南京大学软件学院

新冠肺炎疫情下的网络社会心态分析

数据科学大作业实验报告

小组组长：樊言鹏　小组成员：王子岑 刘晋元

2021-1-24

目录

第一部分 实验部署

1. 实验背景

中国社会正处在深刻而快速的转型期，其中，在社会变迁层面，社会结构的快速分化，以“撕裂”的方式强化了社会团体、阶层之间的张力，使得整体社会结构出现紧张（李汉林、魏钦恭、张彦，2010），并投射在个体心理层面，进一步凸显出公众的社会认知、情绪、信念、意向、行动等对社会治理的重要影响（王俊秀，2014；杨宜音，2006）。 同时，随着互联网应用的不断普及，日益多元复杂的公众情绪，借助网络的力量传播和放大，对社会心态的塑形力量进一步增强，赋予了群体心理及集体行为的极化可能（周晓虹，2014）。当下新型冠状病毒（COVID-19）肆虐全球，给人们的生产和生活产生了极大影响，也形成了疫情下独特的网络社会心态和公众情绪。因此，立足此次新型冠状病毒（COVID-19）重大突发公共卫生事件情境，借助适宜的数据与计量手段，准确并客观地了解公众的网络社会心态与基于此呈现出的行为规律，就可能实现公众的情绪引导，让大众以积极的心态与政府一起应对和处理公共卫生事件及其衍生问题，维护国家与社会的长治久安。

2. 实验要求或目标

实验要求：进行数据爬取、筛选、关键词提取、心态词典建立、心态映射以及数据可视化处理与分析。

目标：在此次新型冠状病毒（COIVD-19）传播这一重大公共卫生事件情景下，以大数据技术深描中国大众的网络社会心态。

3. 实验人员及分工

樊言鹏：学号：191250028，联系方式：18315040075

数据爬取，筛选，关键词提取，心态词典建立，心态映射，数据可视化的校正、实验报告。

刘晋元：学号：191250092，联系方式：18851863820

心态词典建立、数据可视化、实验报告数据分析。

王子岑：学号：191250144，联系方式 18851863310

关键词提取、心态词典构建、初步进行数据可视化、实验报告目录。

4. 实验数据来源及工具使用

数据：

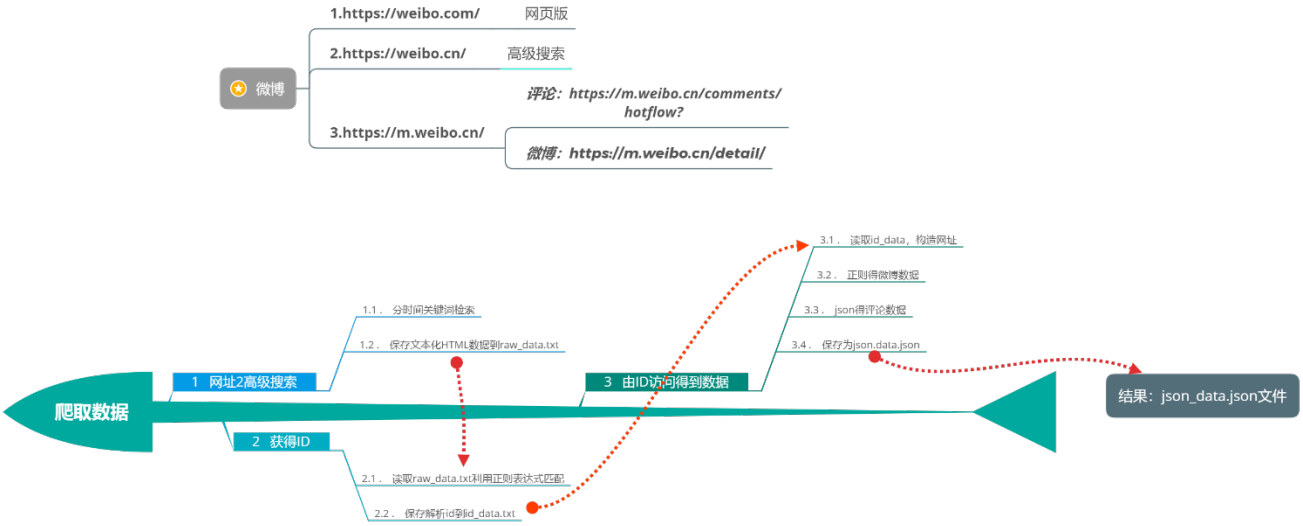
1. 本次实验的语言数据来源为<新浪微博人民日报账号 https://m.weibo.cn/profile/2803301701>
2. 构造心态词典参考<**大连理工大学中文情感词汇本体库**>

工具：

1. 数据获取以及处理阶段使用 python
2. 数据可视化阶段使用 python 和 Minecraft 制图

第二部分 实验步骤

1. 爬取数据



思路：我们的目标是爬取人民日报官方微博的与疫情相关的新闻信息与评论，并进行对重点新闻、评论的拟合和筛选。微博共有 3 个不同的访问网址：https://weibo.com/、https://weibo.cn/、https://m.weibo.cn/，分别通过这几个网址进入人民日报微博主页，监视网页 network 数据，刷新网页后找到网页数据文件，获得请求网址，并获得 cookie 和 User-Agent，在 python 中用 requests 库，用 header 与网址进行访问。我发现移动端的网址，在由 python 发起的请求下都可以访问首页的数据，其它两个网址只会跳转到登录页面，即便已有网页登录的记录，也不能访问。移动端微博

第一页有 20 条评论左右，考虑到对每条微博的评论数量需求并不大且模拟登录面临微博的反机器登录的屏障，我决定在不登录的情况下，爬取每条微博的第一页评论数据。因为移动端的微博的评论默认是按其关注程度自上而下排列，对于人民日报这样的关注度极高的账号，选取第一页即相当于选取了这条微博关注度最高的评论，因此也就不需要再次对评论的重要性做拟合和筛选。现在的问题转化为了如何获得每一条微博，通过不断的对网址发出 requests 请求和查询相关资料，我发现移动端可以通过前缀（https://m.weibo.cn/detail/）加上微博的 ID 号获得唯一确定微博的 html 内容，其中包含此微博的所有信息，有了 html 数据，就能通过正则表达式解析出我们需要的数据。在对应的微博页面上我继续使用了网络监视，每次打开评论页面都会有一个相似的请求网址，经过观察分析可以发现评论的请求网址构造为前缀
(https://m.weibo.cn/comments/hotflow?)+?id=XX&mid=..&max_id=.. ，即也可以通过 id 访问评论页面，并且能得到一个 json 文件。通过 json 文件的解析和网页数据的结构，就可以取出我们需要的第一页所有评论。只要获得微博 id，就能通过移动端获取需要的所有数据，所以现在的问题就转化为了如何获得我们需要的微博 id，https://m.weibo.cn/ 的高级搜索功能每一次可以显示一页的数据，通过对搜索网页结构的分析，我发现高级搜索结果是一个数据格式为 html 的页面，每个页有十条左右的微博，每条微博有一个对评论页面的引用链接，例如：
<"https://weibo.cn/comment/IqI21v3hS?uid=2803301701&amp″>，经过访问检验可以发现 comment 后面一个数据恰好此页面的 id。于是我获得每一页的 html 数据并将其存储在 raw_data 文件夹中，然后通过正则表达式获取所有的引用里面的 id，分阶段存储。之后用所有 id 访问微博的页面，获取所有需要的数据后统一存储为 json 文件。数据爬取的工作到此就结束了。我们使用 python 爬虫作为工具爬取了自 2019.12-2020.6 期间人民日报所发微博的正文及评论，其中微博正文共有 2272 条数据，微博评论有 41009 条。

```
pattern=re.compile('comment/.*?uid')
comment=pattern.findall(rawData)
#对 comment 稍加处理即可获得所有的 id

#将 id 存储，方便访问网页时读取.
store=" ".join(IDlist)
with open('id_data\\1.txt','w') as f:
    f.write(store)
    f.close()
```

2. 数据筛选与拟合

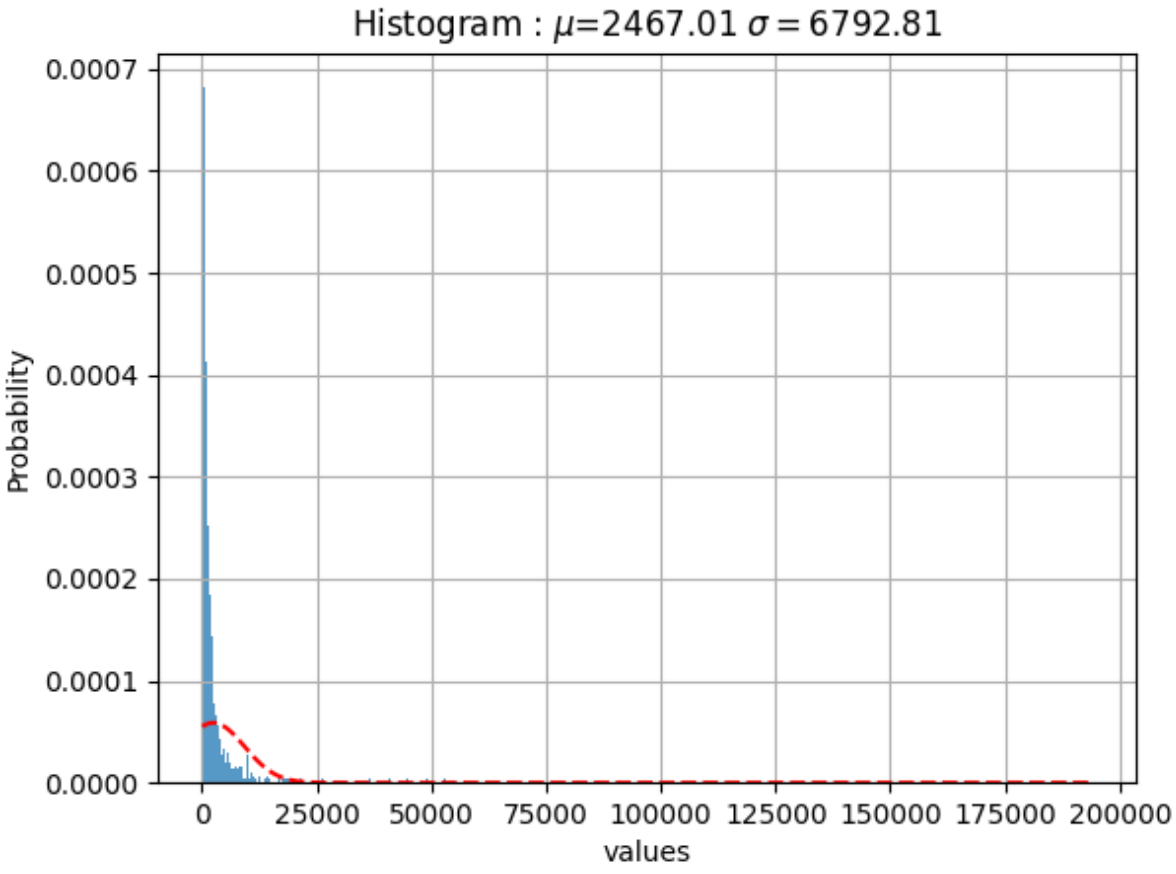
数据筛选：本次数据爬取基于 https://weibo.cn/ 的高级搜索，关键词为疫情，为了进一步筛选数据，我定义其他疫情相关的关键词，检查关键词列表是否在所得新闻内容中，如果没有数据在其中，去掉此条微博，对四个阶段数据分别进行此操作，得到 pure_json_data.json 文件，这是用于关键词提取的基础文件。

##示例

```
select_list = ['武汉','隔离','口罩','新冠', '抗疫']
```

数据拟合：对于各阶段的微博评论数进行了拟合，使用了 `import matplotlib.pyplot as plt` 画图以及 `from scipy import stats` 统计库进行正态分布检验，四个阶段的 p 值都接近零，远远小于 `0.05`，即认为微博评论数量的分布不是正态分布。

第四阶段微博评论数分布（其他阶段见 https://github.com/wsfyp/data_science_byFLW/tree/master/fan_code_data/pic）：



3. 提取关键词

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。字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降（百度百科）。

在使用结巴分词库分词后，我们对结果运用了 TF-IDF 方法：**先计算 IDF**：将所有词语去重，计算总句子数，计算每一个词语的在此阶段的句子中出现次数，由 IDF 构造公式计算每个词语的 IDF 值，以此构造字典。**再计算 TF**：计算每个词再句子中的出现频率，与 IDF 字典对应项相乘后，进行排序，构造 json 文件，获得每个句子排序后的词语。对每句话来将，排名越靠前的词语越能代表这一句话，当心态词典构造完成后，按词语 IDF 从大到小检查是否在词典中，若 hit，那么这条评论就被映射到一种心态。

4. 心态词典建立

思路：构建一个心态字典，按键值将不同的情感心态词（key）映射到心态词（value），要求心态词要足够的全面，映射心态要足够地准确。我们使用了大连理工大学的中文情感词汇本体库，原词典情感词分为 7 个大类。根据分析，我们决定将 7 大类情感映射为五个类心态，即【喜悦，积极，愤怒，伤感，焦虑】最为此次心态分析的 5 大基本心态。

构建方式：读取 txt 文件的情感库，通过正则匹配的方法取出每一组的心态词和对应的心态，构造一个心态词与 5 大心态的对应字典，并将其存储为 json 文件便于心态词典映射。结果为 pic_dic.json。

```
#对应关系：
def get_psy(psy_id):
    if psy_id in ['PA','PE']:#赞扬 安心
        return '喜悦'
    if psy_id in ['PD','PH','PG','PB','PK']:#尊敬 赞扬 相信 喜爱 祝愿
        return '积极'
    if psy_id in ['NA']:#愤怒
```

```
        return '愤怒'
    if psy_id in ['NB','NJ','NH','PF']:#悲伤 失望 愧疚 思
        return '伤感'
    if psy_id in ['NI','NC','NG']:#慌 恐惧 羞
        return '焦虑'
    if psy_id in ['NE','ND','NN','NK','NL']:#烦闷 憎恶 贬责 妒忌 怀疑
        return '厌恶'
    if psy_id in ['PC']:#惊奇
        return '惊奇'
    return '异常'
```

5. Hit 心态词典

思路：基于已经存储的心态词典和分阶段获得的 TF-IDF 关键词数据，将每一句评论中的关键词通过心态字典映射到对应的心态，按此方式便利每一条评论的已排序 TF-IDF 列表，若映射成功或句子中无心态词，则进行下一条评论。（关键词映射）

实现：每一次映射成功，hit_dic 的 key 对应的 value（击中次数）就会加一，以此记录所有评论心态在在 5 中定义心态中的分布情况。

```
#words 为一句评论或新闻中按 TF-IDF 排序的词语
for word in words:
    if word in psy_dic:
        hit_dic[psy_dic[word]]+=1
        break
```

#一旦命中，结束对这个句子词语的遍历，用这个词作为句子的心态关键词，除了保存心态词命中频率，我也记录了对评论中心态关键词的出现次数，并在构建 json 文件之前进行了排序，见 https://github.com/wsfyp/data_science_byFLW/blob/master/fan_code_data/psy_result/keywords1.json

6. 数据可视化

- 通过 psy_result.json 文件结合 Microsoft 绘图得到各阶段心态分布图

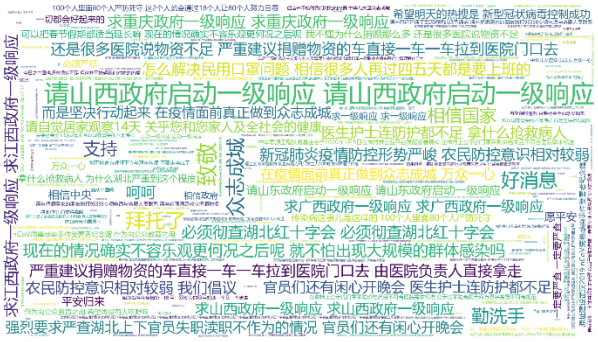
见：第三部分 数据分析

- 通过 python 的 wordcloud 库绘制各阶段心态词词云\句云(只展示部分图片，其他阶段和形式的词云图：https://github.com/wsfyp/data_science_byFLW/tree/master/fan_code_data/pic)

评论词云（第一阶段）：

正文词云（第一阶段）：

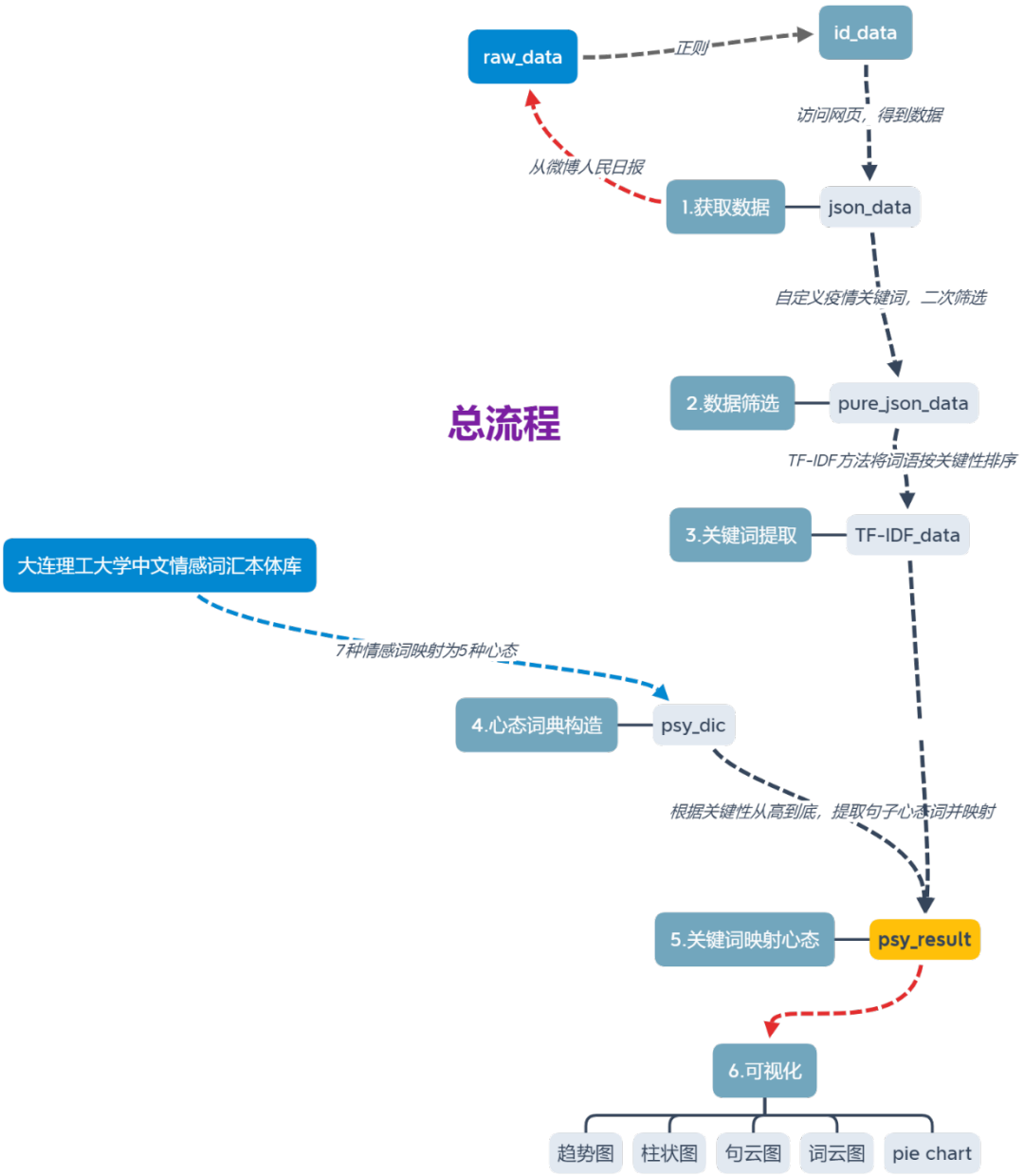
评论句云（第一阶段）：



评论句云（第四阶段）

第二阶段总流程：

包括：数据流向，处理流程



#所有数据代码及数据参见 https://github.com/wsfyp/data_science_byFLW

第三部分 数据分析

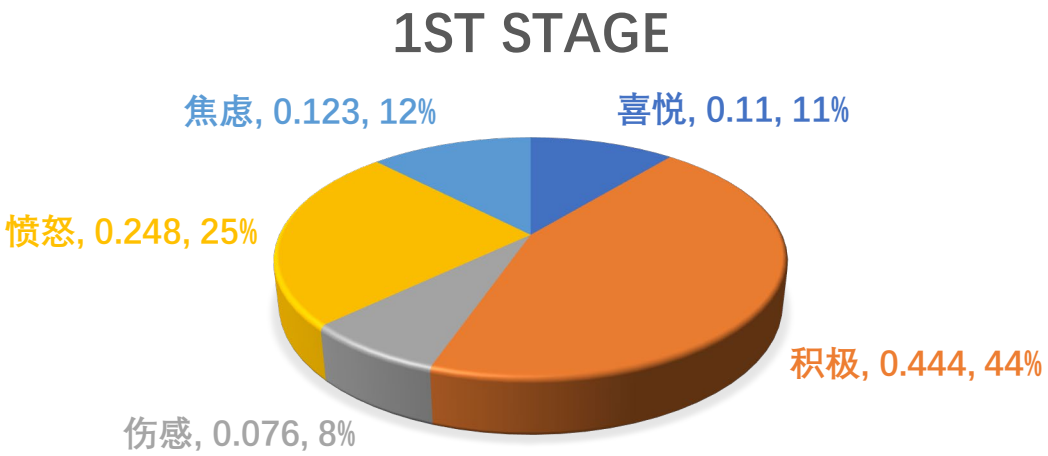
1. 心态关键词分析

对于各阶段心态占比的分析，我们采用了用关键词 hit 心态词典的方法。得到了各阶段新闻正文及评论的心态占比。第一阶段的评论中，喜悦被 hit 到 156 次，积极 630 次，愤怒 351 次，伤感 107 次，焦虑 174 次，这样就得到了第一阶段评论的心态占比。第二阶段，喜悦 512 次，积极 2198 次，愤怒 1429 次，伤感 421 次，焦虑 325 次。第三阶段，喜悦 112 次，积极 442 次，愤怒 303 次，伤感 64 次，焦虑 55 次。第四阶段喜悦 1955 次，积极 6996 次，愤怒 4041 次，伤感 1011 次，焦虑 1150 次。新闻正文各阶段的数据是：第一阶段，喜悦 11 次，积极 65 次，愤怒 26 次，伤感 4 次，焦虑 13 次。第二阶段，喜悦 53 次，积极 267 次，愤怒 74 次，伤感 25 次，焦虑 34 次。第三阶段，喜悦 14 次，积极 56 次，愤怒 15 次，伤感 5 次，焦虑 6 次。第四阶段，喜悦 161 次，积极 797 次，愤怒 258 次，伤感 91 次，焦虑 123 次。这些数据被 hit 后完整的过程及结果放在了 GitHub 中，地址如下：https://github.com/wsfyp/data_science_byFLW

2. 分阶段评论数据分析

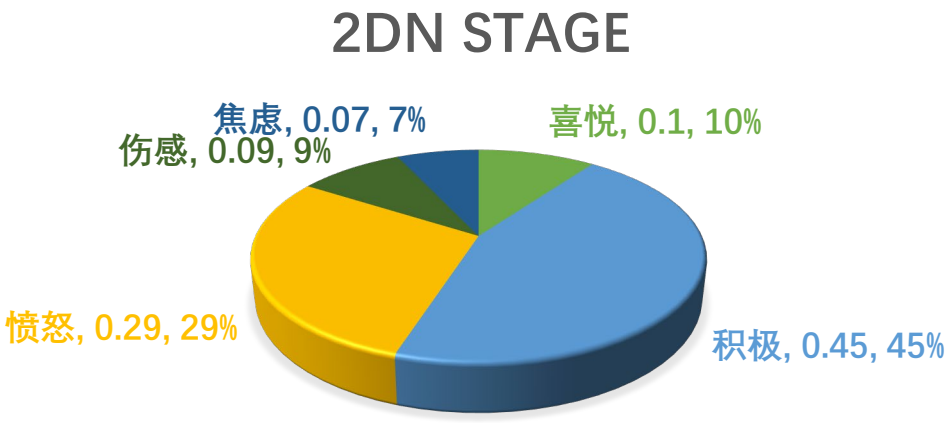
基于已经建立的五类心态的前提下，利用各阶段获取到的评论数据所提取出的关键词来 hit 已经建立的心态词典，将获取的能有效刻画心态的数据作百分比处理，得到五大心态在不同阶段所占比例及变化的过程：

第一阶段(2019.12.8-2020.1.22)：（标志事件：2019.12.8 发现首位肺炎患者，2020.1.22 湖北启动公共卫生事件二级响应）



本阶段属于疫情初发期，在这一阶段疫情刚开始蔓延，并没有受到广泛的关注。所以疫情初期公众的心态并没有因为疫情而发生显著的变化，而到 1.20 日前后疫情彻底爆发，公众的负面情绪开始出现，焦虑和愤怒占据负面心态的主体，但在接近半数的评论中能看到公众积极的一面。说明在面对突发紧急事件时，大部分人能保持冷静积极 的心态，但出现负面的心态是无法避免的。疫情的快速蔓延，以及疫情爆发原因的查明给公众的心态造成了很大的消极影响，焦虑和愤怒的情绪就因此而生了。从饼状图来看，公众的心态仍然较为乐观，正面心态占到 55%，伤感占负面心态的比例较小，初期还未出现死亡病例，焦虑所占比例较高，这是人们对于未知的可能以及危险所产生的自然的心态。而愤怒的出现以及较高的比例是因为疫情扩散速度过快且在初期没有得到有效的管控。

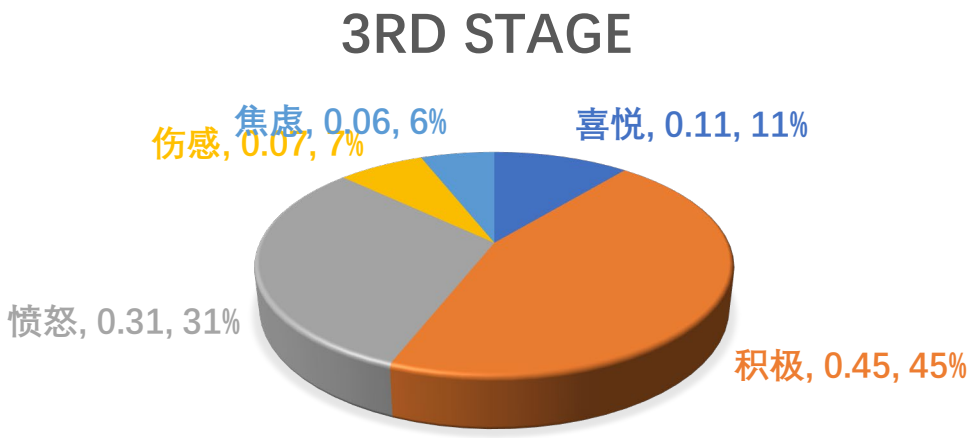
第二阶段(2020.1.23-2020.2.7)：（标志事件：2020.1.23 武汉宣布“封城”，2020.2.7 “吹哨人”李文亮去世）



本阶段属于疫情爆发期。武汉宣布封城，医疗物资匮乏，全国乃至世界各地向湖北各地市捐

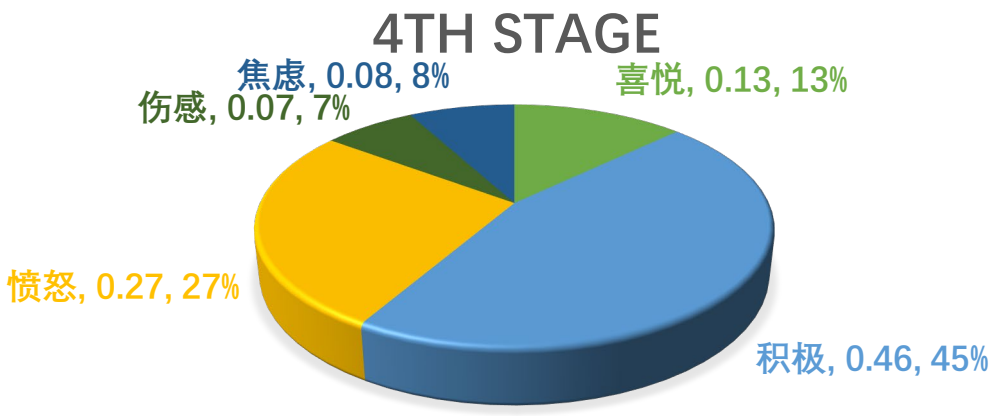
赠医疗物资。这一阶段国家开始强制管控，出动军队援助抗击疫情，使民众焦虑的心态得到缓解。武汉封城、各地的“硬核”防疫措施也在一定程度上安抚了忐忑不安的群众。但这段时期也正因为这些管控措施造成的对群众生活的不便等因素，有人不管不顾防疫要求，与工作人员发生冲突，更有甚者为一己之私，破坏、阻挠防疫工作的展开与进行。这样的令人不齿的行为与各地捐粮捐物的众志成城和团结一心形成了鲜明的对比，网络上公众愤怒的情绪蔓延开来，加诸网络以及流量效应，负面情绪传播的范围更广、程度更深、激起的民愤更大。在新闻以及舆论上造成了很大的负面影响。愤怒的心态比起第一阶段上升了 4 个百分点。虽然正面心态依旧超过半数，但负面心态的转换比例也能看出疫情期间不同的新闻报导对公众的心态有着不同程度的影响。

第三阶段(2020.2.10-2020.2.13)：（标志事件：2020.2.10 19 省对口支援湖北省除武汉外 16 个市州及县级市，2020.2.13 湖北省相关领导变动）



这个阶段是疫情的严格统一管控和物资配给阶段。各省医疗队驰援湖北，医疗资源统一调配，雷神山、火神山医院投入使用，全民抗击疫情。虽然确诊病例不断增加，但国家及时有力的举措让公众焦虑的情绪平静下来，焦虑所占比例继续下降，正面心态比例再次上升，公众对于打赢这场抗击疫情的战争有了极大的信心和决心。但在此期间，医患关系成了新的矛盾激发点，患者在绝望下精神出现问题，做出了过激的行为。新闻报道一出便引起全网公愤，有更多的不配合隔离，阻碍防疫工作的情况出现。一边是医护人员无私奉献地工作，另一边是只顾自己，破坏防疫工作的嚣张言行。不禁让公众的愤怒再次爆发，相较于第二阶段再次上升，达到了 31%。

第四阶段(2020.3.10-6.24)：（标志事件：各省开始有序复工）

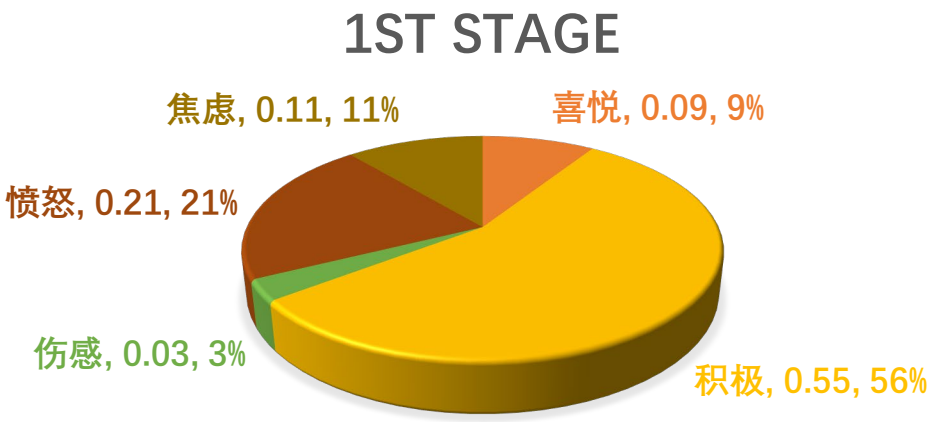


这一阶段疫情已经得到有效控制，各省根据疫情状况开始有序复产复工。疫情的好转使得公众的愤怒情绪逐渐平缓，下降了近 4 个百分点，同时一些私营工商业者终于迎来了疫情期间的转折，复产复工使得他们得以重新投入工作获得收入，结束了一段时期以来收入拮据的窘境。人们终于能够自由出门，不再需要居家隔离，各地的公共设施及场所也逐步开放。疫情居家隔离期间的各种被压抑的需求得以释放，这一阶段正面情绪占比接近 60%。喜悦的心态占比上升符合疫情得到控制有序复工复产的事实。4.8 号武汉解封更是标志着我国新冠疫情度过了最艰难的时期，宣告着我国抗疫已经取得了重大胜利，而这种喜悦也感染了无数人，国外疫情的大爆发与国内复产复工的安全环境的显著对比，使得网络舆情一片向好。

3. 分阶段新闻数据分析

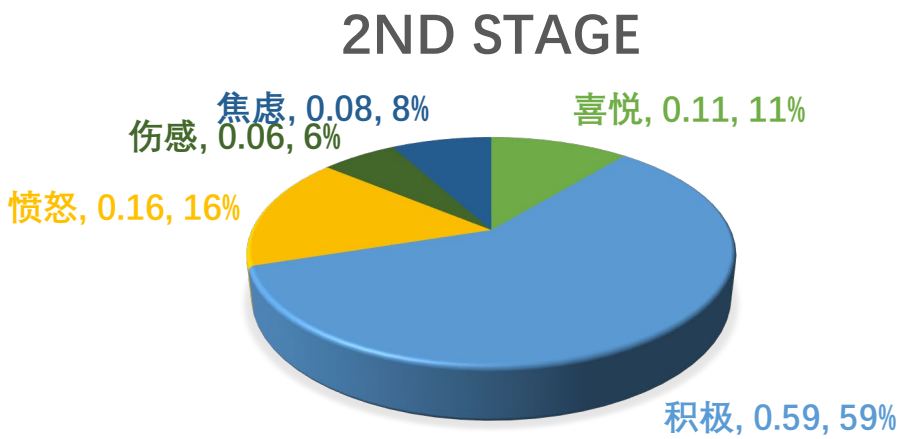
基于已经建立的五类心态的前提下，利用各阶段获取到的新闻数据所提取出的关键词来 hit 已经建立的心态词典，将获取的能有效刻画心态的数据作百分比处理，得到五大心态在不同阶段所占比例及变化的过程：

第一阶段(2019.12.8-2020.1.22)：（标志事件：2019.12.8 发现首位肺炎患者，2020.1.22 湖北启动公共卫生事件二级响应）



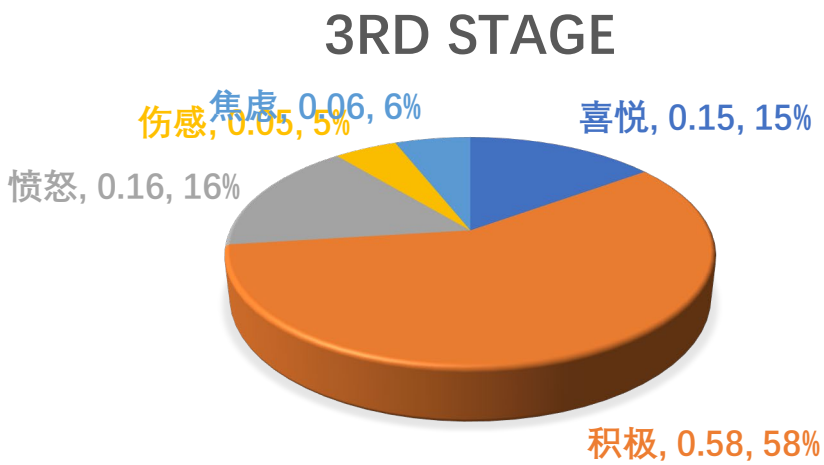
本阶段属于疫情初发期，在这一阶段疫情刚开始蔓延，并没有受到广泛的关注。疫情初期的新闻报道更多的是为了避免恐慌，积极心态占比相当高，是为了引导公众情绪，而愤怒心态占比较高与疫情初期扩散速度过快且没有得到有效的管控。

第二阶段(2020.1.23-2020.2.7)：（标志事件：2020.1.23 武汉宣布“封城”，2020.2.7 “吹哨人”李文亮去世）



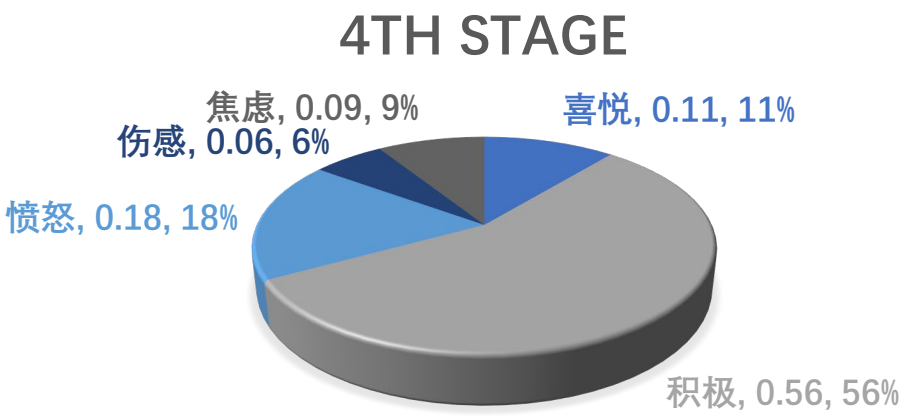
本阶段属于疫情爆发期。武汉宣布封城，医疗物资匮乏，全国乃至世界各地向湖北各地市捐赠医疗物资。这一阶段的新闻报道主要以赞扬捐赠物资，展现疫区情况，传达党和国家政策为主，仍然是以舒缓公民情绪作为主要目的,传播积极向上的正能量。愤怒心态的出现是因为一些特殊的新闻报道，各地出现了一些为一己之私，破坏、阻挠防疫工作的展开与进行的人。在这种特殊的时期，这样的负面新闻出现，在舆论上造成很大的负面影响，对于公众心态也有巨大的影响。

第三阶段(2020.2.10-2020.2.13)：（标志事件：2020.2.10 19 省对口支援湖北省除武汉外 16 个市州及县级市，2020.2.13 湖北省相关领导变动）



这个阶段是疫情的严格统一管控和物资配给阶段。各省医疗队驰援湖北，医疗资源统一调配，雷神山、火神山医院投入使用，全民抗击疫情。这一阶段的新闻主要报道了医护人员的不顾安危，不辞辛劳，为打赢这场防疫战付出了十二分的努力。这样的报道不仅仅让我们感动，更让我们备受鼓舞，知道在这种特殊的时候有这样一群人为了国家和人民英勇奉献，积极乐观的心态感染了千千万万的中国人。

第四阶段(2020.3.10-6.24)：（标志事件：各省开始有序复工）

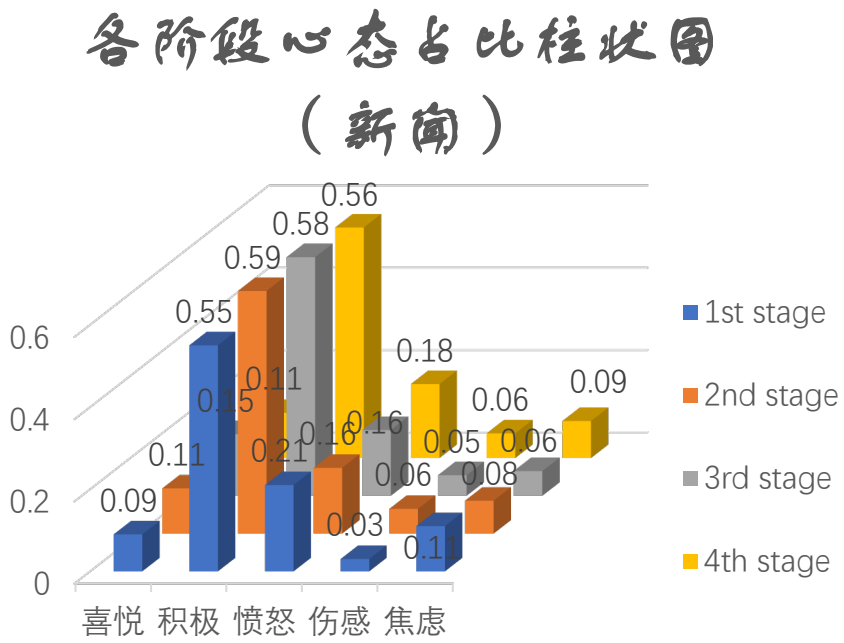


这一阶段疫情已经得到有效控制，各省根据疫情状况开始有序复产复工。报道的重心转向了复工复产，人们终于能够自由出门，不再需要居家隔离，各地的公共设施及场所也逐步开放。这一阶段新闻中的正面心态占比达到了 67%，符合疫情得到控制有序复工复产的事实。4.8 武汉解封的新闻更是让无数人为之感到由衷的喜悦。这阶段的新闻让公众们得到了有效的信息，更为复工复产打下

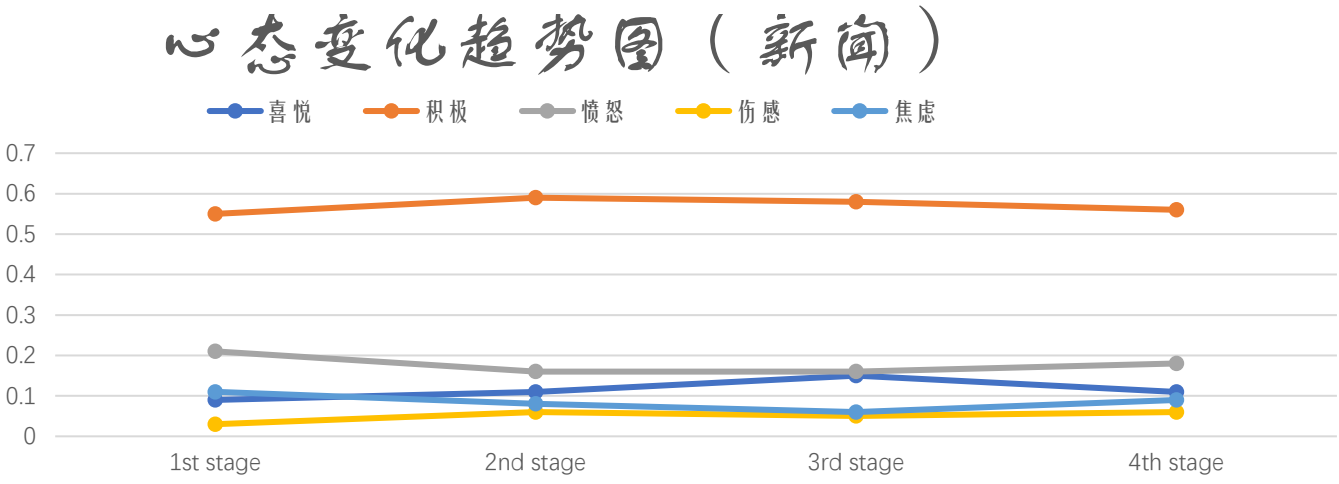
了坚实的心理基础。

4. 总体趋势分析

我们把从新闻中所得的数据进行了总体的分析，并绘制了图表。从**各阶段心态占比柱状图（新闻）**可以看出正面的心态占比始终是大于负面心态的，积极的心态占比始终在 50%以上，这说明新闻的目的是把公众的情绪引导向积极的方向。

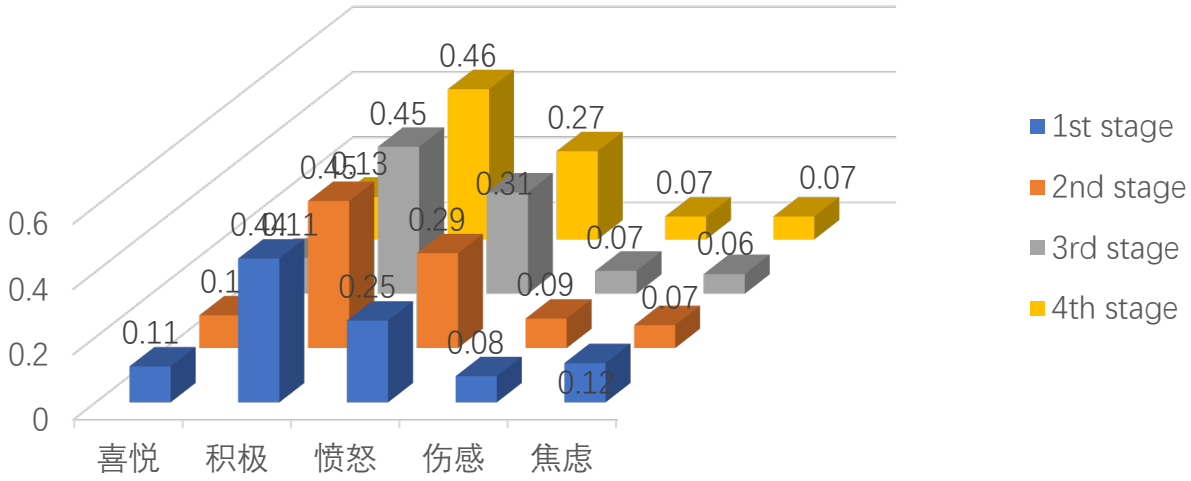


在新闻中负面心态占比非常小，焦虑、伤感的心态占比低至 10%以下。从总体趋势来看，正面心态在第三阶段达到顶峰，达到了 73%，这一阶段正是抗击疫情的攻坚时刻，新闻的鼓舞作用就体现出来了，这也为抗疫成功做出了不小的贡献。



再看**心态变化趋势图（新闻）**，这张图反映了在新闻中所能体现出的心态变化幅度并不大，保持在一个较稳定的状态，符合新闻的真实性。

各阶段心态占比柱状图（评论）



再看分析评论得到的数据-**各阶段心态占比图（评论）**，从总体来看正面心态占比随着阶段的推进在逐步上升，这表明虽然疫情的爆发及恶化让公众产生了负面的心态，甚至占比接近 50%，但大部分人仍然保持着较为积极的心态，对于新冠疫情的爆发以及国家采取的措施保持着乐观的态度。这无疑为抗击疫情起到了正面的影响。再看负面心态的占比，随着阶段内重大事件的发生与一些令人心痛或愤怒的新闻报道出现，以及前三个阶段每日上涨的确诊病例和死亡人数，焦虑伤感等心态始终保持在一定的水平线上，而愤怒心态更是在前三个阶段有逐步上升的趋势，这样的变化也源于在此期间内，公众的网络心态是更易被来自网络的信息所影响，层出不穷的违反抗疫规定，不顾群众利益的新闻出现，让公众们的心态向着愤怒更进一步。

根据绘制的**心态变化趋势图（评论）[见 18 页]**，我们可以清楚的看到积极和喜悦两个正面心态在总体上是上升的，而焦虑这一心态在经历了三个阶段的下降后，在最后一阶段反而有所上升，这可能是因为受疫情影响，复工复产并没有想象中那么简单，焦虑的心态也就随之而来。

将两部分数据放在一起对比可以发现，评论中所反映的真实的公众心态受到了新闻中体现出心态的影响。公众能保持较高的正面心态与新闻报道的正面引导是分不开的。虽然在新闻评论中所展现的负面心态要远大于新闻本体，但也在可控的范围内，并没有超过总心态占比的 50%。