**南京大学数据科学大作业**

**题目：新冠疫情新闻分析**

**组长：徐赫**

**学号： 191250164**

**所在院系：软件学院**

**专业：软件工程**

**手机号码：15061718022**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 序号 | 成员姓名 | 学号 | 所在院系 | 专业 | 手机号码 |
| 1 | 张悦祺 | 191250199 | 软件学院 | 软件工程 | 18851863771 |
| 2 | 单金明 | 191250122 | 软件学院 | 软件工程 | 19831700078 |

**分工：**

**新闻爬取：张悦祺，单金明**

**情感分析：徐赫，张悦祺**

**总结报告：徐赫，张悦祺**

代码开源地址：https://github.com/darcy823/DataScienceTermProject-reNew-

**2021年1月24日**

一．评论部分

代码分析：

数据源来自微博手机端（https://weibo.cn/search）

代码思路：

1. 通过微博搜多界面获取每个时间段“疫情话题下”微博总数和热门微博ID

2. 通过微博ID爬取该微博下热门评论和评论总数

3. 将数据写入txt文件

数据格式：

'评论者主页', '评论内容', '评论获赞数', '评论发布时间',按时间段分别打包成txt数据

调用包/方法：import os

from datetime import datetime, timedelta

import requests

from lxml import etree

import re

from time import sleep

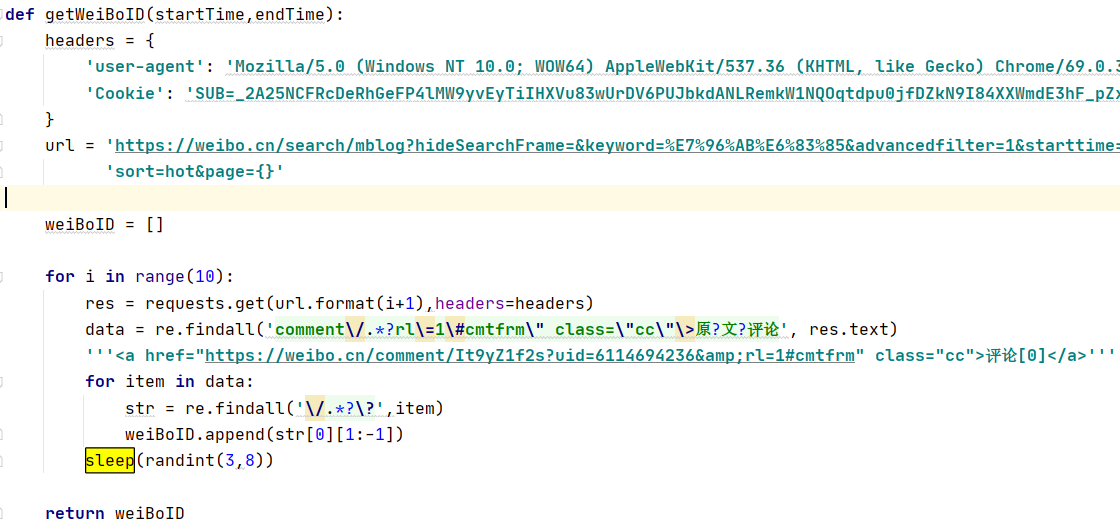
from random import randint

具体实现方法（代码文档）：

CommentsGrab文件下包含7个方法，分别是主函数main（），运行函数run（），热门微博ID抓取函数getWeiBoID(),微博评论获取函数WeiboCommentScrapy()，评论信息解析函数 get\_one\_comment\_struct()，时间解析函数parse\_time()，以及写函数write(result,starttime)。

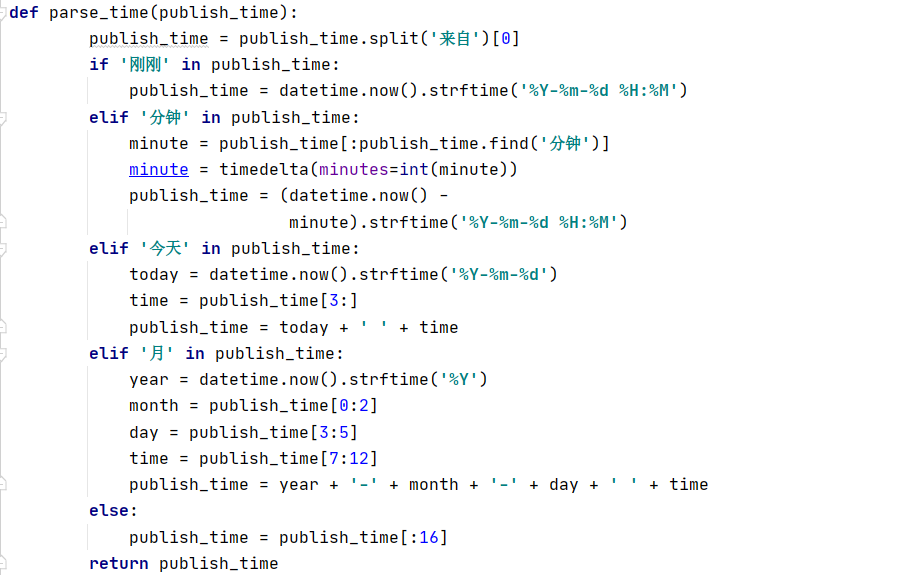
其中，main（）函数中存储时间分段，并调用run（）函数启动程序。run（）函数依次对每个时间段进行热门评论抓取，获得热评，内容解析，写入数据。

getWeiBoID()通过cookies进行浏览器模拟，并调用requests获取搜索源代码，结合正则表达式提炼微博关键信息，形成列表并返回给run（）。



getWeiBoID()

run（）函数将微博ID传入WeiboCommentScrapy()通过lxml库进行解析得到python对象并获得URL。通过get\_one\_comment\_struct()和parse\_time()将数据格式转化成格式化数据用于下一步的处理。

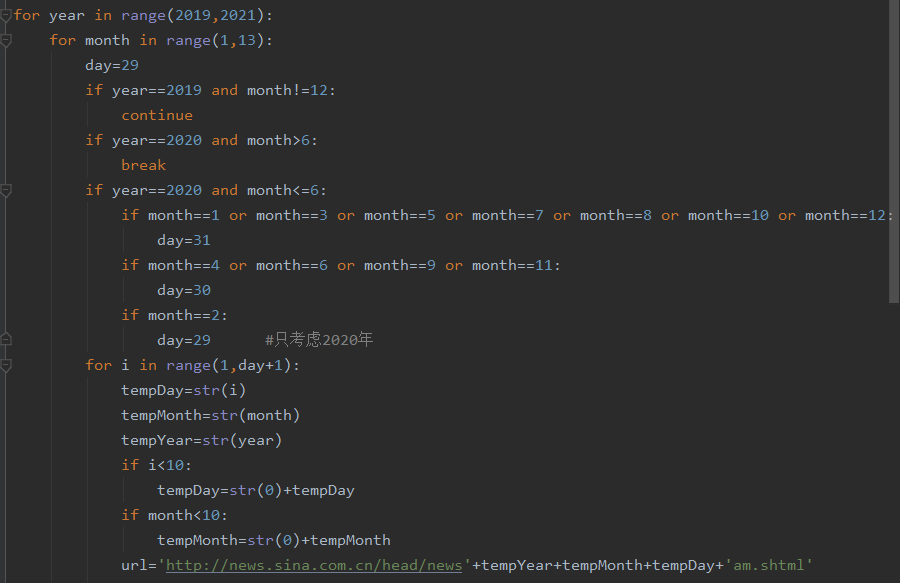


parse\_time()

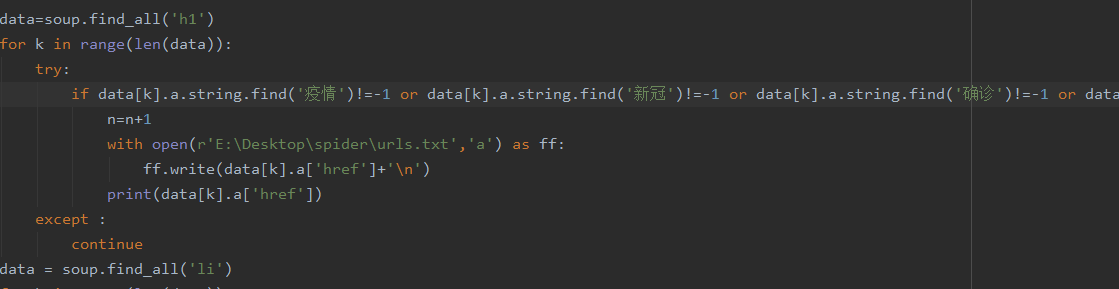
中间穿插sleep（）函数防止操作频繁导致IP禁封。

二．新闻部分

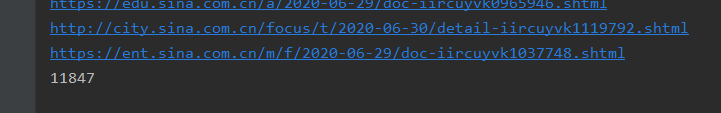
我们组选择爬取新浪新闻2019年12月到2020年6月这七个月关于此次新冠疫情的新闻。根据规律获得每一天新浪新闻的首页的链接。（默认新浪新闻首页的新闻算重点新闻）（对应代码文件为spider.py）



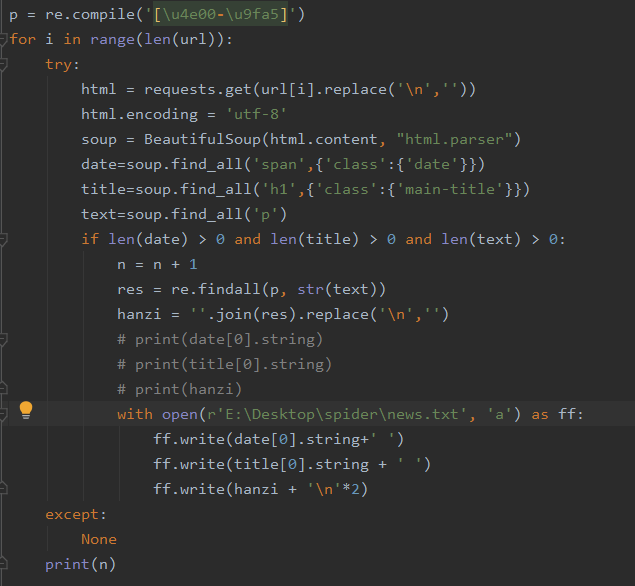
首页中有很多新闻的链接，然后根据标签中的关键词选择出跟疫情有关的新闻，关键词选择“疫情”，“新冠”，“确诊”，“隔离”，“武汉”等。



共获得11847条新闻链接



最后根据每一个链接爬取下对应的新闻,通过爬取属性值为date,main-title的日期，标题以及文章。（对应代码文件为singleSpider.py）



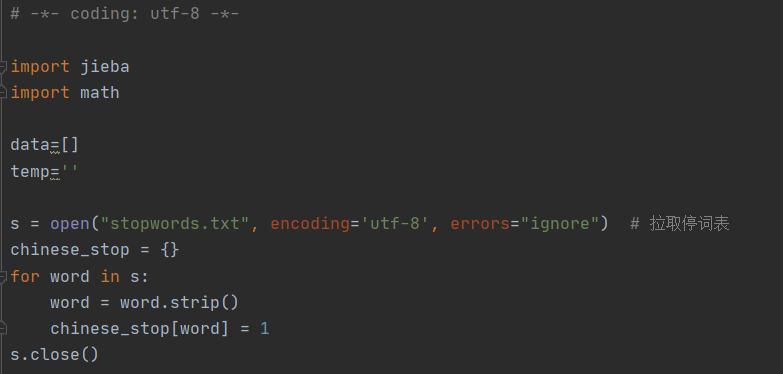
下一步操作就是对爬取的数据进行自然语言处理。这一过程中涉及到关键词的提取，关键词的词频统计，对每条新闻进行情感分析。

由于汉语和英语的语法不同，连字成词且词与词之间没有空格分隔，这导致我们对汉语进行自然语言处理之前需要进行分词操作。所谓分词，就是讲一句话，一段话，乃至一篇文章分割成为若干独立的词汇。而在本次的大作业中，我们选择使用Hanlp来实现分词操作以及后续的自然语言处理。

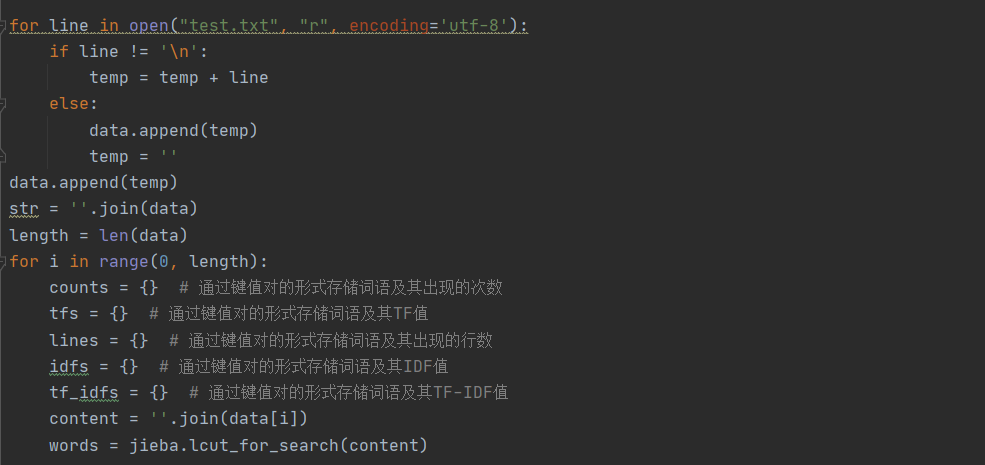
Hanlp是由哈工大开发的开源的自然语言处理库，不同于一些简陋的分词类库，HanLP精心优化了内部数据结构和IO接口，做到了毫秒级的冷启动、千万字符每秒的处理速度，而内存最低仅需120MB。不同于市面上的商业工具，HanLP提供训练模块，可以在用户的语料上训练模型并替换默认模型，以适应不同的领域。项目主页上提供了详细的文档，以及在一些开源语料上训练的模型。我们使用的pyhanlp是用python包装了HanLp的java接口。

在本次大作业中，我们主要使用了hanlp的三个功能：词频统计，关键词提取以及情感分析。接下来是具体的代码实现。除hanlp以外，我们还尝试使用jieba分词。

在对数据进行词频统计和关键词提取的过程中，我们尝试了两种做法互为验证。首先是手动提取关键词。这部分的实现有如下几个步骤：首先，我们拉取停用词表。这里我们使用的是hanlp内置的停用词表。停用词表的作用在于计算词频（TF）和逆文档频率（IDF）时将一些无意义的标点符号（如“，”、“。”、“？”）以及无实际意义的字（如“和”、“或”、“与”）去除。



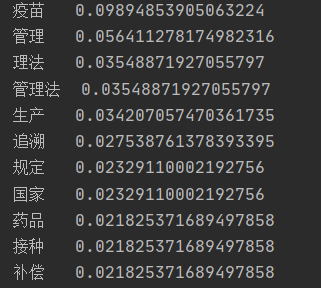
其次，使用分词工具对新闻进行分词处理（可以使用hanlp分词或jieba分词，因为后续第二种方法主要依赖hanlp，故手动提取关键词的过程我们选择使用jieba分词）因为我们所



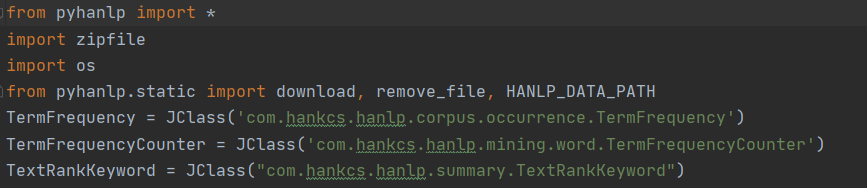
爬取的数据是将每条新闻以空行间隔开。因此我们进行分词以及关键词提取操作时借助for循环对每条新闻进行操作，以免每条新闻的总字数不同影响关键词提取的精确度。最后，我们将停用词表中的词剔除，对TF,IDF,以及TF-IDF进行计算，并将TF-IDF按值大小由高到低排列，值越高代表其越关键。

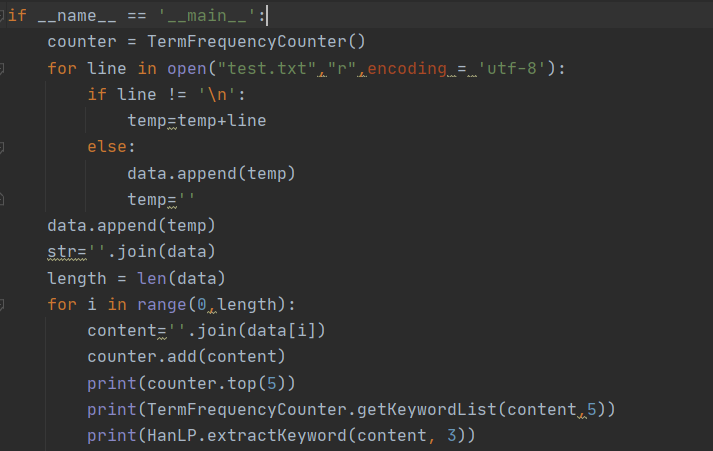


可以得到如下的关键词。

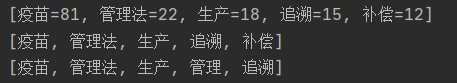


并且作为验证，我们还使用了hanlp中封装的TermFrequencyCounter类用来统计文档的词频，以及TextRankKeyword类来进行关键词提取。

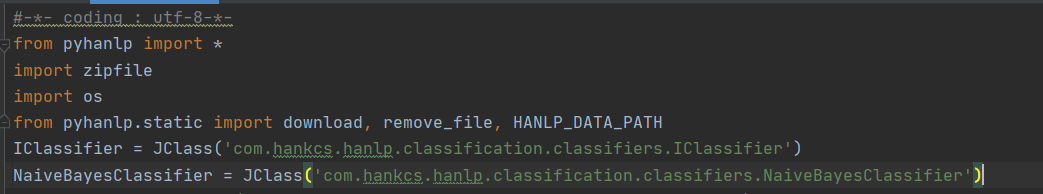


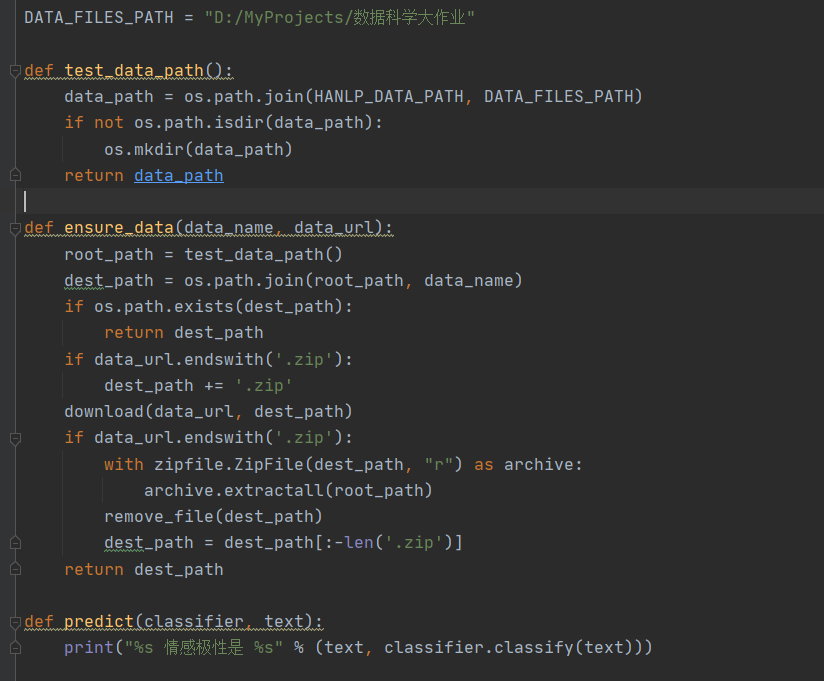


得出的关键词与手动计算的关键词相符，验证了关键词提取的正确性。

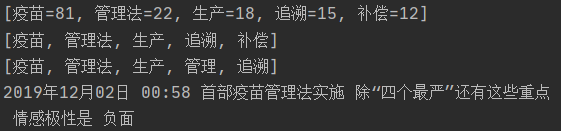


对于每条新闻的情感分析，我们发现了hanlp封装的IClassifier和NaiveBayesClassifier类可以作为十分有效的工具。Hanlp的情感分析需要语料库作为依据，只要我们拥有同样格式的语料，那么我们可以使用这个分类器做任何我们需要的文本分类。并且我们可以利用文本分类在情感极性语料上训练的模型做浅层情感分析。目前公开的情感分析语料库有：中文情感挖掘语料-ChnSentiCorp，语料发布者为谭松波。代码实现如下：首先调用hanlp库，其次设置文件路径并配置语料库，最后使用文本分类进行情感分析得出情感极性。







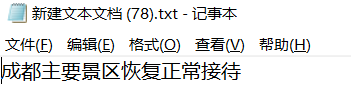
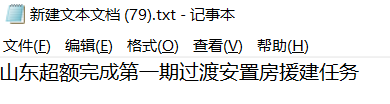


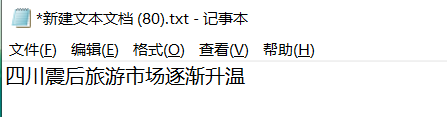
此时我们发现，我们所调用的语料库，只有正面和负面两个分类，于是我们尝试在此基础上建立自己的机器学习样本。

样本大体分为两类，积极与消极。进一步细分为乐观，坚定，感恩，自豪，忧虑，不满，害怕，愤怒。

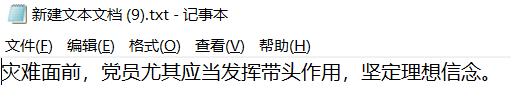
首先是积极部分。（只展示部分示例）

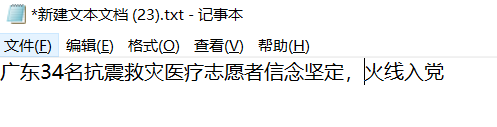
乐观的样本包括

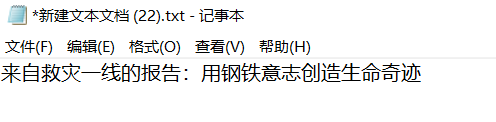
 



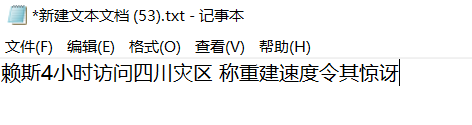
坚定的样本包括

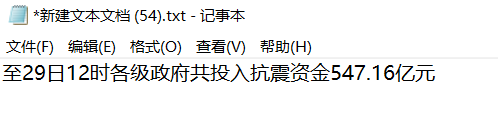


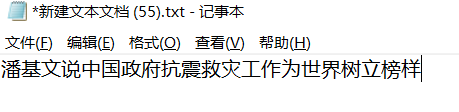




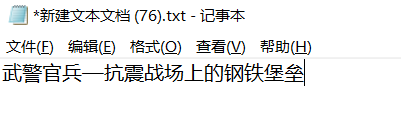
自豪的样本包括

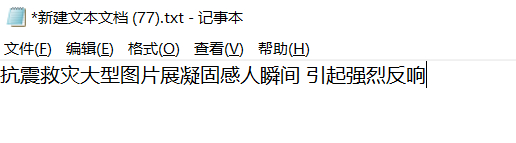


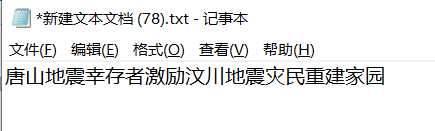


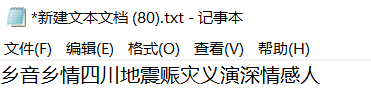


感恩的样本包括



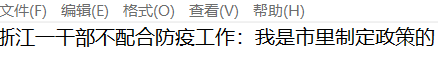




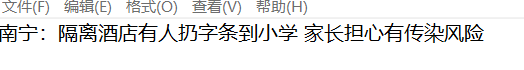
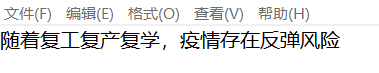


其次是消极部分，根据消极程度将之分为忧虑，不满，害怕和愤怒

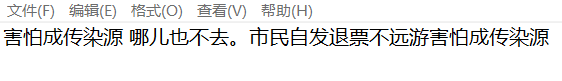
不满：



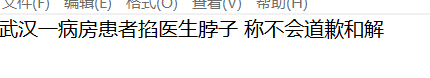
忧虑：



害怕：

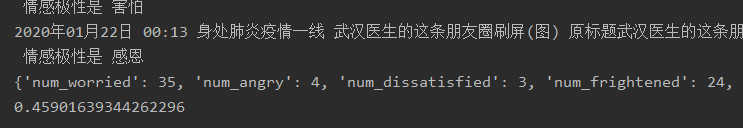


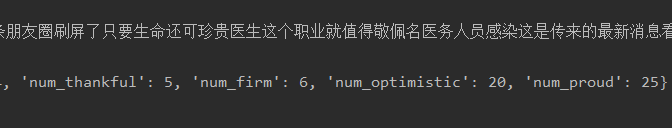
愤怒：



下面进行新闻的情绪分析，分为四个阶段。

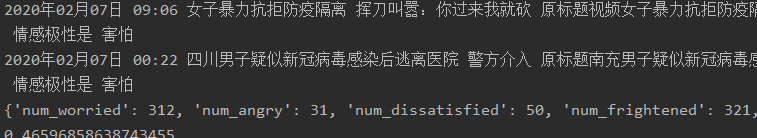
第一阶段2019.12.01-2020.1.22

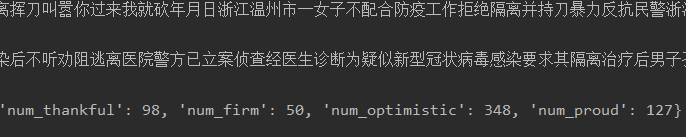




第一阶段为不重视和无奈扩散阶段。该阶段的标志事件是2019年12月8日首例肺炎患者确诊以及2020年1月22日湖北启动公共卫生2级响应。。群众此时出现的情绪以负面的忧虑和恐惧为主，由图中数据可知恐惧占比20%忧虑占比29%。由于2020年1月20日钟南山院士明确提出新冠病毒肯定可以“人传人”。因此在2019.12.8到2020.1.20的时间区间内，人们对于新冠病毒只之甚少，出现了真实信息与谣言混杂的情况（例如“吹哨人”李文亮医生由于发布疫情真实消息在2020年1月3日，因“在互联网发布不实言论”，而被辖区派出所提出警示和训诫）。人们对疫情的确凿信息知之甚少，因而倍感忧虑。除此之外由于武汉政府初期反应不及时，政策跟进速度慢且存有侥幸心理，并未进行有效的人群疏散工作，导致在1月9日出现首例死亡病例之后仍未对进出武汉的流动人群以及武汉市民大规模聚集进行控制导致疫情传播加速，以及后期钟南山院士确认人传人现象存在，增添了群众的恐慌情绪。

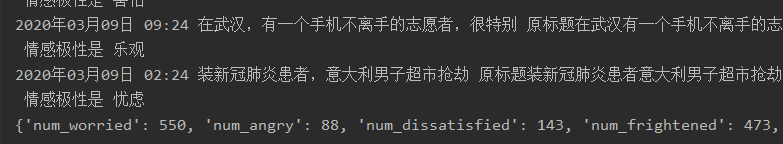
第二阶段 2020.1.23-2020.2.7

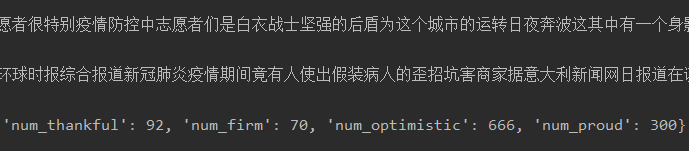




第二阶段是资源缺乏阶段。该阶段的标志是为2020.1.23武汉宣布“封城”和2020.2.7“吹哨人”李文亮去世。这一阶段仍然是负面情绪大于正面情绪。并且恐惧情绪占比上升达到了24%，这一阶段出现了医护人员感染牺牲（1月25日梁武东医生和2月7日李文亮医生）并且病毒开始从武汉蔓延，包括西藏这样的偏远地区（1月29日）以及菲律宾（2月2日，境外首例）这样的境外地区都出现了病例，而2月5日出生仅30小时的婴儿确认感染，更加剧了人们的恐惧情绪。但是值得注意的是，这一阶段正面情绪中的乐观比例达到了26%，这是源于1月25日中央成立了疫情领导小组以及雷神山火神山医院的建设和交付，人们对于党和中央政府的信任以及对基础医疗设施建设的中国速度的信赖是乐观情绪的主要来源。

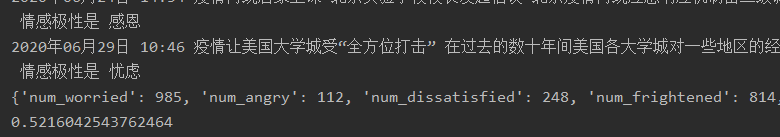
第三阶段 2020.2.8-2020.3.9

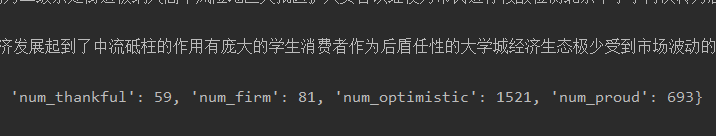




这一阶段开始标志是湖北政府相关领导的变动。这一阶段下大众情绪基本稳定，和上一阶段相比，持乐观情绪的群众占比进一步增长达到28%。这是因为这一阶段全面抗疫步入正轨，武汉保卫战进入总攻，各省市对口援助湖北的开展。政府领导力和社会主义制度优越性开始显现，抗疫物资集中生产，统筹调配，党员带头冲锋，轻伤不下火线，各种最美逆行者的出现都增长了人们的乐观情绪。这一阶段，每日新增病例逐渐下降，甚至每日新增病例仅两位数，治愈率也在稳步上升。与此同时，国外疫情开始严峻，西方各国应对疫情的丑态初现，对比之下也增长了群众的自豪情绪。

第四阶段2020.3.10-2020.6.30





这一阶段标志性事件是各省市开始有序复工复产。这一阶段正面情绪超过负面情绪，占比过半。这一阶段人们的生活逐步回归正轨，各地有序复工复产复课，在常态化防疫的保护下，中国开始逐步走出疫情影响，人们对于疫情也从无知恐惧，转变为乐观积极。乐观情绪占比超过三成达到了34%。而这一阶段的忧虑，恐惧等负面情绪的来源，主要来自对境外输入病例及无症状感染病例的恐惧与忧虑。这一阶段，以美国为首的西方国家疫情愈演愈烈，与国内逐步走出疫情影响形成更加鲜明的对比。对祖国的自豪情绪进一步上升致15%，这是对社会主义制度和道路的自信自豪，对党和国家坚强领导的自信自豪。

下面从总体来看各种情绪占比在四个阶段的变化，首先是消极情绪。

忧虑的情绪占比四个阶段稳步下降，说明民众从一开始的慌乱到逐渐接受了疫情事实，并严格遵守国家相关防疫规定。害怕情绪只在第二节点大幅增长，因为一开始民众不能完全接受疫情，更多持怀疑态度，没有意识到它的传播之快，范围之广。一波大面积感染后，人们才从忧虑转为害怕，害怕被传染，害怕已经被感染，在外打工的人害怕给家人带来麻烦而不敢回家等等。而愤怒与不满占比始终较低，说明人民还是十分信任他们的国家。另一方面说明在党的领导下，国家政策的制定与实施速度之快，效率之高属实让人安心。

接着是积极情绪

一开始民众对疫情的看法不是很充分，故更多体现在一种相信国家会平静如往常，可以说是一种自信，自豪感，坚定的情绪也最为高昂。当进入第二阶段时打破了这种平静，但是国家的政策十分及时，火神山医院的建立等举措挽回民众的信任，心态开始走向乐观，并且在这一阶段感恩情绪达到顶峰，正是因为无数的最美逆行者在为我们战斗。从第二到第三阶段开始，抗疫工作开始稳住步调，民众的自豪感直线上升，相信不久就会达到之前的程度并超越之。

纵观疫情开始后的整个半年，虽然整体新闻为消极居多，有三个阶段的消极率都超过了50%。但是，这是建立在我们搜索的关键词为“疫情”，“确诊”，“隔离”等本身就带负面情绪的基础上，其中的积极率能有近乎半数已经难能可贵，第四阶段甚至实现了反超，这正是因为中国这个国家的伟大以及中国人民的坚韧不拔，乐观精神。

疫情虽然尚未结束，河北黑龙江的疫情似乎又有抬头的趋势。但历经1998年的洪灾，2008年汶川地震，以及武汉疫情，历史已经证明并将一次又一次向我们证明，中国共产党是真正为人民服务的政党，是真正把人民生命健康放在首位的政党。党的坚强领导，统筹规划是我们克服万难的坚强保障。多难兴邦，道阻且长。我们有理由相信，在党和国家的坚强领导下，中华民族的伟大复兴终有一日必定实现。