COVID-19背景下的网络社会心态及公众情绪分析

一、小组成员信息及分工按排

1.项目GitHub开源地址

- DataScienceBigWork

2. 成员信息及分工按排

- 成员一：胡笑然(组长) 191250049@smail.nju.edu.cn

√ 参与讨论大作业思路

√ 通过随机抽样的新闻，依据上述的论文，构建2019-12-08~2020-02-20的关键词与心态情绪词的映射

√ 对所有构建的关键词和心态情绪词进行汇总，并人工构建心态词典

√ 对实验所得数据进行分析，并得出结论

√ 实验报告框架以及大量内容撰写

- 成员二：吴子玥 191250153@smail.nju.edu.cn

√ 参与讨论大作业思路

√ 爬虫编写

√ 通过随机抽样的新闻，依据上述的论文，构建2020-02-21~2020-04-20的关键词与心态情绪词的映射

√ 对TF-IDF进行分析并抽取关键词进行排序，绘制TF-IDF-Ranking条形图

√ 基于排序后的TF-IDF绘制Top关键词词云

√ 基于心态词典实现的SnowNLP情感分析，统计各情感分数段出现的频率得到的柱状图

√ 对实验所得数据进行分析并得出结论

√ 参与实验报告的撰写

- 成员三：贺伟 191250044@smail.nju.edu.cn

√ 参与讨论并确定大作业思路走向

√ 爬虫编写、获取所有新闻

√ 分词代码编写，将所有数据进行清洗、去停词、并将所有数据预处理

√ 对所获新闻数据进行数量与时间的分析，发现新闻数量是随着疫情的爆发和时间进行变化，并不符合高斯分布，绘制出（新闻数量-每两周）的直方图

√ 随机抽样代码的编写，取定比例0.2，获取随机抽样的样本

√ 从软件学报和知网中检索10+篇论文，对心态词典的构建提供思路

√ 通过随机抽样的新闻，依据上述的论文，构建2020-04-21~2020-06-20的关键词与心态情绪词的映射

√ TF-IDF计算代码的编写，依据分词的数据，将新闻数据分为四个阶段，分别计算好TF-IDF

√ 通过构建好的心态词典，编写代码，提取关键词的TF-IDF，并将其依据心态词典的5类，构建5维心态情绪向量

√ 编写代码提取出5维心态词典的TF-IDF，并编写绘制（5维心态+1维(数量/时间)）的散点图的代码

√ 通过分析四个阶段的6维散点图发现，其中积极的心态和感激的心态具有高度相关性，消极的心态和质疑的心态也有高度的相关性，所以认为可以将5维的情感向量进行PCA降维，降至三维度(积极、消极、严阵以待)

√ 编写PCA降维后数据的雷达图

√ 对上述绘制的图形进行分析并得出基础结论

√ 参与实验报告的撰写

二、问题背景及简单建模

1.问题背景:

新冠肺炎疫情突如其来，极大的影响了人民的日常生活和情感心态。在网络视角下，公众的情感表达更为显性，在社交媒体平台的活跃度和议题参与度明显增强。[1]情感作为新闻舆论的重要元素，对于网络舆论的生成以及社会心态的演变具有重大影响。随着互联网的发展以及中国网民的数量迅速增加，网络社会心态极大的反映了公众的情绪以及内心感受，有助于了解公众的社会心理需求，维持社会健康良好发展。新冠疫情治理期间，除了举国上下万众一心共同抗疫的主旋律，同时也有因疫情爆发产生的恐慌、焦虑、质疑的心态，也有听从指挥、居家抗议、自我隔离的积极、乐观的心态等等。本次我们以2020年初中国国内的新冠疫情为例，对重大突发公共卫生事件下的网络社会心态及公众情绪进行分析。

关键词：社会网络 公众心态 情感分析 新冠疫情

2.模型建立

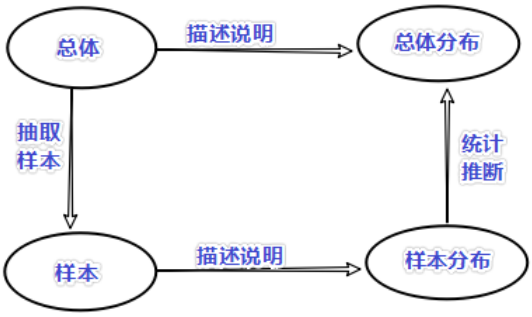
有统计抽样基本规律：

设$(X_1，X_2，\ldots，X_n)$是来自总体X的一个简单随机样本，将其一个观测值$(x_1，x_2，\ldots x_n)$的分量按从小到大的顺序排列成

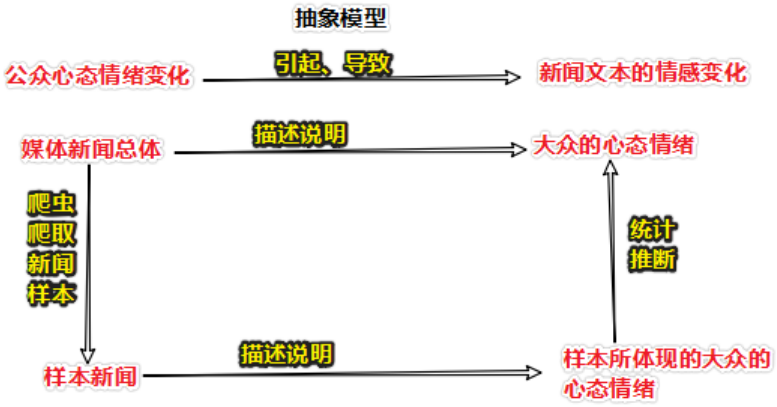
$x_{(1)}<x_{(2)}<\ldots<x_{(n)}$，其中$x_{(i)}(i=1,2,\ldots,r)$出现的频数为$n_i(n_1+n_2+\ldots+n_r=n)$，记

$$F(n)\begin{cases}0, & x\le x_{(1)}\\ \frac{k}{n}, & x_{(k)}<x\le x_{(k+1)},k=1,\ldots,n-1\\ 1, & x>x_{(n)}\end{cases}$$

为随机样本的经验分布函数，则有：$n\to\infty,F(x)\to$总体分布函数。有有样本和总体之间的规律：

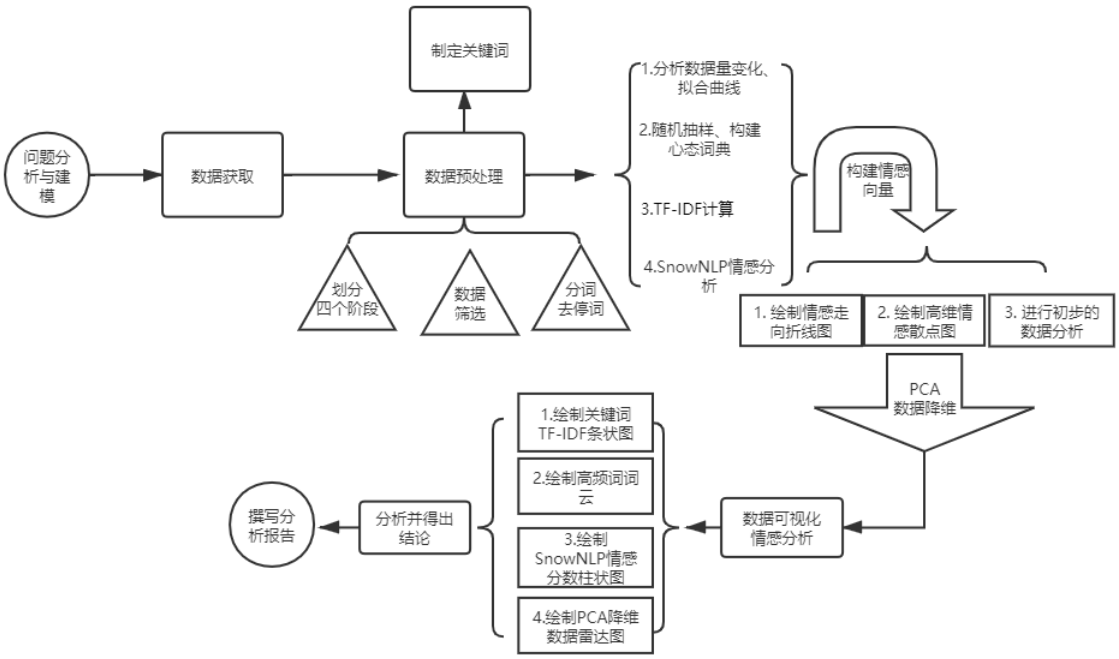


情感作为新闻的基本元素，那么公众的心态情感的变化，会引起描述公众情感的变化。因为我们可以通过对媒体新闻的文本做情感分析来统计推断公众的心态情感的变化。简单建立模型如下：



因此，我们对于公众心态情绪的分析可以通过抽取媒体新闻，对新闻进行文本情感的分析，从而推断出大众的心态情绪变化。

3.问题求解的主要过程



关键词：建模、关键词、数据处理、随机抽样、TF-IDF、分词、情感向量、统计图、PCA降维、数据可视化、SnowNLP

三、研究方法及实验过程：

在本次分析研究中，我们运用了多种方法，主要方法如下：

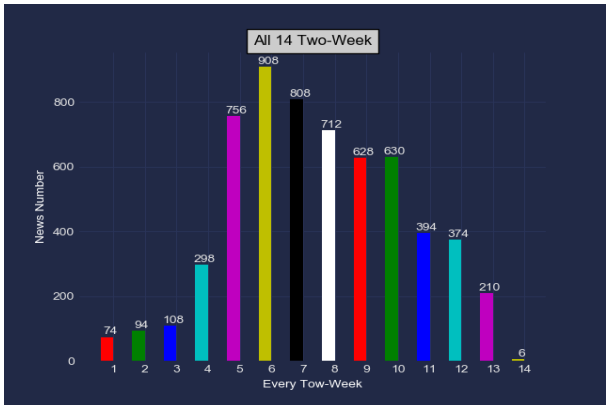
1.关键词[2]的制定

从研究主题[3]，基于高索引量、低竞争性、有效性三个标准来筛选过滤数据的关键词，对获取的新闻数据进行过滤，尽量减少非有效数据对实验过程和实验结果的干扰。我们选取的是新浪新闻和人民日报中从2019年12月8日至2020年6月20日的新闻，并且根据自定义的KeyWord.txt根据标题对其进行了筛选了，通过判断标题中是否含有关键字（如“疫情”“防控”等），保留了最可能和疫情相关的新闻。

2.数据量度量与曲线拟合

进行数据量度量和曲线拟合的原因：对于获取的数据，我们需要从数据量变化情况以及原因来判断数据是否有效。

前提：情感是新闻的基本元素，所以关于研究主题所以引起的情感心态的变化，会导致相关新闻数据量的变化，二者呈高度正相关。而随着疫情的爆发以及新冠病毒的蔓延，党和国家出台相关政策来指导抗疫，导致公众的情绪和心态并不会符和正态分布，所以新闻文本数量的变化情况也不会符和正态分布。



过程：我们对获取的的数据统计了每2周的数据量的变化情况，并进行了高斯分布拟合，发现不符合正态分布。于是绘制出（数据量-每两周）的直方图，发现总体图形变化符和常识及客观规律。

结论：所获得的数据有效，能够通过这些数据来统计推断大众心态情绪的变化。

3.数据清洗

原始数据是大段的新闻文本，并不能通过机器直接分析出情感。并且原始数据含有大量的无意义信息，会干扰情感分析，并且加大了计算资源的消耗。所以我们通过对原始数据分词，利用常用停词表对新闻文本数据去停词，极大的减少了文本中无效信息的干扰以及复杂度的降低。

基于对新闻依据(新闻文本数量-每两周)直方图和主要事件对新闻文本划分时间段，主要划分为以下四段：

1. 2019.12.08--2020.01.22：不重视与无奈扩散阶段
2. 2020.01.23--2020.02.07：资源缺乏阶段
3. 2020.02.08--2020.02.29：严格统一管控和物资配给阶段
4. 2020.03.01--2020.06.20：有序复工阶段

并对数据进行初步整合。

4. 随机抽样[4]

对于关键词的制定以及心态词典的构建，都会涉及到从已获得的数据中抽取样本这一过程。为了尽量减少人为操作的影响，降低抽样误差，所以我们采取随机抽样的方法，抽取总样本占比0.2的数据量进行人工标记处理。

5. 心态词典[5]的构建

基于已有关于文本情感分析以及大众心态的相关研究论文，初始确定心态情感词的分类。将情感分析归纳为 3 项层层递进的研究任务,即情感信息的抽取、情感信息的分类以及情感信息的检索与归纳。[6]

小组成员利用随机抽样的样本，对样本新闻进行划分，分工合作。按照句子、词两层进行两次的心态情感映射，并最后归纳总结出5类心态情感的词典。

6.SnowNLP情感分析

对数据文本中的情感词统计排序，基于已现有的SnowNLP情感分析模型，对四个阶段中公众的心态情绪变化做一个宏观分析，直观的分析了解大众心态情绪在疫情的蔓延、党和政府出台的政策、谣言传播、部分外国媒体的恶意中伤等因素的影响下，大众心态情绪的整体水平，便于为后续实验操作进展提供思路。

7.基于心态词典的TF-IDF的聚类[7]构建情感向量

由我们构建的抽象模型知，对于公众情绪心态的分析已经转化为对于抽取数据文本的情感的分析。TF-IDF是一种统计方法，用以评估一字词对于一个文件集或一个语料库中的其中一份文件的重要程度。[8]所以我们可以通过对于数据文本的关键词与心态词典映射关系，基于TF-IDF进行情感的聚类。

设心态词典中的心态情感可划分为K类，对于第i类心态有n个关键词映射，那么对单个文本第i类心态情感$E_i,(i=1,2,\ldots,K)$可聚类表示为：$E_i=\sum_1^n(TF-IDF_i)$；所以文本有K个不同的心态情感，所以构建该文本的K维的情感向量$Vector$表示为$Vector=[E_1,E_2,\ldots,E_K]$。

8.PCA数据降维[9]

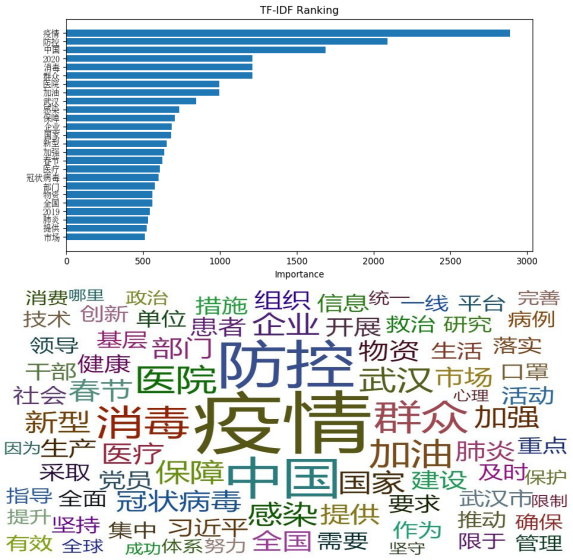
数据文本中的情感是多样的，即情感向量是高维向量，为了突出主要特征情感变化，所以我们需要想办法对大量的K维情感向量进行降维，采用PCA数据降维的方法。在实验过程中，通过情感向量随着新闻数量变化的散点图发现”积极、自信“的心态同”感激、感谢“的心态呈高度正相关，”消极、担心、恐慌“的心态同”质疑、愤怒“的心态也呈高度相关，所以我们可以通过PCA对原有的5维情感向量进行降维，降至三个维度，减少情感向量维度带来的分析难度以及对主情感的干扰。

9. 数据可视化

对于问题的分析建模以及实验，最终目的是为了获得实验的结果和结论。所以我们采用数据可视化的方法将整个实验过程中的实验所得数据进行可视化操作，包括但不限于绘制统计图表等。

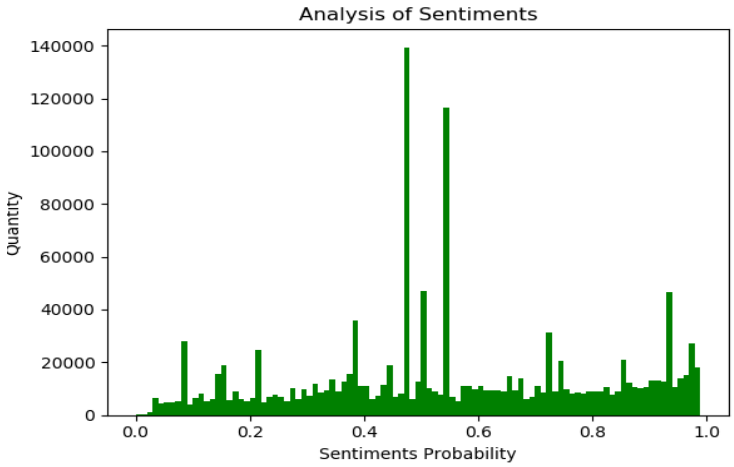
四、数据可视化及数据分析

1. 基于TF-IDF提取关键词后的词云和TF-IDF-Ranking条状图



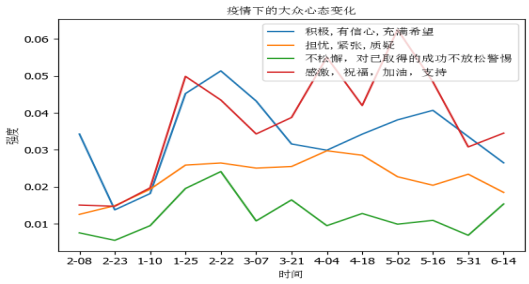
在我们所筛选的新闻中，“疫情”、“防控”、“中国”等词无疑出现了最多次，这是与我们的预计相符的，因为我们利用自定义的关键词（KeyWord.txt）根据标题对新闻进行了筛选。除了“疫情”“防控”“中国”“2020”等客观的，不带有感情色彩的词之外，如“加油”“加强”“保障”“全力”等带有感情或者希冀的词语也出现了较多次，可以观察出，在疫情期间，社会心态总体是积极、向上的，面对疫情，无论是公众还是国家，也都在进行奋力的抗疫斗争。并且疫情，防控，消毒，医疗，物资等是人们尤为关注的话题，体现了百姓对疫情的关注度之高，重视程度之高。同时，词云图中亮眼的加油，坚持也体现了百姓对一线人员的尊敬，以及对政府工作的信任支持。

2.基于SnowNLP情感分析的情感分数段频率柱状图



将2019.12~2020.6月的新闻进行分词处理后，利用基于心态词典实现的SnowNLP情感分析，统计各情感分数段出现的频率得到的柱状图，其中越接近1为积极，接近0为消极。可以看见大于0.5的比重更多，这也能说明总体上积极的心态占比更重。此外，0.5左右的占比最多，这可以说明我们对于疫情有着清醒的认识，在保持积极心态的同时也能正视问题，严阵以待。

3. 基于时间段分类的TF-IDF情感变化折线图



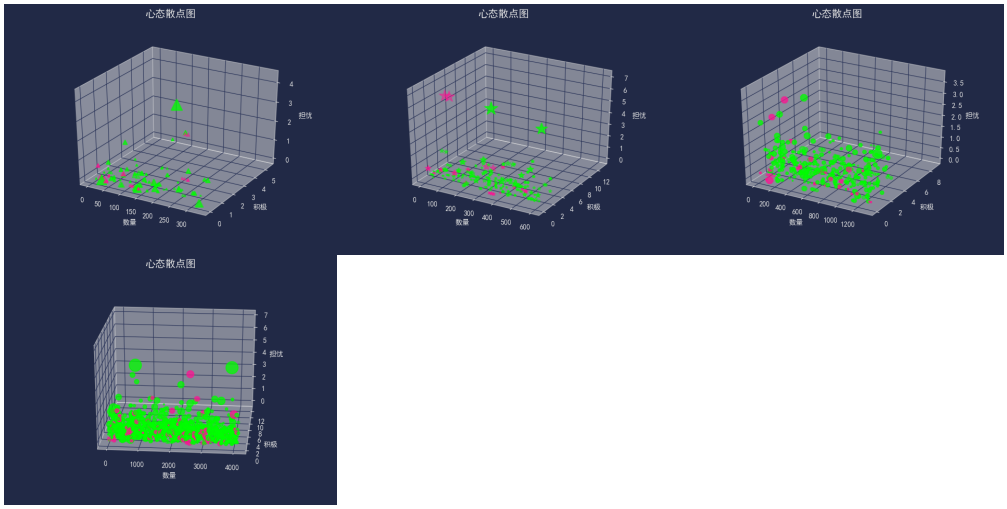
在疫情爆发初期（2019.12.08--2020.01.25），由于最初对于疫情的不了解和轻视，发生了一些让民众对政府和部分组织感到不信任的事情，也由于最初物资的紧缺和形势的危急混乱，一些本来可以避免的损失未能被避免，因此，这段时间里，担忧和紧张的情绪逐渐上涨，几乎和积极的情绪齐头并进。而在2020.01.20左右，当国家真正开始重视疫情，并且各地给予湖北武汉以援助，物资短缺的情况得到缓解之后，积极以及感激的情绪达到了顶点，担忧和质疑的情绪虽然仍然存在且还在增长，但是与积极的情绪已经不可同日而语。这生动地反映了民众心态的变化，由疫情初始阶段的不安和担忧，已经逐渐转变为了积极乐观的心态。而在2020.03.21抗疫基本取得成功之后，更多的感情则由原来的担忧、紧张变成了对于在抗疫斗争中无私奉献的医护人员的感激与敬佩。此外，需要注意的是，不松懈，坚定抗疫的情绪从最初一直持续到了最后，这也是我国抗疫能成功的原因之一。

3.基于时间段分类的情感散点图

通过按照重要事件分为四个时间段，依次绘制各个时间段内的初始5维情感向量 + 数量变化的散点图：

注意高维散点图的各元素含义：

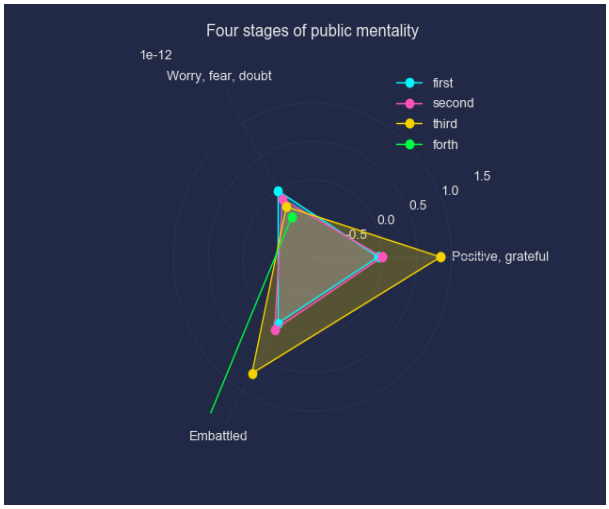
1. X轴：表示积极的心态情绪状况；
2. Y轴：数量轴即时间轴，二者是有一个正相关；
3. Z轴：表示担忧的心态情绪状况；
4. 图形的大小：表示质疑、焦虑、恐慌的心态情绪状况；
5. 图形的蓝绿色：表示感激、祝福、加油、支持的心态情绪状况
6. 图形的深粉红色：表示严阵以待、自觉抗疫、不松懈的心态情绪状况



通过对这四幅图的横向和纵向比较可以得知：

1. 在疫情发现初期：图形数量较少、并且面积也比较小，体现了人们大都比较平常，心态情绪相对平和，也在一定程度上说明了对于“未知肺炎”的不重视。
2. 在第二个阶段（资源相对缺乏的状况）：很明显的，图形面积迅速增大，表现了大众对于疫情没能及时制止通告的质疑，恐慌以及愤怒。同时深粉红色图形明显增加，表现了人民大众在了解疫情的严重后自觉抗疫、不松懈的心理状态。
3. 在严格统一管控和物资配给阶段：很明显能感受到图形的蓝绿色大幅度增加，表现了人民大众对于政府统筹抗疫的决心以及自信。于此同时，图形面积的大小和高度也在增加，这一方面也受国际疫情迅速恶化的影响，并且国际上一些不怀好心的人大肆宣传这是由于中国武汉导致的病情。
4. 有序复工阶段：随着举国上下万众一心的抗疫，国内疫情基本得到控制，公众的心态逐渐恢复为正面积极的状态，同时随着公司的有序复工，严阵以待、自觉抗疫、不松懈的心态情绪状况仍将达到高潮。
5. 很重要一点，我们发现，积极、自信的心态和加油、感激的心态具有很强的正相关性，而质疑、恐慌的心态和担忧焦虑的心态也具有很强的相关性，故此，我们可以通过PCA将这5维的情感降维至3维。

4.基于PCA降维，对四个时间段内情感求和的心态情感雷达图



1. 通过降维后的雷达图，很明显的可以看出随着抗疫的进展以及对于新冠肺炎的普及，大众对于新冠肺炎的恐惧明显减少
2. 同时随着党和国家的统筹抗疫，严厉打击抗疫期间违法违纪的官员，全国一盘棋，显著的增强了公众的对党和国家的自信以及对于医护人员的感激之情。
3. 注意到，第四阶段，随着复工复产的进行，工人们严阵以待，自觉隔离，居家办公或定点办公，体现了国家宏观政策对于公众心态的导引。但第四阶段的图形并没有闭合，在进行PCA数据降维时，丢失了部分数据信息。

五、结论及研究意义

2020年初暴发的新冠肺炎疫情是新中国成立以来发生的传播速度最快、范围最广、影响最大的重大突发公共卫生事件，它在挑战人们日常生活秩序的同时，也给人们的社会心态造成严重的负面冲击。通过分析疫情期间的新闻，我们完成了对疫情爆发期间社会心态变化过程的认识。从最开始的担忧紧张乃至质疑，到最后的积极乐观和平和，心态的变化总体上符合我们根据经验做出的预设的预设，且可以用事实来进行解释。

从上述研究可以看出，面对重大突发性公共卫生事件的爆发，会迅速影响公众的心态和情绪，扰乱人们日常生活秩序。同时谣言、新闻媒体的报导、国家宏观政策的引导都会极大的影响人们心态的走向，因此，为了积极的引导公众的心态情绪，我们需要做到：

1. 不信谣、不传谣
2. 辩证的看待新闻媒体的报导
3. 国家及时迅速的出台相关治理政策

4. 积极的舆论引导
5. 坚定保证日常生活所需

来保证社会公共秩序的良好运转和有效的疫情防控治理。

六、参考文献及引用

1. 刘海明，宋婷.共情传播的度量：重大公共卫生事件报道的共振与纠偏[A].新闻界，2020(10). ↩

2. 关键词，附件1. ↩

3. COVID-19背景下的网络社会心态及公众情绪分析 ↩

4. 随机抽样,附件2. ↩

5. 心态词典 附件3. ↩

6. 赵妍妍, 秦兵, 刘挺. 文本情感分析[J]. 软件学报, 2010. ↩

7. 聚类分析 ↩

8. TF-IDF,维基百科 ↩

9. PCA ↩