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。

思路方法：在刻画的此次新型冠状病毒（ COVID-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。

实验人员及分工：

樊言鹏，组长，数据爬取，筛选，关键词提取，心态词典建立，心态映射，获得分阶段心态数据，数据可视化的校正、实验报告

刘晋元，组员，参与心态词典建立，数据可视化和实验报告阶段数据分析

王子岑，组员，参与关键词提取，参与心态词典初步构建，初步进行数据可视化、实验报告目录撰写

（按姓氏排名）

数据源选择：

1. 微博地址、微博发布时间、标题、正文、赞数、评论数、转发数、评论。来源于 微博账号-人民日报：

<https://m.weibo.cn/u/2803301701?uid=2803301701&t=0&luicode=10000011&lfid=100103type%3D1%26q%3D%E4%BA%BA%E6%B0%91%E6%97%A5%E6%8A%A5>

1. 心态词典构建数据参考<**大连理工大学中文情感词汇本体库>**

工具选择：

数据获取分析导出阶段使用python

数据可视化阶段使用python和Microsoft制图

实验步骤：

数据爬取，

思路：我们的目标是爬取人民日报官方微博的疫情相关的新闻于评论，并进行对重点新闻、评论的拟合筛选。

微博有3个不同的访问网址：

[https://weibo.com/](https://weibo.com/、)

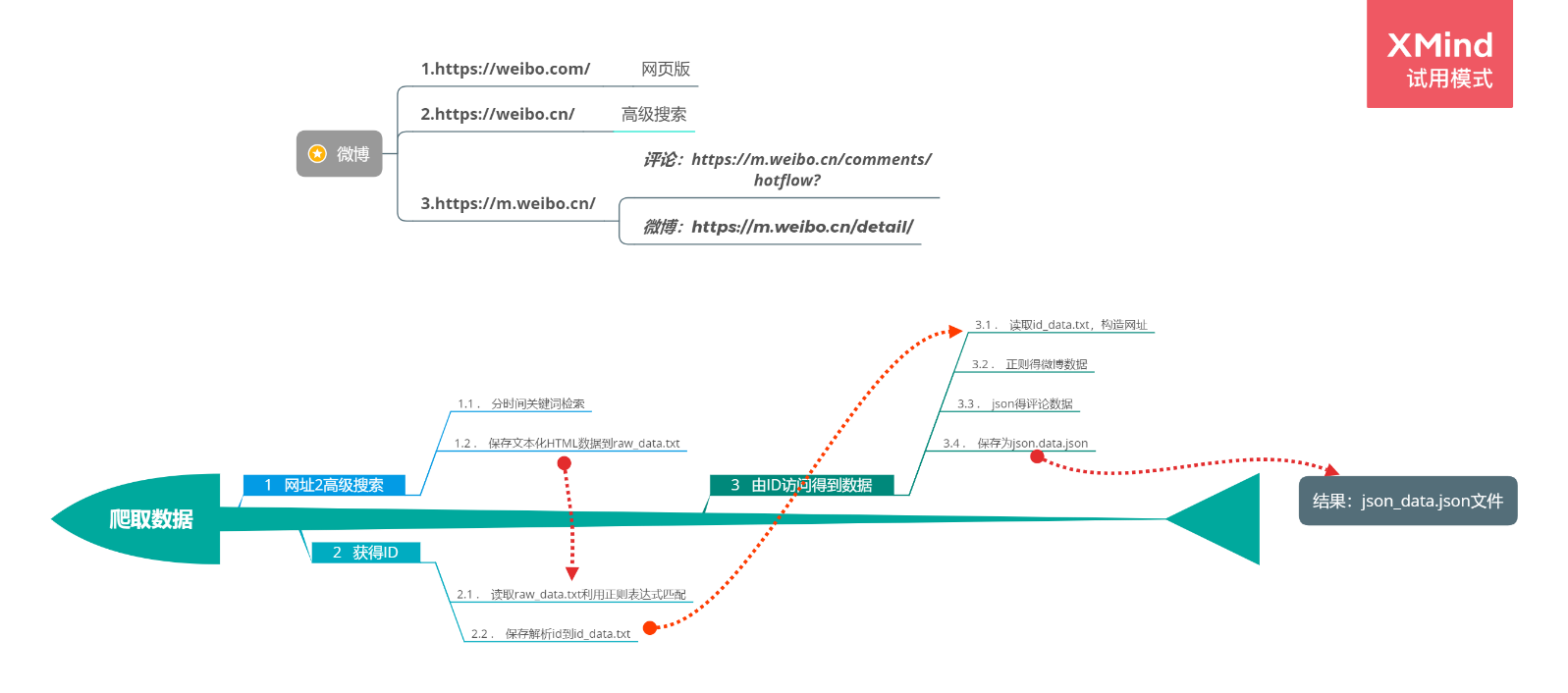
[https:// weibo.cn/](https://m.weibo.cn/%20%20%20https://m.weibo.com/)

<https://m.weibo.cn/>

思路：通过这几个网站找到人民日报的微博，通过F12监视network，找到request url，并设置cookie和**User-Agent，然后用**python的request，我发现无论是哪一个网址，在由python发起的请求下都只能访问首页的数据，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和模拟登陆可能面临微博的反机器登陆的屏障，我选择了不进行登陆，而通过指访问第一页评论数据进行爬取，这样做也有一个好处，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，选取第一页即选取了这条微博最关键的评论，因此不需要再次对评论的重要性做拟合于筛选。那么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request请求，我发现移动端可以通过前缀（https://m.weibo.cn/detail/）加上微博的ID号获得新闻html内容，其中包含此微博的所有信息，通过正则表达式可以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，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=XX&max_id=XX>,即可以通过id访问评论页面，得到一个json文件，然后通过json文件的解析和网页数据的分布，取出我们需要的第一页所有评论。只要获得ID，就能通过移动端获取数据，现在的问题就在于如何获得我们需要的微博id？[https://m.weibo.cn/](https://m.weibo.cn/   https://m.weibo.com/)的高级搜索功能每一次可以显示一页的数据，通过对数据的分析，我发现每一个html页面里面含有对移动端的引用，我获得每一页的html数据之后，将每个阶段的数据存储到txt文件，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存储为json文件。数据爬取的工作到此就结束了。

代码实现参见<https://github.com/wsfyp/wb_crawler>



数据筛选，

本次数据爬取基于[https:// weibo.cn/](https://m.weibo.cn/   https://m.weibo.com/)的高级搜索，关键词为疫情，为了进一步筛选数据，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，去掉此数据，对四阶段数据进行此操作，得到pure\_json\_data.json文件，这是关键词提取的基础文件。

关键词提取

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，那么这个句子就被映射到一种心态。

心态词典

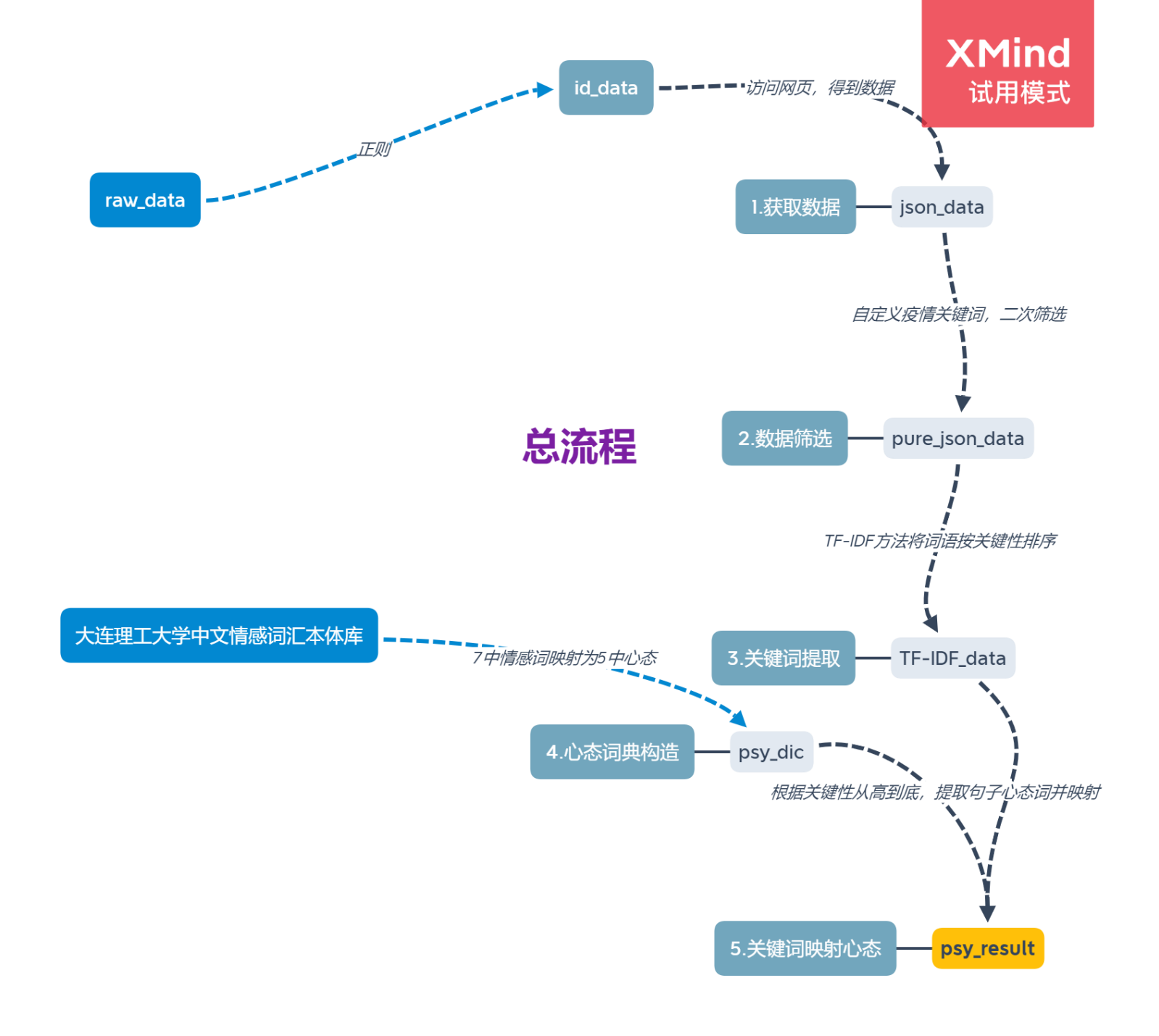
思路：构建一个心态字典，按键值将不同的情感心态词（key）映射到心态词（value），要求心态词要足够的全面，映射心态要足够地准确。我们使用了大连理工大学的中文情感词汇本体库，情感词分为7个大类，根据分析，我们决定将7大类映射为五个小类，即【喜悦，积极，愤怒，伤感，焦虑】最为此次心态分析的5大基本心态。

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。

关键词映射，

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。

具体实现：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。



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

4阶段评论柱状图：4阶段正文柱状图

4阶段评论趋势图：4阶段正文趋势图

通过python的wordcloud库绘制各阶段心态词词云

评论词云：

正文词云：