小组信息：人数：2 人；

组员：

姓名：张晋轩　学号：191250191　邮箱：191250191@smail.nju.edu.cn

职责：新闻及评论数据爬取

姓名：康森　学号：191250065　邮箱：191250065@smail.nju.edu.cn

职责：数据处理及分析

一、 研究问题：COVID - 1 9 背景下的网络社会心态及公众情绪分析

1. 认识：随这互联网的不断发展、网上交流渠道的不断丰富与拓展，日益复杂的公众情绪与思维得以有丰富的渠道进行抒发和交流，极大地推动了社会心态的塑造，并且有着使群体心理及集体行为极化的可能。在当前疫情横行的特殊时期，人们对于新冠疫情的关注使得社会心态的变化十分显著，形成了特殊时期的特殊网络社会心态与公众情绪。因此，立足于此次新冠病毒肆虐的特殊时期，本小组借助了一定的数据与计量手段，收集并测量了网络上新冠病毒相关新闻的内容及其评论，希望通过此种方式研究社会舆论与心态随疫情发展情况的变化情况，并客观了解公众的网络心态与行为变化规律。
2. 研究出发点：首先通过爬虫获取百度新闻、微博新闻的内容及评论，再通过心态词典对评论的情绪进行分析、提取，获得不同时间段的评论情绪的数据，进而对社会心态及公众情绪进行分析。

二、 代码：

1. 爬虫开源地址：https://github.com/yuan-su-xuan/NewsCrawler
   i. Crawler: 用于爬取百度上的新闻内容的链接
   ii. WeiBoCrawler: 用于爬取微博上的新闻内容加评论
   iii. TextGetter: 用于访问爬虫爬下来的链接进而获取新闻内容
   iv. CommentsGetter: 选取时间段进行爬取
2. 数据分析代码开源地址：
   i. https://github.com/suyiiis/learn/tree/master/%E5%A4%A7%E4%BA%8C%E4%B8%8A/%E6%95%B0%E6%8D%AE%E7%A7%91%E5%AD%A6%E5%9F%

BA%E7%A1%80/%E5%A4%A7%E4%BD%9C%E4%B8%9A/data_analysis

ii. Jieba:用于分词

iii. Nltk：用于计算 TF

iv. Matplotlib：用于绘制图像

三、 研究方法：

1. 数据分析方法：

使用 jieba 分词

采用 TF 逆序寻找常见心态词

字典映射统计心态频率

各阶段心态百分比化

单一心态的变化趋势可视化

采用 Matplotlib 可视化

2. 数据集：

爬虫所爬下来的四个阶段数据

3. 详细说明：

（在爬虫爬数据时，筛选过热评，故不再对评论做高斯分布的拟合）

1. 使用 jieba 分词的全分词模式来获取尽可能多的情绪词汇，并通过停用词去除某些可能一词多分的情绪词
2. 建立心态词典时，采用 TF 逆序寻找常见心态词，并在每个阶段随机抽取千条评论，补充心态词典的网络用语词汇表
3. 通过字典映射的方式统计心态频率，计算各阶段的心态占比
4. 使用 Matplotlib，将各阶段的心态占比绘制为饼状图，可以一目了然的看到各阶段网民们的舆论倾向
5. 使用 Matplotlib，将心态的占比随时间的变化绘制为折线图，更直观地看出网民们随着疫情变化时的心态变化
6. **因为疫情的回弹，所以我们考虑到需要分析后续网民们的心态变化，所以我们还做了拓展工具，可以方便快捷地爬某个时间段的数据并进行心态分析，和心态走向折线图的绘制**

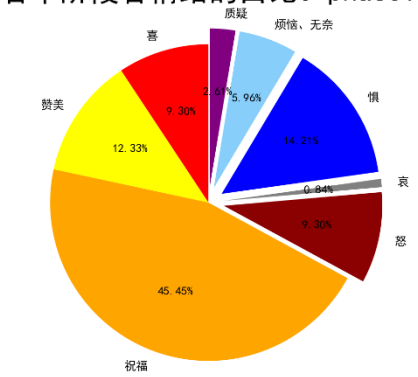
四、 案例分析：在本次研究中……

1. 从四个阶段的饼状图可以看出，不论是哪个阶段，正面的心态都占了一大部分约为 60%~70%，而负面情绪占小部分。这比较符合网络上的主流声音是积极向上，但仍有负面消极的情绪
2. 恐惧和愤怒占负面情绪部分的绝大多数
3. 通过折线图可以看到：
   i. 疫情爆发之初，人们更倾向于祝福那些疫情重灾区的人们。但随着国家的努力，疫情的好转，祝福的占比呈下降趋势
   ii. 疫情刚爆发时，人们没有太多的事情、对象去赞赏，故赞美的占比相对较低，然而第二阶段各地政府的迅速反应，比如武汉封城等，人们的赞美陡然上升，后续两个阶段也只是有所波动
   iii. 人们在面临未知的事物是总会有从心底油然而生的恐惧，而恐惧占了第一阶段负面情绪的绝大部分，但随着疫情的真相逐渐解开，各地政府和医护人员的努力，人们的恐惧逐步下降
   iv. 第一阶段时，“愤怒”占了负面情绪一部分，主要原因是当时人们不了解疫情的真相，不明白疫情的来源，而且临近春节，不能回家让人们相当的愤怒。所以当时人们的矛头指向了武汉吃蝙蝠的那些人，网民们更是对吃蝙蝠的人不吝指责
   v. 在二三四阶段中，各地封城，新春佳节的结束让人们回归冷静，故而愤怒的情绪占比逐渐下降。随之而来的则是人们对死者的哀伤，和对医护人员，为疫情奋斗的工作人员的同情这类负面情绪的上涨。而在疫情好转，这类情绪也逐步下降。
   vi. 总体来看，随着国家和政府的努力，疫情期间积极正面的情绪占比逐步上涨，而负面情绪逐步消退。

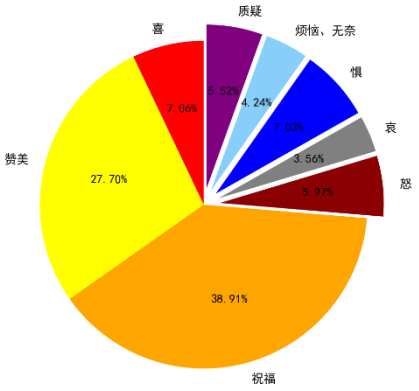
五、 课程意见：暂无

六、 附录：（补充说明的图表、数据）

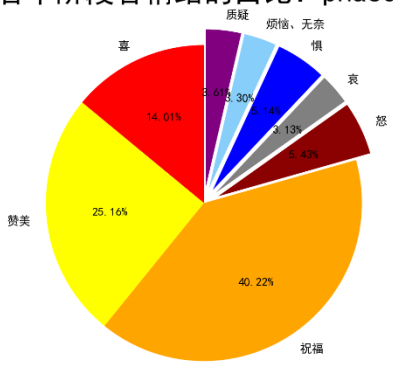
各个阶段各情绪的占比：phase1



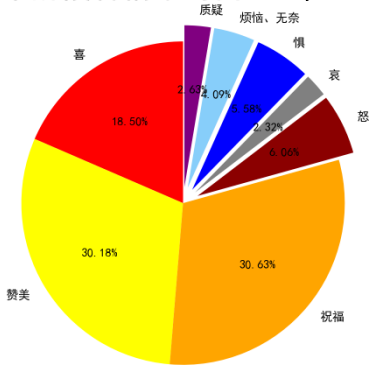

各个阶段各情绪的占比：phase2

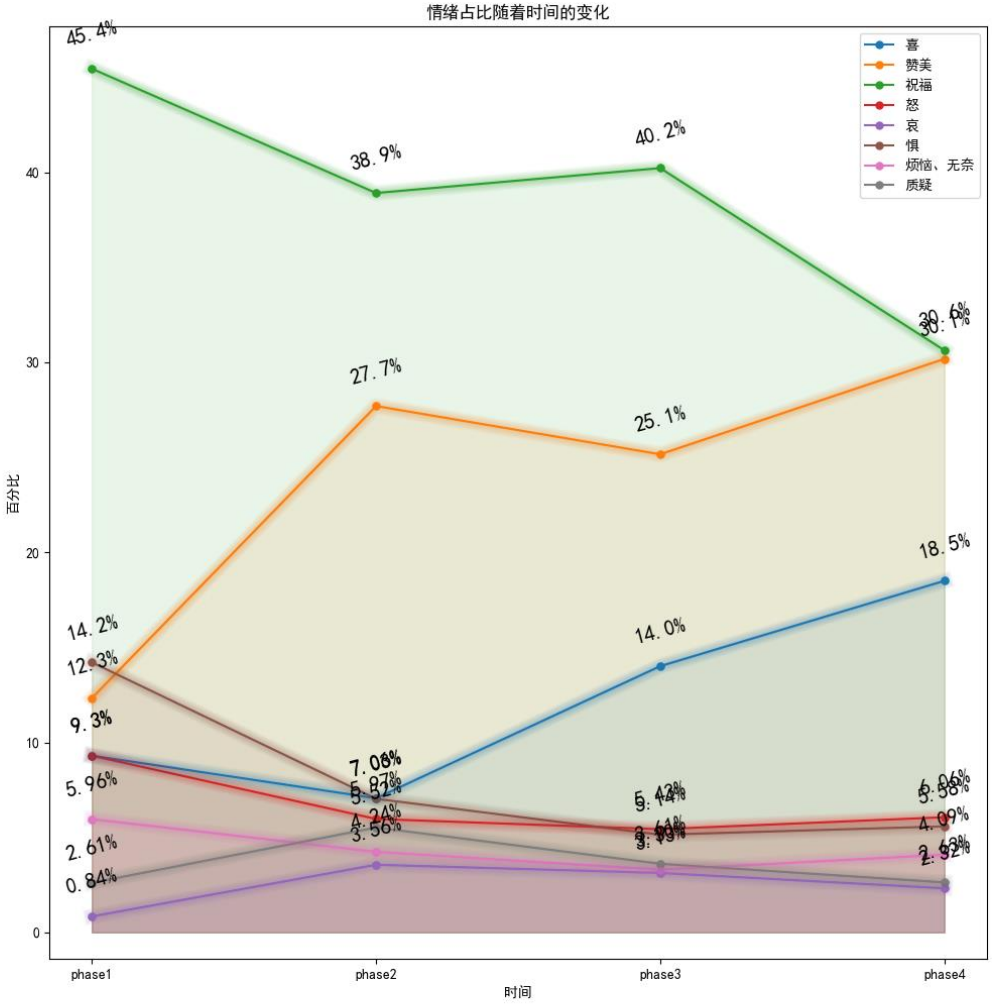


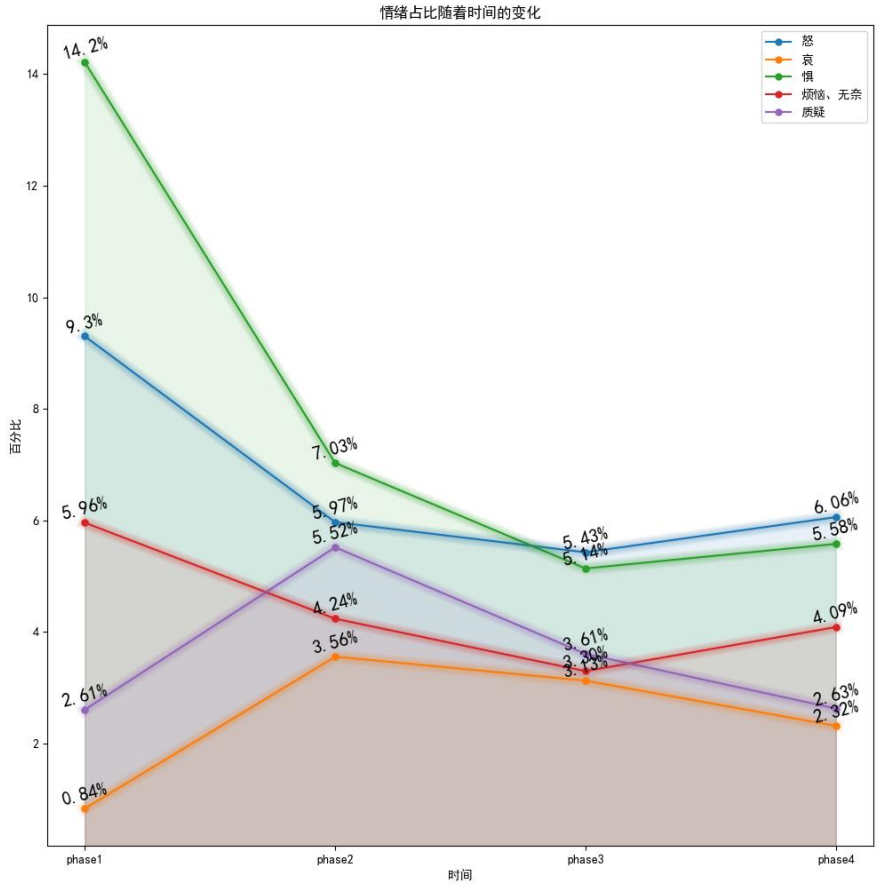

各个阶段各情绪的占比：phase3



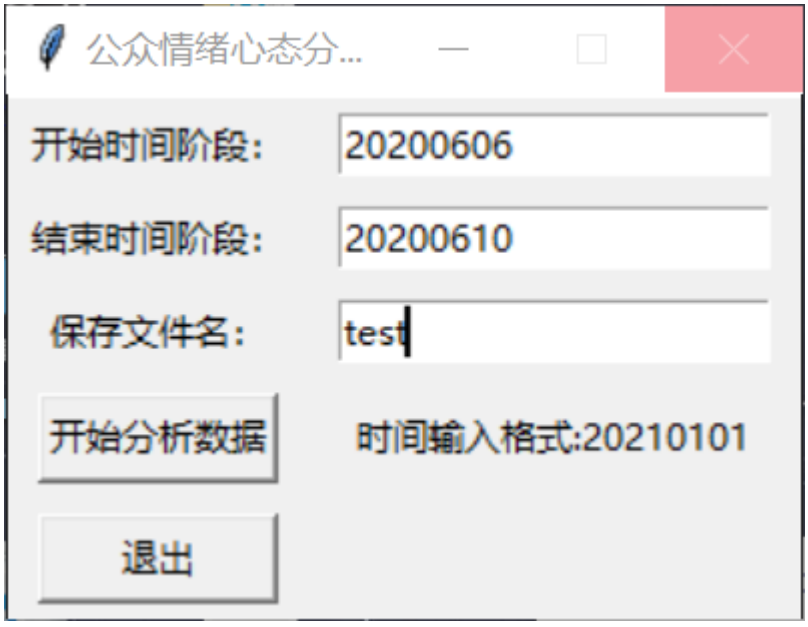

各个阶段各情绪的占比：phase4




情绪占比随着时间的变化
喜
赞美
祝福
怒
哀
惧
烦恼、无奈
质疑
百分比
时间
phase1
phase2
phase3
phase4
0
10
20
30
40
45.4%
38.9%
40.2%
27.7%
25.1%
18.5%
14.0%
14.2%
9.3%
5.96%
2.61%
0.84%
3.56%


情绪占比随着时间的变化
怒
哀
惧
烦恼、无奈
质疑
百分比
时间
phase1
phase2
phase3
phase4
14.2%
9.3%
5.96%
2.61%
0.84%
7.03%
5.97%
5.52%
4.24%
3.56%
5.43%
5.14%
3.61%
3.30%
3.13%
6.06%
5.58%
4.09%
2.63%
2.32%

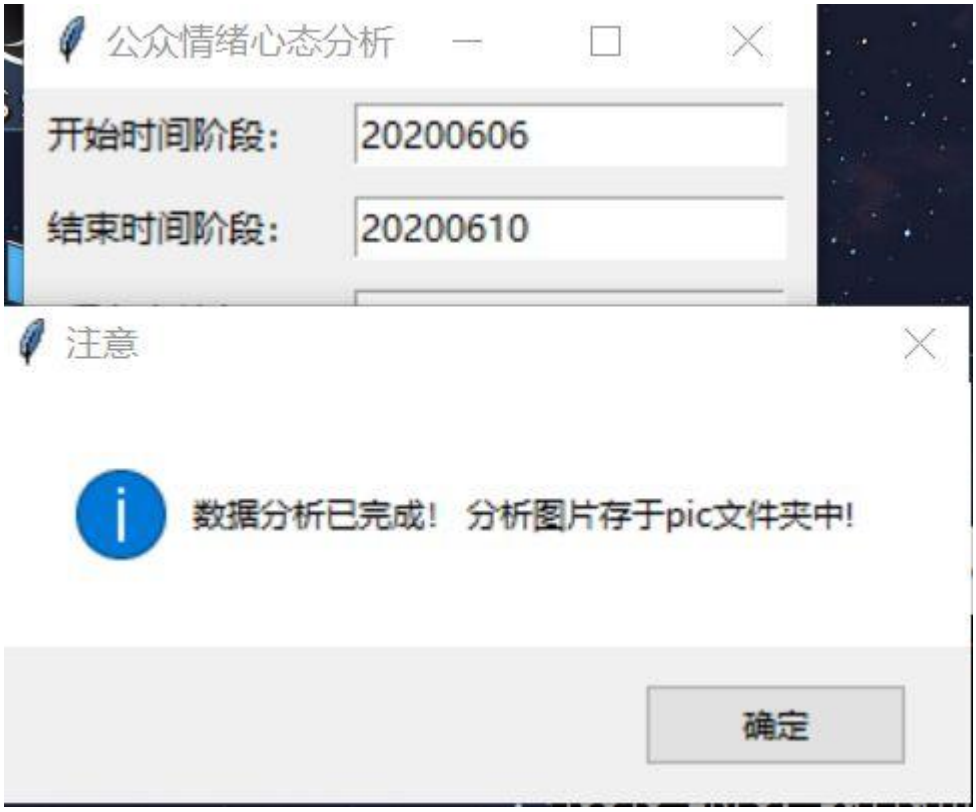
***我们设计的程序：**


公众情绪心态分...
开始时间阶段:
20200606
结束时间阶段:
20200610
保存文件名:
test
开始分析数据
时间输入格式:20210101
退出

• 这里输入起始时间和结束时间，点击开始分析就可以分析啦

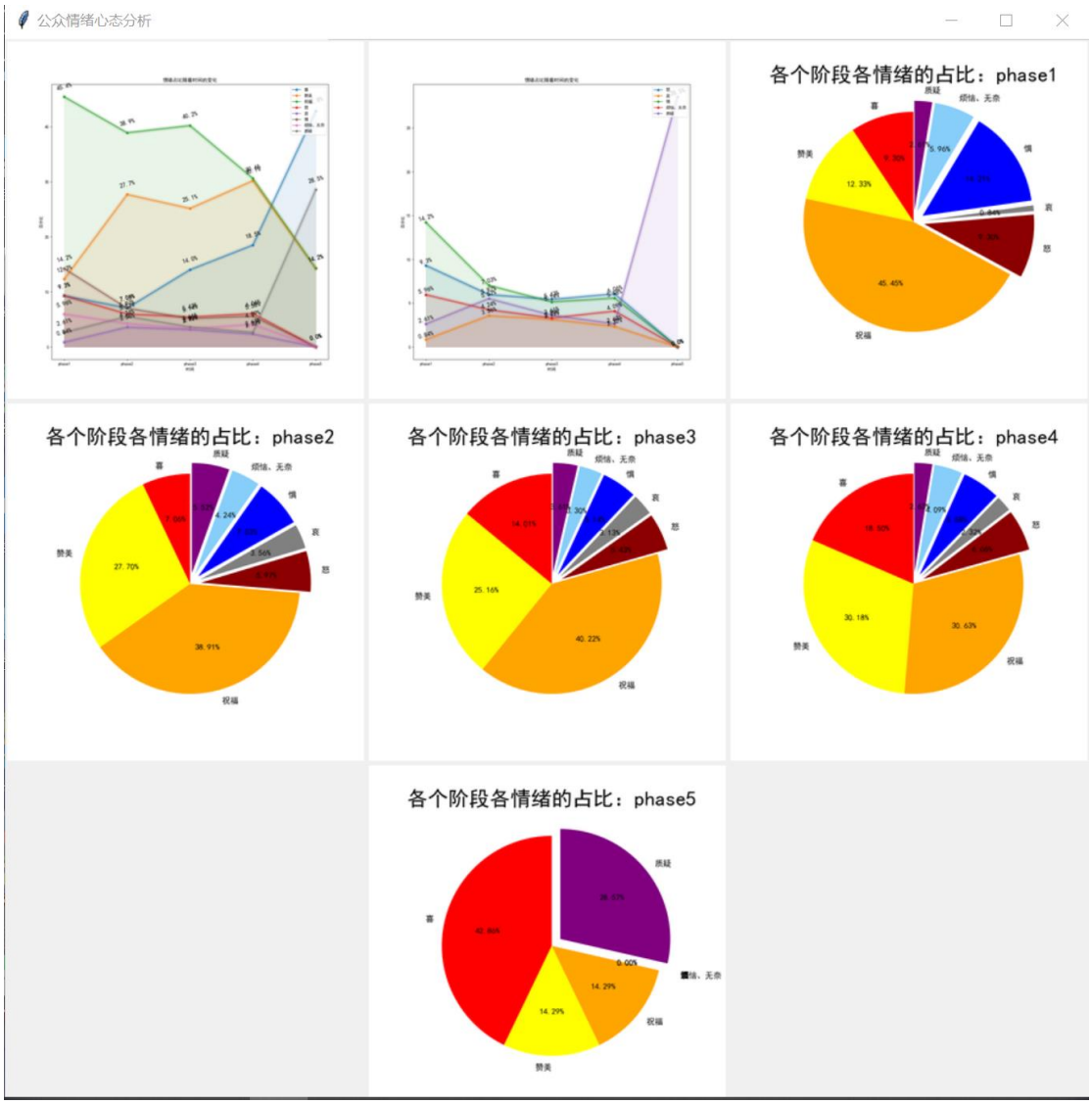
• 为了便于展示，我们减少了数据爬取量.

• 尽管如此，为了躲避微博的防爬，我们爬取 5 天的数据并分析仍需 2 分钟




• 这是爬取成功的提示

• 点击确定，即可看到分析图表




- 操作十分简单。若想看到清晰的大图，可以在 pic 文件夹里查看
- 这里的 Phase5 为测试数据

工具开源地址：https://github.com/suyiiis/learn/tree/master/%E5%A4%A7%E4%BA%8C%E4%B8%8A/%E6%95%B0%E6%8D%AE%E7%A7%91%E5%AD%A6%E5%9F%BA%E7%A1%80/%E5%A4%A7%E4%BD%9C%E4%B8%9A/emotionAnalysisTool

项目开源地址：https://github.com/suyiiis/learn/tree/master/%E5%A4%A7%E4%BA%8C%E4%B8%8A/%E6%95%B0%E6%8D%AE%E7%A7%91%E5%AD%A6%E5%9F%BA%E7%A1%80/%E5%A4%A7%E4%BD%9C%E4%B8%9A