2020-12-25~2021-1-12

数据获取

在此次新型冠状病毒（COVID-19）传播这一重大公共卫生事件情境下，为了深描中国大众的网络社会心态变化，分析从疫情最初不被重视，到逐渐扩散、爆发，再到全国上下联防联控、万众一心，新冠肺炎疫情得到有效控制，最后全国逐渐复工复产、经济逐渐恢复这一过程中中国大众网络社会心态的体现和分布情况，我们决定选择获取从2019年12月中旬至2020年6月的官方新闻文本，以及从2020年1月到6月初的各类疫情新闻的“热评”。除此之外，我们注意到：在2020年12月中旬至今（2021年1月19日左右），北方疫情出现了较为严重的反弹，包括河北藁城、北京大兴、黑龙江黑河、绥化望奎、惠七等地均出现了大量的确诊病例和无症状感染者，在全国逐渐基本回到了正常生活轨道的时刻，疫情突然再度反弹，无疑会对大众心态产生较大的冲击，因此我们额外选取了2020年12月中旬至2021年1月中旬的微博及“热评”，分析在疫情回弹的形势下，中国大众网络社会心态又会发生什么样的变化。

为了获取最真实、最能反映大众心态的新闻评论，我们斟酌再三，最终决定选择获取新浪微博相关正文内容及热评，原因及相关做法如下：

**一、主要原因**

如今，新浪微博是中国网民的重要活动平台，微博上的许多“大V用户”拥有数以千万计甚至上亿计的粉丝量，其发布的微博具有权威性和引导性，其微博下的评论也可以反映出中国大众的主流网络社会心态。另外，微博平台不似其他众多平台，在微博评论中，网友可以不拘于条框，将自己内心真实想法用“评论”的方式表达出来，而微博上给评论“点赞”的功能则能够反映其他网友对于这条评论的认可程度，点赞数较多的前几条（约5~8条）评论将成为“热评”，而使用“热评”作数据的好处有：1、评论内容真实可信，可有效避免“机器粉”刷屏；2、评论内容为多数网友所赞同；3、减少了后期筛选数据的工作量，免去了数据“拟合”的过滤筛查，我们可以直接使用大量“热评”数据进行分词和提取。在本次作业中，我们爬取了热点新闻下相对应的“热评”，这些评论不仅点赞数最多，且反映出来的网络社会心态也最真实；

**二、URL选择**

网页版微博（weibo.com）可以方便地获取对应时间日期段的微博内容（可以直接指定月份爬取），可以减小后期从大量数据中再划分数据的任务量，相应地最难获取；手机端微博（m.weibo.com）数据爬取难度相比前者更小，但无法指定对应日期的微博内容，后期数据处理时需要另外进行事件阶段性划分；移动端微博（weibo.cn）最易爬取，但数据最为粗糙，后期处理时更加需要精细的筛选。综上，我们最终选择爬取网页版微博的数据；

**三、数据源选择**

权威性的重点新闻及其下的评论数据是我们爬取的目标，何谓“重点新闻”呢？通俗理解就是指评论量、点赞量及转发量等远超其他新闻的新闻。我们横向对比了新浪微博上的“大V”新闻官煤，包括：央视新闻、人民日报、人民网、荔枝网、新华视点、中国青年报等，通过对每家媒体的微博内容进行抽样调查，发现只有央视新闻和人民日报满足这一要求，我们分别对人民日报和央视新闻的微博点赞数和评论数进行了四次抽样，每次随机抽取100条微博内容，并获取其下的评论数和点赞数，得到如下图的结果：

图表

已生成极高可信度的说明

（央视新闻微博点赞、评论抽样分布图1）

图表, 直方图

已生成极高可信度的说明

（人民日报微博点赞、评论抽样分布图2）

图表

已生成极高可信度的说明

（央视新闻微博点赞、评论抽样分布图3）

图表, 直方图

已生成极高可信度的说明

（人民日报微博点赞、评论抽样分布图4）

除个别微博热度过高，评论和点赞数异常高之外，我们发现：获得这两家媒体平均每条微博的评论量可以达到4000~8000，平均点赞量可以达到8000~20000，且这一水平较为稳定。而对于热点微博事件评论量可以达到超过10000，而点赞量更可以达到数十万，如图中“红线”的高凸处。而对比其他媒体的只有1000~2000的点赞量和几百评论量来说，央视新闻、人民日报发布的微博可以称为“重点新闻”，而对于如此庞大的数据量，我们有充分的理由认为，央视新闻和人民日报这两个新浪“大V用户”的新闻内容及相关热评可以反映中国大众网络社会心态；

**四、数据爬取**

虽然网页版微博（weibo.com）可以方便地获取对应时间日期段的微博内容，但是是三者中最难获取的，由于微博反爬机制的存在，我们最终采用“模拟人浏览网页”的方案进行网页源码获取，即采用Python并引入Selenium库，通过半自动化操作来获取动态加载的网页，而后再对网页源代码进行解析、正则提取，最后获得了相关微博的正文、点赞数、转发数、评论数，并爬取了所有的“热评”，最终累计获取微博19000余条，微博“热评”共1210000余条。除微博外，我们在其他平台爬取的纯新闻数据共30000余条。我们接下来的任务就是对这些数据进行进一步的筛选和提取。

2021-1-13~2021-1-21

事件阶段划分

在获取了大量数据后，我们对这些数据进行了第一次筛选——先进行一次粗筛选，先根据如下事件阶段标准对数据进行划分：

**第一阶段：**2019年12月末——2020年1月下旬，疫情从刚显苗头、未得重视到全国人民逐渐意识到新冠肺炎疫情的严重性（由于2019年12月中旬以前关于疫情的报道相当少，因此并未选择获取那时的数据）；

**第二阶段：**2020年1月末——2月上旬——由于人民逐渐意识到疫情严重和逐渐主动居家、不走亲访友，但随之而来的便是物资匮乏、物价上涨，此时好多网友纷纷表示物资“太贵了”、“买不到”，食品、用品等价格上涨；

**第三阶段：**2020年2月中旬至下旬，政府严格统一管控、物资配给，使得居民生活得到了保障，人民逐渐从恐慌到安心，全国上下有序联防联控；

**第四阶段：**2020年3月以后，各省开始有序复工复产……进一步划分，我们发现，在2020年3月之中，我国逐渐聚焦于复工复产的稳健进行、各地防控工作的稳定开展，我们迎来了逆行天使和抗疫英雄们的凯旋，各省各地欢迎医护人员们“回家”……

**第五阶段：**2020年4月~5月，我国疫情形势已得到了稳定良好的控制，而国外疫情情势逐渐变得十分严峻，全世界范围内病魔肆虐，中国境外除澳门均为高风险地区，我国人民开始对外国疫情表示出了担忧等情绪……2020年下半年，我国疫情防控工作一直顺利进行，各地偶有病例也迅速隔离、追根溯源，疫情一直得到了有效稳定的控制；

**第六阶段（额外分析）：**入冬以后，从2020年12月中旬开始至今，中国北方（如河北、吉林、黑龙江、北京等地）疫情由于聚集性活动和低温等原因严重反弹，最严重时日新增确诊病例可达近100例，而这正是大学生即将进入寒假、新春佳节将至的时刻，网友们对能否回家过年、大学生能否正常返乡表现出了担忧，全国上下为北方防疫工作加油鼓劲……

关键词提取

为了更加细致地分析中国网络大众在疫情防控至复工复产再到疫情反弹这一过程中的心态趋势，我们需要从他们的大量评论中提取关键词和情绪核心词，相关提取方法如下：

**一、分词**

由于获取到的“热评”均为长度不等的句子，因此首先需要粗略地将这些句子划分为词语，我们引入了一个强力的工具——jieba（“结巴”）分词库（https://github.com/fxsjy/jieba/），该库的主要功能是将长句切分成词语，同时用户可以创建自己的新词词典，以便分词更加精确。此外，jieba库包含基于TF-IDF算法的关键词抽取功能，但该函数只能传入待提取文本，也就是说此函数运作时是先分词，再基于TF-IDF算法进行词频分析和逆文本频率指数分析，最终得出Top-K个关键词，因此如果使用该函数，我们只会得到最终分析出来的关键词，其分析过程中的分词成果对于我们是透明的，因此最终我们选择只使用jieba库的分词功能，以便更清晰地观察特定词语在评论中出现的情况，后续再自己写词频统计和逆文本频率指数分析的相关代码并作图。

在初次分词的过程中，我们发现：句子中存在的大量中文标点（如，。、【】等）、大量无意义单字（如的、啊、以等许多单字）、无法反映网络心态的词汇（好像、看着、身边、网络等）对中国大众网络社会心态分析无益，因此我们在原有停词表模板（近3000条停用词）基础上通过一遍遍筛选过滤，添加了近200个新的停用词，这样即保证了文本数据的有效性，又不至于将文本数据过滤得过于干净而丧失有效数据；与此同时，我们发现jieba库虽然有新词识别算法，但是对于疫情期间出现的特定词汇如“抗疫”、“火神山”、“雷神山”、“境外输入”、“新型冠状病毒”、“输入病例”、“一级响应”、“逆行天使”、“核酸检测”等不能进行很好的识别和分词，为此，我们建立了新词词典并添加了200余条疫情期间网友们常用的新词和特定词汇，在分词时便能更好地屏蔽无意义符号和词语，分出和收录更多能切实反映大众心态的词汇了，同时我们还在新词表中添加了中国地名共400余个。

**二、关键词提取**

接下来我们对划分好的词汇进行关键词提取。

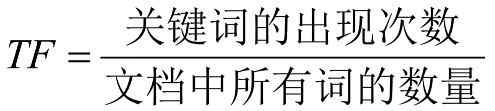
在剔除近3000个停用词及引入新词词典后，我们将61000余条“热评”语句划分成了对心态分析有利的词汇，我们引入Python中collections库中的Counter类，向Counter类中传入列表并将其传入dict()方法中，即可自动生成一个“关键词-频数”的字典，对于每个关键词，Counter类都会自动计算出其在所有词汇中出现的总次数，而后再调用sorted()函数进行排序，我们就得到了逆序排列的“热评”高频词表了。

下一步我们选择自己实现词频(TF: Term Frequency)和逆文本频率指数(IDF: Inverse Document Frequency)的分析过程，以便更好地了解关键词在各个月份的疫情微博“热评”中的分布情况。

TF-IDF是一种统计方法，用来评估词语对于一个文件集或一个语料库中一份文件的重要程度，其在文件中出现次数越高，可以初步认为该词语重要性越大，但同时会随着它在语料库中出现的频率成反比下降。

词频（TF）是指某一个给定的词语在文件中出现的频率，逆文本频率指数（IDF）是一个词语普遍重要性的度量，而由于特定词语可能在文件中一次都没出现过，因此实际计算时应将分母加1。但不能单纯地认为词频越高，重要性越大，因为语料库中某一文件可能会反复出现该词语而使TF值偏高，因此我们引入TF-IDF值来对词语重要性进行考量：一个词语的TF-IDF值即为一个文件或语料库中该词语的TF值与IDF值的乘积，某一词语越重要，其在该文件内即具有高词语频率，在整个文件集合中即具有低文件频率，这样便会产生高权重的TF-IDF词语，进而筛选出我们需要的“关键词”。非常值得注意的是，我们之所以要不断完善停词表（stop words），就是因为类似“的”、“以”之类的无意义词可能会占据总词频的大部头，而它们对项目分析几乎无用，所以应将其忽略掉。

根据TF和IDF的计算公式（IDF计算时分母加1），我们取某一阶段的所有评论语句构成该阶段的完整语料库，认为每一条热评构成一个文件，在计算IDF时，需要判断某词语是否在某条“热评”（即语料库中的文件）中出现。

 图片包含 示意图

已生成极高可信度的说明

首先我们统计、计算、绘制并观察各个阶段“热评”词汇的Term Count（词数）、Term Frequency（TF：词频）、Inverse Document Frequency（IDF：逆文本频率指数）和TF-IDF综合值的条形图和折线图。

图表

已生成极高可信度的说明

（一阶段词语TF（条形）及IDF（折线）统计图）

图表, 直方图

已生成极高可信度的说明

（一阶段词语词数（条形）及TF-IDF（折线）统计图）

图表

已生成极高可信度的说明

（二阶段词语TF（条形）及IDF（折线）统计图）

图表, 直方图

已生成极高可信度的说明

（二阶段词语词数（条形）及TF-IDF（折线）统计图）

图表, 折线图

已生成极高可信度的说明

（三阶段词语TF（条形）及IDF（折线）统计图）

图表, 直方图

已生成极高可信度的说明

（三阶段词语词数（条形）及TF-IDF（折线）统计图）

图表

已生成极高可信度的说明

（四阶段词语TF（条形）及IDF（折线）统计图）

图表, 直方图

已生成极高可信度的说明

（四阶段词语词数（条形）及TF-IDF（折线）统计图）

图表, 折线图

已生成极高可信度的说明（五阶段词语TF（条形）及IDF（折线）统计图）

图表, 直方图

已生成极高可信度的说明（五阶段词语词数（条形）及TF-IDF（折线）统计图）

图片包含 物体, 游戏机, 华美, 一群

已生成极高可信度的说明

（六阶段词语TF（条形）及IDF（折线）统计图）

图表

已生成极高可信度的说明

（六阶段词语词数（条形）及TF-IDF（折线）统计图）

通过统计图，我们非常直观地看到，并不是TF越大的词语，其TF-IDF值就越高，因为还会有IDF值的制约。通过对每个阶段TF & IDF统计图和Term Count & TF-IDF统计图的综合分析，选取每个阶段TF-IDF权重前100的词汇，我们就得到了每个阶段前100个关键词——

图片包含 表格

已生成极高可信度的说明

图片包含 表格

已生成极高可信度的说明

图片包含 表格

已生成极高可信度的说明

表格

已生成极高可信度的说明

**三、心态分析**

**（一）心态词典建立与关键词碰撞计划方案**

得到了每个阶段的关键词，我们就可以对这些关键词进行心态分析了。

我们首先将心态词所表达的心态归一化，即将心态具体量化，映射到(0.00, 1.00)区间上，对每个心态赋一个“心态值”，即建立一个“心态词(attitude)→心态值(value)”的映射。具体做法是：

以“平静”（中立）心态为中线，赋值为0.50；

“心态值”小于0.50的心态词被认定为“负面心态”，包括：绝望 0.00，愤怒 0.05，恐惧 0.10，厌恶&嘲讽 0.15，悲伤 0.20，担忧 0.25，焦虑 0.30，慌张 0.35，惊讶 0.40，迷茫 0.45。其中越接近0.00，表示负面心态越严重；

“心态值”大于0.50的心态词被认定为是“正面心态”，包括：积极 0.55，鼓励 0.60，祝愿 0.65，敬佩 0.70，感激 0.75，乐观 0.80，期待 0.85，渴望 0.90，喜悦 0.95，热爱 1.00。其中越接近1.00，表示正面心态越积极。

而后我们建立“心态词典”，即建立一个“核心词(key)→心态值(value)”的映射，通过对各阶段评论内容的浏览和分析，并借助提取后的逆序词频表，我们人工建立了一个600余条的“情绪/关键词→心态值”的心态词典。额外地，我们向心态词典中加入了各地地名，经过大量评论分析后我们认为，评论中出现各地地名可以经典地反映出网友对这些地疫情的“焦虑 0.30”、“担忧0.25”和“鼓励0.60”的心态。我们将之前得到的各个阶段的关键词与心态词典进行碰撞，若某关键词与心态词典的键重合，则可由“键”对应的“值”找到该关键词对应心态词，进而根据该关键词的词频数（Term Count）得到对应心态词的词频数，最后即可得到每个碰撞成功的关键词对应心态词的词频数，进而分析各阶段中国大众的网络社会心态。

**（二）关键词与心态词典的碰撞**

接下来我们按照上述方案进行关键词向心态词典的碰撞，并得到各个阶段每个心态的分布情况（横轴为心态词，纵轴为心态频数（条形）和心态词频率（折线））：

图表, 直方图

已生成极高可信度的说明

（一阶段）

图表, 直方图

已生成极高可信度的说明

（二阶段）

图表, 直方图

已生成极高可信度的说明

（三阶段）

图表, 直方图

已生成极高可信度的说明

（四阶段）

图表, 条形图, 直方图

已生成极高可信度的说明

（五阶段）

图表, 直方图

已生成极高可信度的说明

（六阶段）

通过对六个阶段心态频数及频率统计图的观察分析，我们发现：每个阶段中，中国大众网络社会心态分布情况大致相同，即每个阶段的“心态图形”都近似呈一种“驼峰”状，而积极心态（心态值＞0.5）和消极心态（心态值＜0.5）中都有某1~2种心态在图中为“驼峰”，即为特定阶段中中国大众表现出的主流网络社会心态。初步观察来看，在每个阶段的心态图中，积极心态的“鼓励”和“敬佩”占比较高，而消极心态的“担忧”和“焦虑”占比较高，由此我们初步认为：疫情背景下，中国大众网络社会心态主要表现为“鼓励”、“敬佩”、“担忧”和“焦虑”。下面我们具体分析在整个六个阶段中，每种心态占比的变化情况。

我们观察各个阶段积极心态（心态值＞0.5）的占比及变化。

（插入图片）

我们观察各个阶段消极心态（心态值＜0.5）的占比及变化。

（插入图片）