总计数据为{}条".format(total)) 35. **print**("备注:其中1表示乐观，0表示中立，-1表示消极") 37. #可视化数据: 38. #设置坐标： 39. x=1 40. #设置长度： 41. w=2 42. #并列柱状图 43. plt.bar(x,a,w,label='positive') 44. plt.bar(x+w,b,w,label='objective') 45. plt.bar(x+w\*2,c,w,label='negative') 46. plt.legend() 47. plt.show() |

**六、实验结果**

|  |
| --- |
| 将数据可视化输出结果 |
| 图表 1分析统计: {0: 600, -1: 3063, 1: 6324} 总计数据为9987条（去除无效、格式错误数据13条）。  备注:其中1表示positive，0表示objective，-1表示negative。 |

分析结果所得，在抽样调研的1万条对象数据中，评估每一用户所发布的文本内容的情绪分析值，其中“乐观”的情绪标签有6324条数据，“中立”的情绪标签数据有600条，同时有3063条“消极”情绪标签的数据（去除无效、格式错误数据13条）。网络情绪总体倾向于积极，但是消极情绪占比也接近1/3，这部分消极情绪标签的数据尤其值得重视。对于数据的时期而言，疫情事件影响公众的时期可归于情绪的酝酿期间，众多网民的情绪是处于弥散状态的，后续存在发展成任何类型社会情绪的可能性，最后哪种情绪成为主流，还有诸多因素需要考虑。

**七、总结与思考**

本次实验课题研究以新冠肺炎疫情相关的微博新闻及其评论作为研究对象，在建立面向疫情的疫情事件以及面向网民的情绪的基础上，深入分析疫情期间出现的重点事件对网民情绪的影响，为各级政府准确掌握网络舆论情况，科学高效地做好防控宣传和舆情引导工作提供了延绵之力的支撑，针对疫情相关话题开展网民情绪识别的任务，为探索微博这一媒体平台面向疫情事件的网络情绪呈现现状及进行相应的舆情治理提供参考。

在处理已给定收集的数据的过程中，我能直观地感受到数据量之多，处理之复杂，深刻认识到像Python这样的高级语言的价值所在，能够利用计算机处理成千上万条复杂的文本数据，利用第三方库jieba、snowNLP库等更能发挥Python语言在处理数据方面上的举足轻重的效用，尤其是在分析中文文本这一方面上，Python的优势更令人折服，直至最后数据处理输出，利用matplotlib第三方库简洁而直观地展示数据处理的结果，绘制柱状图的数据展示，进一步提供了更具参考价值的数据报告结果，众多第三方库和开放包容的Python语言环境相辅相成，为诸多从业数值分析的工作人员提供了一个稳定而高效的平台。

在本次课题研究的过程中，学期课程教授给我的各种知识如库的引用、文件的读取使用等知识点能够很好地运用实践上，这对于作为学生的我来说，这是一个十分合乎预期且很有价值的一次课题研究，能够将在平日所吸收学习的知识点运用起来，转化为自己技能与能力的一部分，我想这才是真真正正学有所成的一部分，而不是一味地学习理论知识和一些皮毛，做应试教育。通过这次课题研究，也有拓宽视野的一面，本着研究过程，让我在学习Python领域拓宽了更广的视野，在CSDN社区论坛上与各式各样的程序员、学生、爱好者交流讨论，学习研究更多的第三方库，使用更加实用高效的运行环境与平台，这都是十分具有价值的收获所在。

新冠疫情这一波及全球的突发公共事件引起了极大关注和广泛的情绪反应。并对人们生活生产的方方面面产生了重要影响，并引发国内舆论的广泛关注，众多网民参与疫情相关话题的讨论。在研究所给定的数据处理输出后，实验结果表明，在抽样调研的1万条对象数据中，评估每一用户所发布的文本内容的情绪分析值，其中“乐观”的情绪标签有6324条数据，“中立”的情绪标签数据有600条，同时有3063条“消极”情绪标签的数据（去除无效、格式错误数据13条）。网络情绪总体倾向于积极，但是消极情绪占比也接近1/3，这部分消极情绪标签的数据尤其值得重视。对于数据的时期而言，疫情事件影响公众的时期可归于情绪的酝酿期间，众多网民的情绪是处于弥散状态的，后续存在发展成任何类型社会情绪的可能性，最后哪种情绪成为主流，还有诸多因素需要考虑。重大突发公共卫生事件本身即为严重的负向刺激，事件爆发后，若政府应对处置不当，媒体信息缺失或信息封闭，虚假或恐怖信息蔓延，则在负向事件刺激、政府处理效力低下、媒体功能缺位的多重作用下，人们很容易产生焦虑、紧张等心理，形成群体舆论性心理恐慌。随着公共事件危机的大规模蔓延，公众情绪波动不断加剧，往往会在消极情绪的累积中出现情绪失衡。特定的事件引起网民关注并集体表达个体情绪，会进而形成较为稳定的集体情绪，同时网络情绪会反作用于事件本身，使得原有特定事件的发生发展改变原有的规律。直至现在，疫情尚未结束，新的疫情不断演变来袭，在政策朝令夕改的同时，我们更要重视公共事件之下群众的情绪疏导与舆情治理，关注、重视群众情绪，调解与舒缓公众的恐慌和消极心理，避免消极情绪发展为主流，有利于稳定社会且持续、高效地推进疫情防控工作的开展。