COVID-19背景下的网络社会心态及公众情绪分析

贺伟 191250044@smail.nju.edu.cn
胡笑然 191250049@smail.nju.edu.cn
吴子玥 191250153@smail.nju.edu.cn

GitHub开源地址:https://github.com/hewei-nju/DataScienceBigWork

摘要： 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

一、问题背景及简单建模：

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

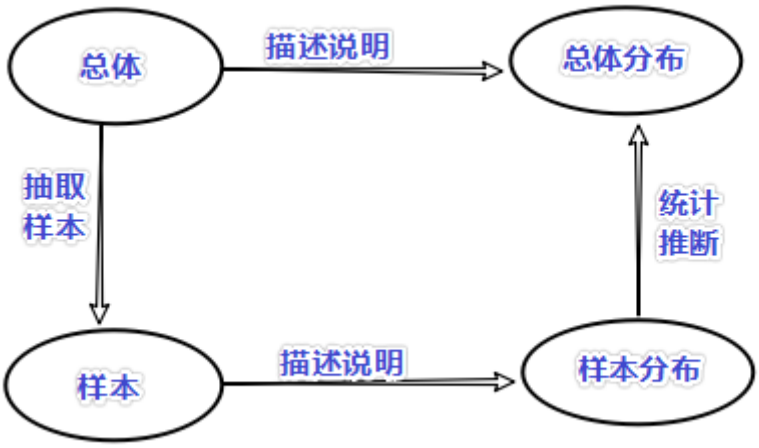
$$x_{(i)}\ (i=1,2,\cdots,r)$$

出现的频数为

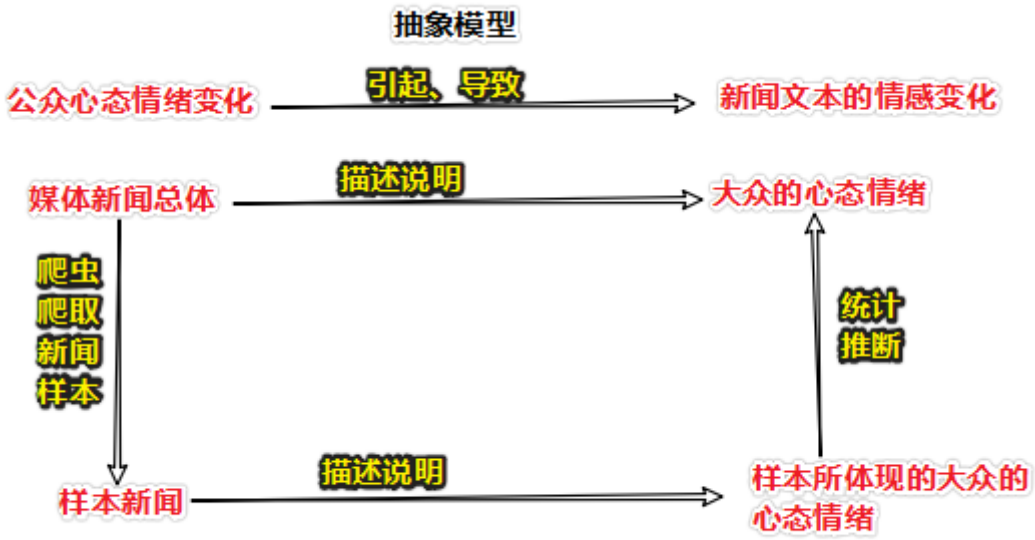
$$n_i\ (n_1+n_2+\cdots+n_r=n)$$

，记 $$F_n(x)=\begin{cases}0, & x\le x_{(1)}\\ \dfrac{k}{n}, & x_{(k)}<x\le x_{(k+1)},k=1,\cdots,n-1\\ 1, & x>x_{(n)}\end{cases}$$

为随机样本的经验分布函数，则有：$n->\infty,F_n(x)->$总体分布。有抽样分布的规律：



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：

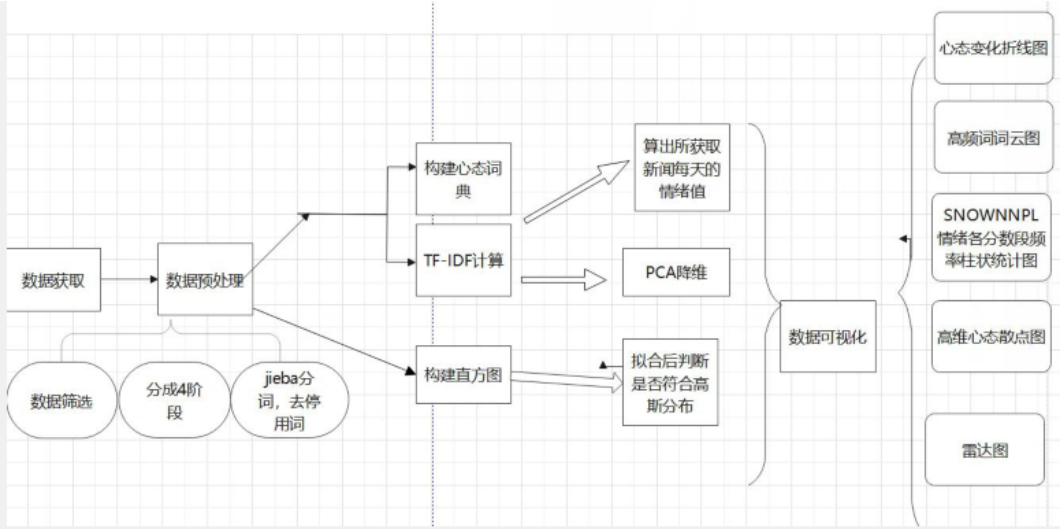


因此，我们对于公众心态情绪的分析可以通过抽取媒体新闻，对新闻进行文本情感的分析，从而推断出大众的心态情绪变化。

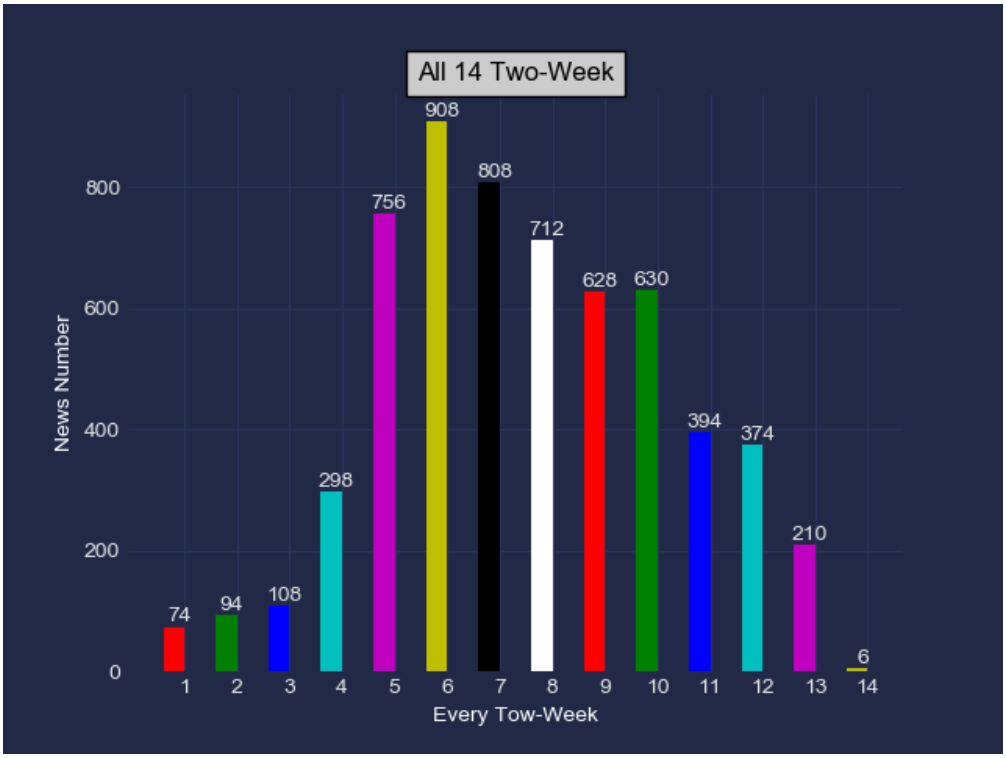
二、研究方法：

1. 研究流程：本次作业中，我们运用多种方法，按以下流程进行了研究：
    1. 下载新闻文本，按照一定格式保存至txt文档。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. 对所有的新闻文本按照时间轴进行分类；根据疫情的发展过程，我们将这段时间`(2019.12.8-2020.6.20)`分成了四个时间段：
        - 2019.12.8--2020.1.22：不重视与无奈扩散阶段
        - 2020.1.23--2020.2：资源缺乏阶段
        - 2020.2--2020.3：严格统一管控和物资配给阶段
        - 2020.3-2020.6.20：有序复工阶段

3. 如上分层之后，我们在每层中通过随机函数随机抽取样本新闻（总共占新闻总量的20%），并人工对这些样本进行了分析：
    - 分析样本，构建心态词典。我们对上一步中随机抽取的10%的样本新闻进行了人工分析，找出了其中含有情绪的语句，并提取了情绪词，将其与特定的情绪相关联，形成映射，构建了针对此次事件的心态词典。
    - 词频分析（计算TF-IDF）。在建立心态词典的同时，我们还对经过分词之后的新闻样本进行了词频分析，并且在对样本进行了分析之后，我们将构建出的心态词典用于总体，分析出了整个事件发展的过程中，社会心态和民众情绪的变化，具体的数据将于下文中进行分析。
    - 情感量化后大数据分析。
4. 流程图:



2. 研究方法：
    1. 获取新闻文本数据：
        - 通过使用python编写爬虫，对新闻文本进行标签提取，获取新闻文本等有用信息，并通过`json`格式进行存储和交互。
    2. 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，去掉部分无意义的停用词(停词表)
    3. 数据分析:
        - 先通过对数据获取来源进行分析，由于数据是取自新闻媒体，并且是由于疫情导向引起的情感变化，通过高斯分布拟合后，发现并不符和正态分布，故绘制出其每两周新闻数量的直方图如下：

- 

  - 有直方图知：随着疫情的爆发，大众的心态变化巨大，从而导致媒体新闻发布量激增；随着国家出台相关政策进行协调抗疫，人民的心态慢慢缓和。
  - 随机抽样[2]：将数据进行简单的清洗后，我们只是得到了初步的数据，并不能直接获取文本所表示的情感。并且情感在描述性新闻中占比并不大，所以我们需要提取出能够表现心态情绪的关键词进行心态词典的构建。为了尽量减少人为因素的干扰，我们通过电脑进行了随机抽样，并对抽取的样本进行仔细地分析来构建心态词典[3]。
  - TF-IDF[4]词频统计：在获取心态词典后，我们目的是通过文本来分析出疫情期间各件事情对于大众心态情绪地影响。所以，我们需要分析出新闻文本中关键词的TF-IDF来量化、特征化大众的心态情感的变化。
  - PCA[5]数据降维：通过查看附件[3] (心态词典)知，我们将大众的情感归纳为了5类，故此，每个文本都可以量化为一个5维的情感向量[^6]，为了更加真实的推测出大众的心态情绪的变化，我们在加大了获取的新闻样本，即我们有大量的5维情感向量。所以我们想到通过PCA对情感向量进行降维，提取出主要特征进行分析。

4. 数据可视化：

   1. **词云**：调用`PyEcharts`中的`WordCloud`子包画图。通过前期计算，得到词频高于55的高频关键词，利用`PyEcharts`进行词云可视化分析。
   2. **情感各分数段出现频率柱状图**：将12~6月的新闻进行分词处理后，利用基于情感词典实现的`SnowNLP`情感分析，越接近1为积极，接近0为消极，统计各情感分数段出现的频率并绘制对应的柱状图。，。代码如下：
   3. **TF-IDF统计可视化**：通过TF-IDF计算过滤掉一些常见的却无关紧要本的词语，保留影响整个文本的重要字词后,将TF-IDF统计可视化。
   4. **心态词典与情感分析**：人工建立心态词典，分别为
      - 积极,有信心,充满希望；
      - 担忧,紧张,质疑；
      - 不松懈，对已取得的成功不放松警惕；
      - 严阵以待；

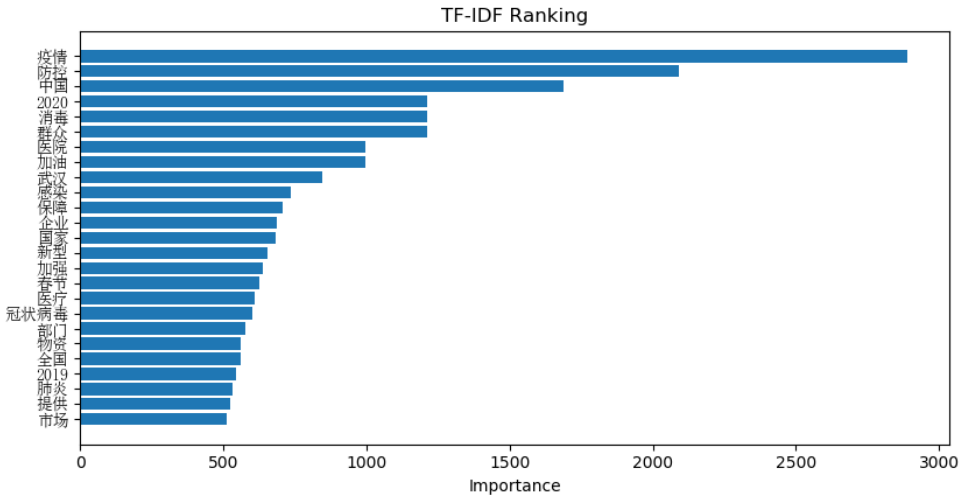
- 感激，祝福，加油，支持5个基础心态list。

5. 对已获得的新闻算出每一天多条新闻的词语的频率并排序，对应心态词典统计出当天五个基础心态的情绪值。

三、数据分析

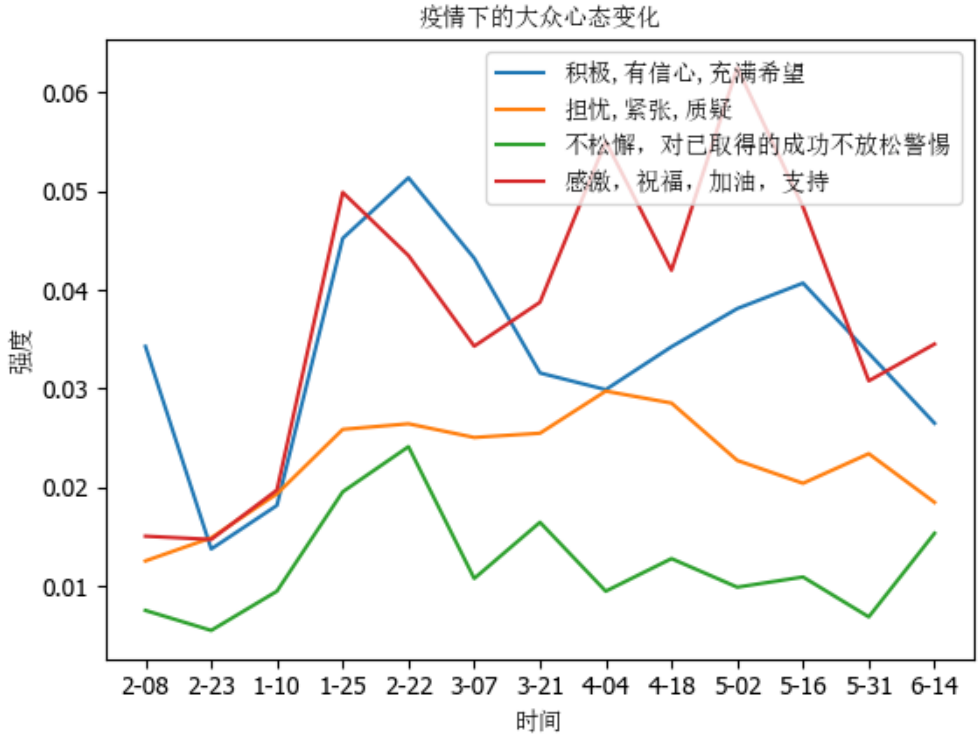
1. 在我们所筛选的新闻中，“疫情”、“防控”、“中国”等词无疑出现了最多次，这是与我们的预计相符的，因为我们利用自定义的关键词（KeyWord.txt）根据标题对新闻进行过了筛选。除了“疫情”“防控”“中国”“2020”等客观的，不带有感情色彩的词之外，如“加油”“加强”“保障”“全力”等带有感情或者希冀的词语也出现了较多次，可以观察出，在疫情期间，社会心态总体是积极、向上的，面对疫情，无论是公众还是国家，也都在进行奋力的抗疫斗争。

2. TF-IDF直方图 & 词云：



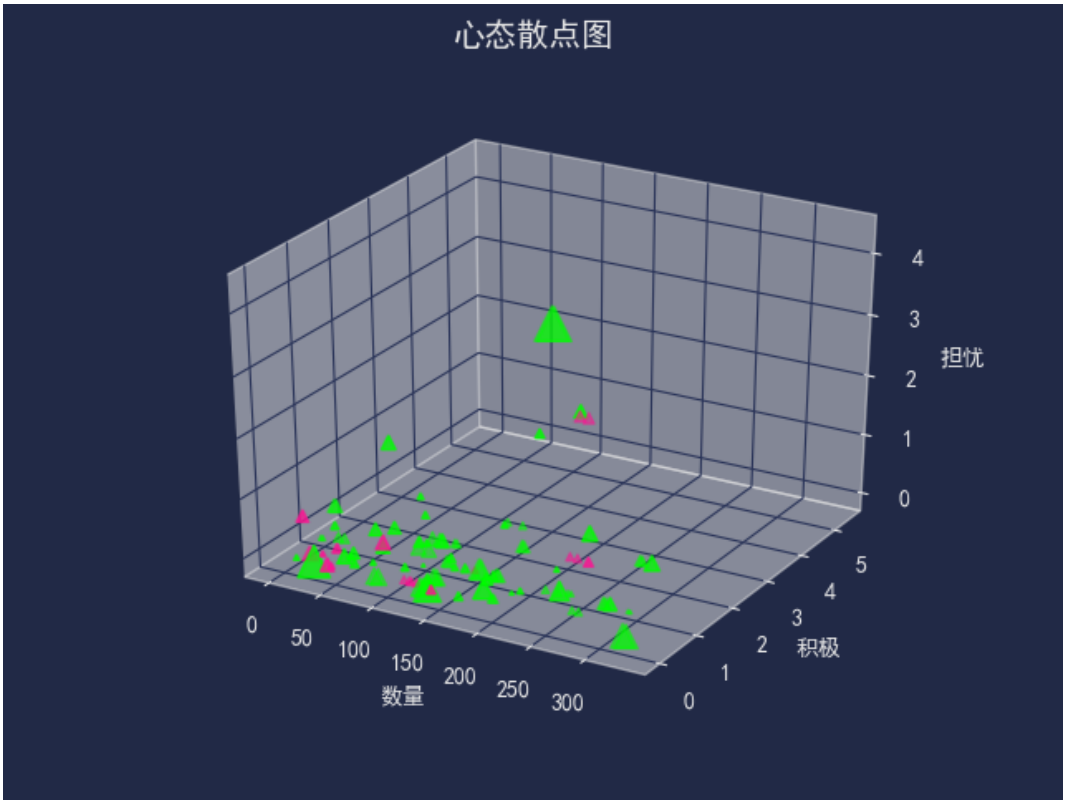
- 如图所得，疫情，防控，消毒，医疗，物资等是人们尤为关注的话题，体现了百姓对疫情的关注度之高，重视程度之高。同时，词云图中显眼的加油，坚持也体现了百姓对一线人员的尊敬，以及对政府工作的信任支持。

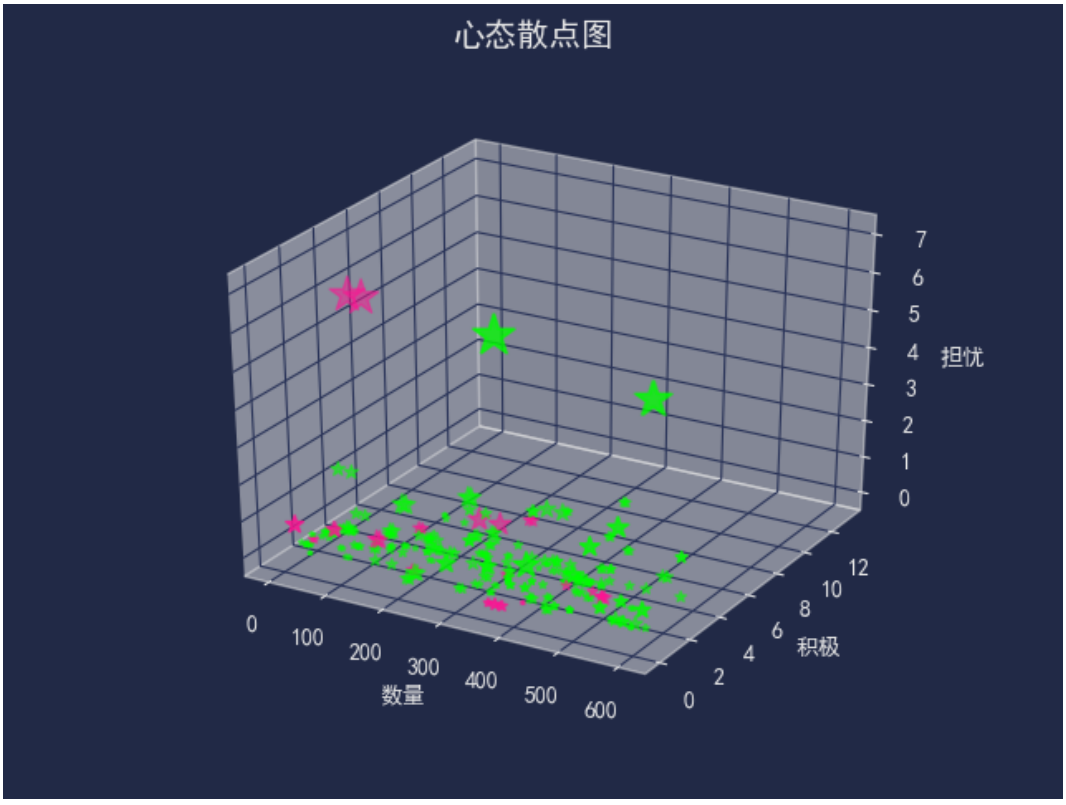
3.



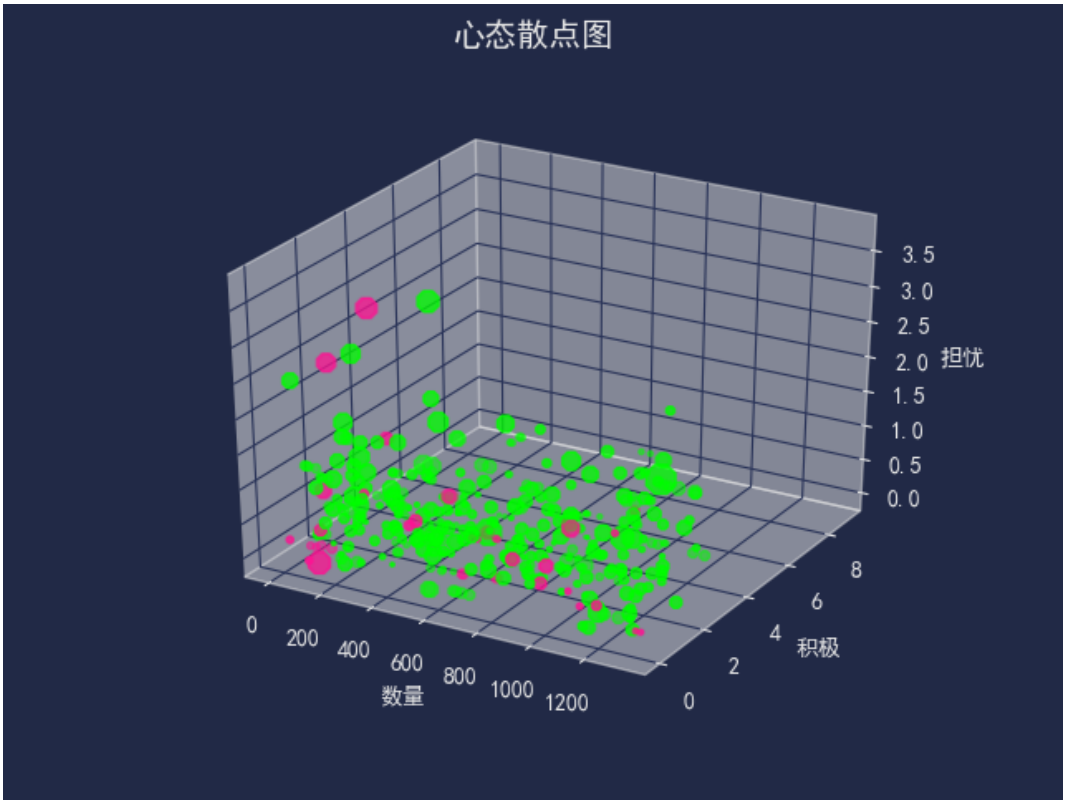
- 在疫情爆发初期（2019.12.8-2020.1.25），由于最初对于疫情的不了解和轻视，发生了一些让民众对政府和部分组织感到不信任的事情，也由于最初物资的紧缺和形势的危急混乱，一些本来可以避免的损失未能被避免，因此，这段时间里，担忧和紧张的情绪逐渐上涨，几乎和积极的情绪齐头并进。而在1.20左右，当国家真正开始重视疫情，并且各地给予湖北武汉以援助，物资短缺的情况得到缓解之后，积极以及感激的情绪达到了顶点，担忧和质疑的情绪虽然仍然存在且还在增长，但是与积极的情绪已经不可同日而语。这生动地反映了民众心态的变化，由疫情初始阶段的不安和担忧，已经逐渐转变为了积极乐观的心态。而在3月21日抗疫基本取得成功之后，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。

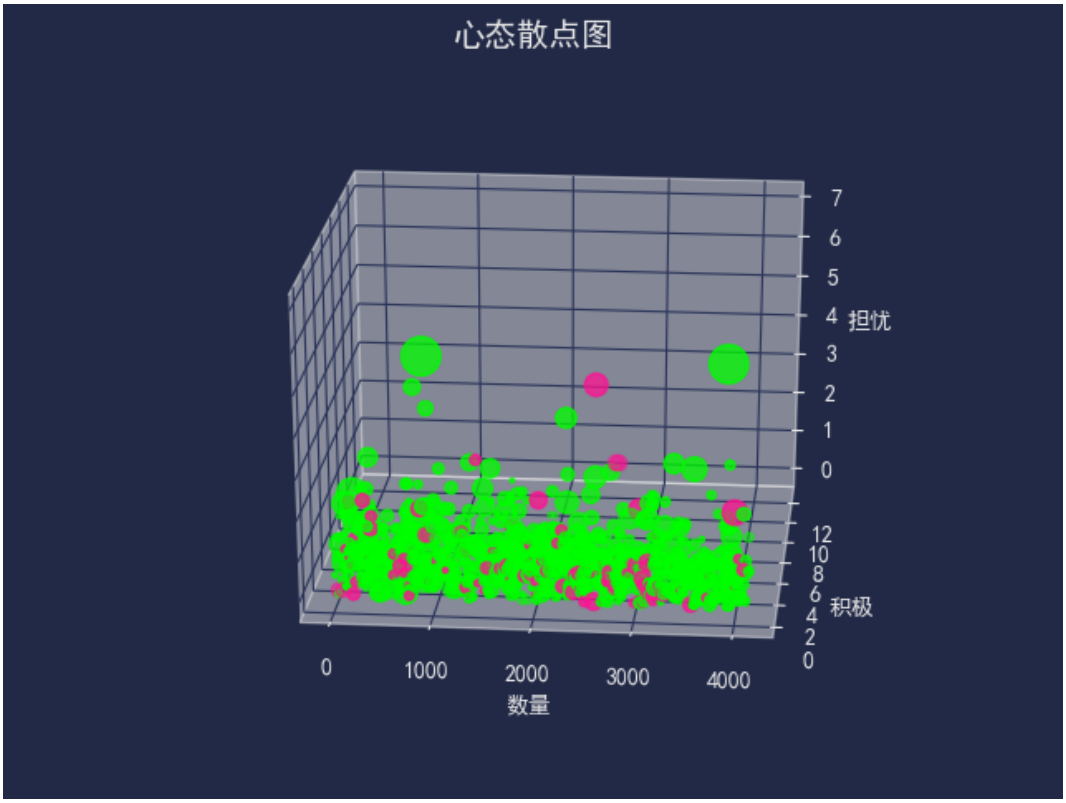
4. 通过按照重要事件分为四个时间段，依次绘制各个时间段内的初始5维情感向量 + 时间(数量）的散点图：
    - X轴：表示积极的心态情绪状况；Y轴：数量轴即时间轴，二者是有一个正相关；Z轴：表示担忧的心态情绪状况
    - 图形的大小：表示质疑、焦虑、恐慌的心态情绪状况；图形的蓝绿色：表示感激、祝福、加油、支持的心态情绪状况
    - 图形的深粉红色：表示严阵以待、自觉抗疫、不松懈的心态情绪状况
    -


心态散点图
担忧
积极
数量
0
1
2
3
4
5
0
50
100
150
200
250
300

- 
心态散点图
担忧
积极
数量
0
1
2
3
4
5
6
7
0
2
4
6
8
10
12
0
100
200
300
400
500
600

-



- 

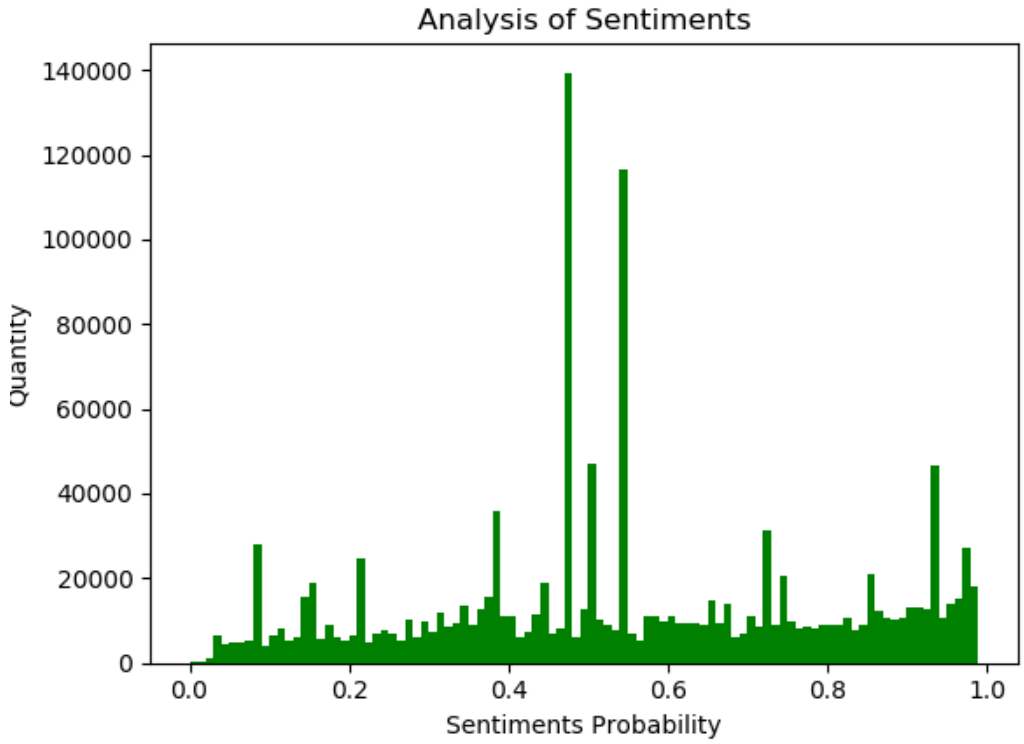
- 通过对这四幅图的横向和纵向比较可以得知：

    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，得到如下雷达图：

- 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
- 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。
- 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。

5.



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。

四、结论

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。除此以外，我们也得到了一些更为细节的结论，这些结果在上面的分析中均有提及。通过此次作业，我们认识到了大数据技术在社会心理学上运用的广阔前景，也了解到数据科学可以服务社会的方方面面。

五、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. 随机抽样 & 附件1. ↩

3. 心态词典 附件2. ↩ ↩

4. TF-IDF 附件3. ↩

5. PCA ↩