# 数据科学大作业报告

小组信息

人数：3 人

191250009_陈家志_191250009@smail.nju.edu.cn

191250008_陈家伟_191250008@smail.nju.edu.cn

191250084_林均勘_191250084@smail.nju.edu.cn

小组分工：

陈家志：爬虫+数据存取与筛选+心态词典构建

陈家伟：数据可视化+心态词典构建

林均勘：关键词提取+心态词典构建+TF-IDF+数据可视化

## 1. 研究问题：

### 计算社会学——心态分析

认识：计算社会学(computational sociology)不同于以往借助社会调查抽样数据进行描述和经典模型回归分析的定量研究，而是借助复杂模型和社会计算工具对复杂社会现象与过程进行描述、解释和预测的定量社会学新领域。作为社会学研究的一种全新范式，计算社会学研究的对象主要集中在复杂的网络现象与社会过程的联系之中。在如今社会的进程中，每个人都会在网络上留下自己的痕迹，而对于痕迹的把握，在一定程度上也可以反映某种角度下的社会结构特征。当利用计算社会学进行心态分析的时候，首先需要把握的是传统社会学如何对社会的总体心态进行把握，不难发现，传统社会学的心态分析，底层逻辑是在研究个体的言语描述，但因为传统社会学研究的能力范围有限，无法掌握大量的具体数据，所以往往更多是其抽象化后的社会舆论。但借助于计算社会学，我们可以轻易地获取每个人的网络言论，即获取了心态分析中的大量基础数据，较之传统社会学更加精准且科学化。计算社会学也有其不能及之处，由于网络平台的分化，个人言论基于网络和现实中的分歧，计算社会学——心态

分析这个方法所得到的基本数据，并不一定能完全描述社会的总体面貌，但正如前文所说，这只是提供某种角度下的社会结构特征，计算社会学不能一次解决传统社会学遗留下的问题，但其可贵之处也在于能提供新颖的角度。

出发点：本次实验中，我们也多方面获取了众多的评论数据，力求能够尽量全面地把握疫情期间社会群体的网络言论。我们从疫情期间的各种社会新闻下的网络评论出发，探寻这些大量言论下构成的社会心态总体样貌。

## 2. 开源地址：

开源地址：https://gitee.com/chen_jia_zhi/data-science-homework

**对应关系**

Client、ReptireStrategy ： 网络爬虫

Analyzer

— analyzer ： 用于给评论打标签以及将正文与词典碰撞得到情绪向量

— dailyAnalyzer ： 用于分析每日情绪占比以及抽取每日主要情绪

— Optimizer ： 拟合分布-数据筛选（）

— provincialAnalysis：根据关键评论分析每一个省份的情绪占比，以及抽取各阶段情绪情况

— TF-IDFanalyzer ： 基于 TF-IDF 关键词提取

— KeyWordExtracter ： 词典修正

data

— 原始数据（数据库）的 JSON、CSV 导出

— 包括未经筛选的 30W 原始评论集， 8K 原始新闻集供核验

Map —— 数据可视化

— dailyEmo：将每天的代表情绪占总比制作成饼图

— EmoMap：将心态在各个省级行政区的分布制作成地图形式（包括深浅的实现）

— lineChart：将每天的各类情绪权重绘制成折线图

— wordCloudtest：将情绪词制作成词云

Else —— 辅助部分（如 JSON 格式文件读写方法、数据库的存取接口）

character-master——供参考的停用词表，没有直接使用上

demo

—jsondemo:用于解析 json 文件的工具类

—scipydemo：

util：项目的阑尾

—

**实现逻辑**

1. Client 启动爬虫部分（引用 ReptireStrategy）爬取新闻及其评论，截取评论的点赞数和新闻的评论总数借助 databaseHelper 存入数据库

【但因 2021 年后新浪新闻界面改版故现在无法从 Client 正常实现，从新闻 url 获取评论的 commentSelector 仍正常工作】

2. 数据库内简单处理后导出为 csv/json 格式存放于 data 文件夹
3. Analyzer-optimizer 读取 data 内数据进行拟合获取上分位点
4. 借由上分位点从数据库获取筛选后的数据存放于 data 文件夹
5. 评论分词后进行词典碰撞，得到心态（评论心态分析），也就是给每一个评论数据打上标签
6. KeyWordExtracter 计算 hit 次数来筛除低频词
7. 人工补充疫情限定词汇，并通过 TF-idf 回填关联度
8. 根据前阶段打好标签的心态数据从省份，时间，不同情绪等过个角度分析得到结果
9. 将结果进行数据可视化直观体现

1. 用数据中的地区与心态地图中的各省级行政区匹配，并对应设置好的相应颜色或深浅

## 3. 研究方法（数据分析方法）

（1） 数据获取

- 数据集：新浪新闻 API 2020.1.1-2020.6.1 日以“新冠/肺炎”为关键词的新闻及其评论（—30W 原始评论 8K 原始新闻）
- 爬虫：基于 BeautifulSoup 的网页静态解析+基于 Request 头请求与 JSON 解析的网页动态解析
- 数据存储：mysql

（2）数据筛选：实现数据集从 30W 到 1W 的**精确化重点化**

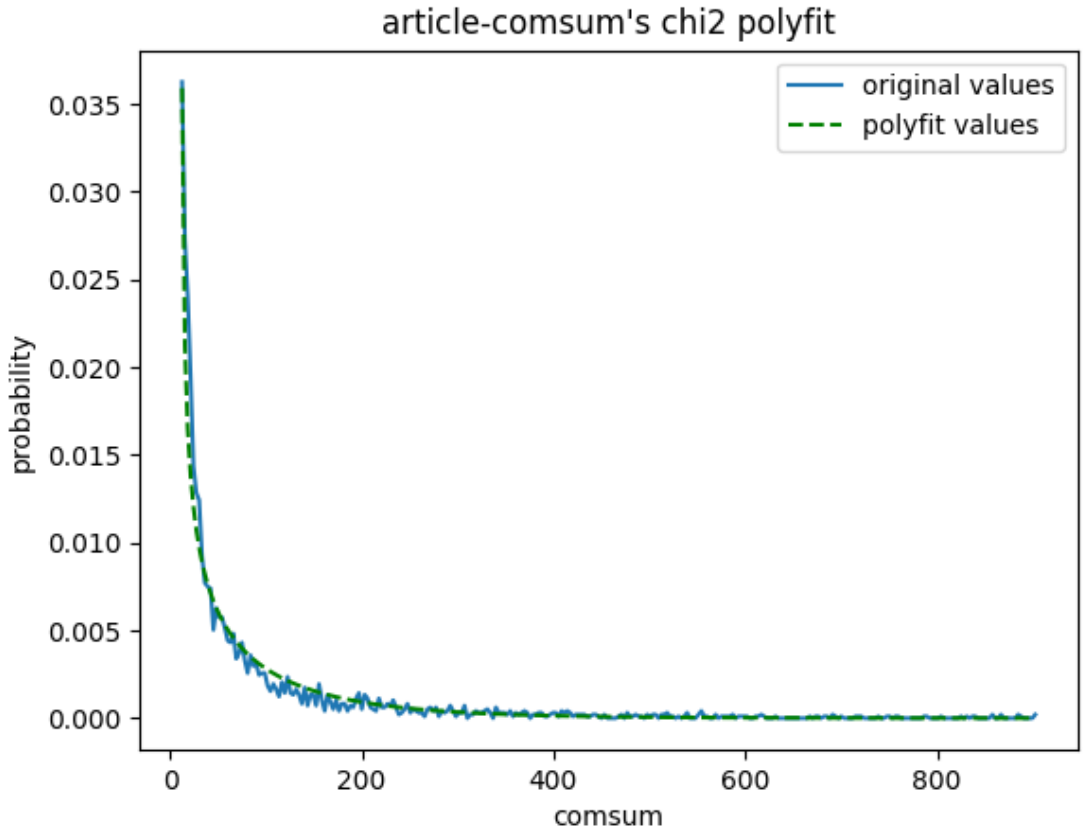
- **基于评论字长的筛选：我们认为字长小于 10 的评论是无效的评论 22W-->11W**
- **基于 Scipy 的分布拟合，**

  **剔除离群点**后（因剔除离群点故数据减少量与分位点选取不完全一致）

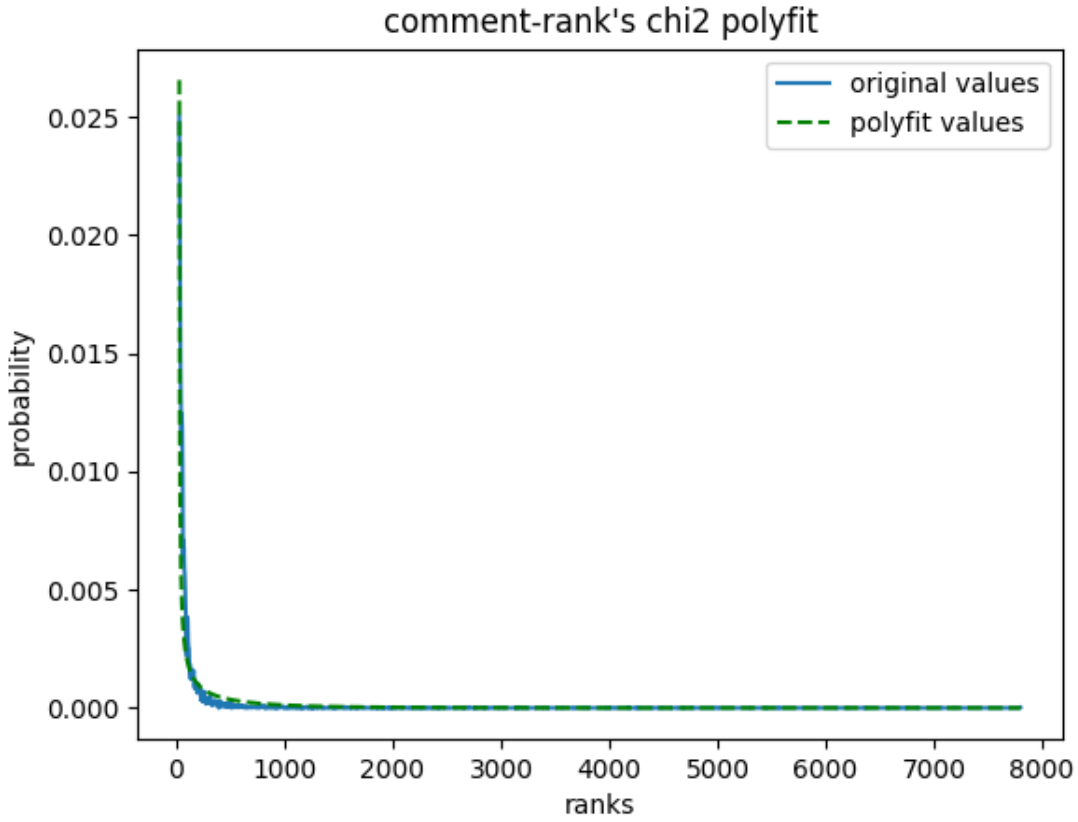
  通过卡方拟合优度检验 p 值决定使用哪种分布拟合（结论是卡方分布）

  最终通过画图来直观确认拟合效果

重点新闻筛查：拟合卡方分布选取 0.3 分位点（评论数前 30%，评论数>88）**30W-->22W**




重点评论的筛查：选取 0.6 分位点（点赞数>60）11W-->1W




**（3）心态字典构建：**

原始资料：

1. 加拿大国家研究委员会（简称 NRC）的专家创建的 NRC 通用词典 8K
2. 对新闻正文评论经由(Jieba、Hanlp)分词后，记录每个词汇出现次数，然后在高频次的词汇中，人工补充额外的心态词 （200+），并且用 TF-idf 方法量化词语与对应情感的关联度（修正情感极性）

处理思路：

NRC 虽是专家产品，但外文到中文翻译会有偏差，也无法涵盖语义更多元丰富的中文，不过作为心态词典的基底还是非常有效方便，只是需要进行词典修正去除其中无效的部分，同时补充疫情限定的以及中文限定的心态词

词典修正：

〈1〉物理修正：将爬取**所有**的评论进行分词然后碰撞，记录命中次数，去除命中次数

过低的心态词 （8K-->4K）

<2>TF-IDF：

目的：**提供词库以及权重，量化情感倾向**

具体实现：

首先解析 json 文件，得到评论数据集以及情感数据集

根据前阶段得出的关键词以及与它相关联的情感构造向量，向量格式为：[key,emotion]初始时先假定所有的 emotion（keyword 与对应情绪的关联度）均为 1

由于 emotion 存储了 9 个情绪维度，因此同时构建了词典用于映射值

之后与评论样本碰撞，计算出词频（利用 jieba 库以及已有的情绪词库，先分词再统计所有有意义的词的总数）和 idf，将两者相乘得到 TF-idf 值

再将存储 TF-idf 的列表从大到小排序，排好序后给每个向量加上它的序号，然后取排序的上四分位点之前的数据标记关联度为 1，处于上四分位至二分位点之间的数据标记关联度为 0.75，以此类推,标记好所有词语与它对应情绪的关联度

将所有新扩充词标记后就得到了词典扩充词语部分的关联度值，从而得到了 2020 年初新冠疫情场景下的情绪词典

（4）数据可视化：

对于疫情期间出现的情绪关键词我们做了**词云**处理，当然其中去除了一些无意义词汇，使之看起来更加直观。（词云背景为口罩=.=）

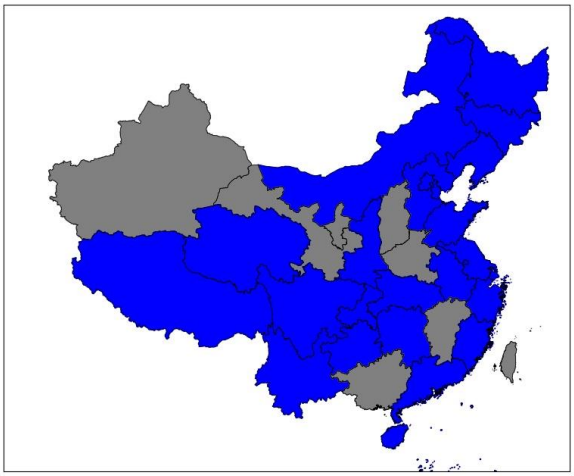

基于 basemap 第三方库实现中国各省心态可视化：

数据获取：根据前一阶段得到的特定词典以及 json 文件的解析结果，构建出根据省份，情绪以及日期归类的情绪向量，其中有做归一化处理，便于量化某个情绪的强烈程度

首先对疫情总体阶段进行总结得到心态地图

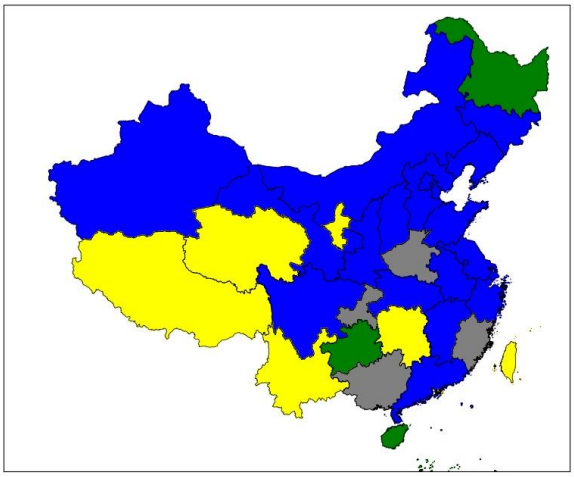
首先对各省份的心态进行分析，然后得到各省份对应的标志性心态

心态对应颜色（joy：黄色，sadness：黑色，anticipation：绿色，fear:灰色，anger：红色，hopeful：橙色，trust：蓝色，disgust：紫色，surprise：青色）

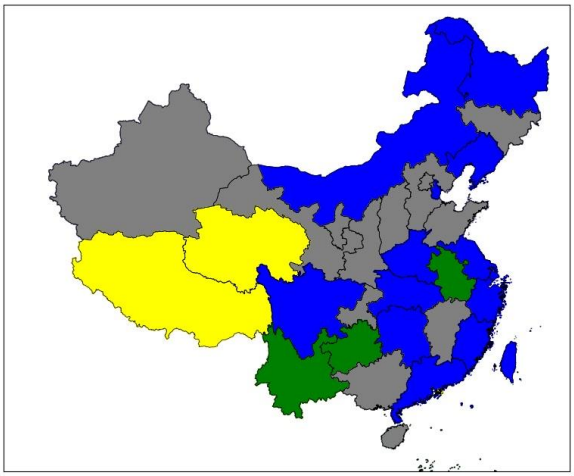


并且在疫情的四个阶段也做同样的分析

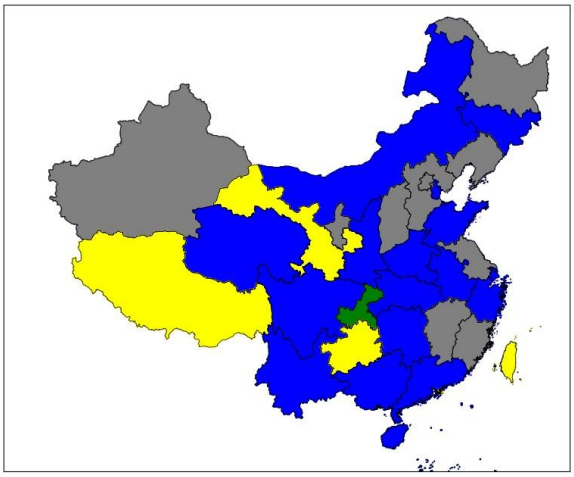
第一阶段：



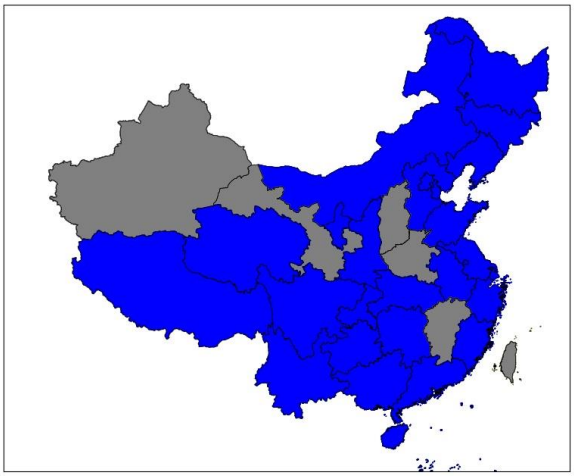
第二阶段



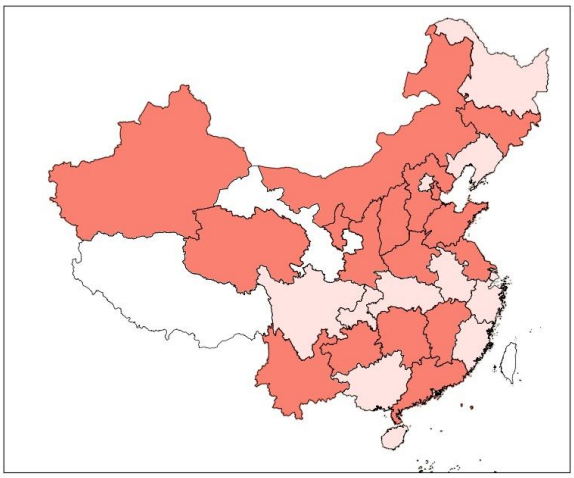
第三阶段



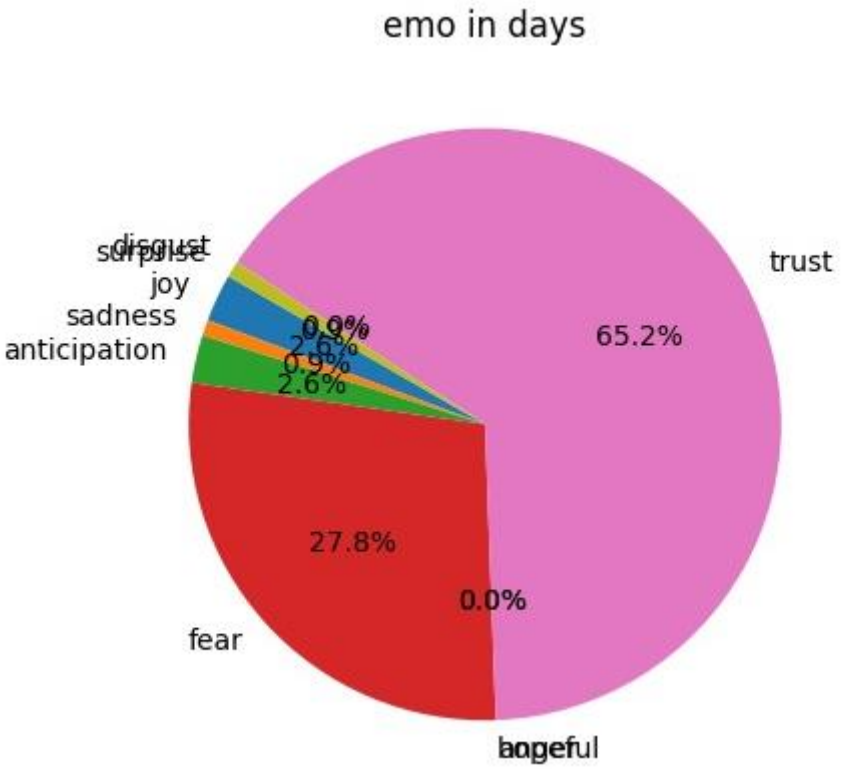
第四阶段



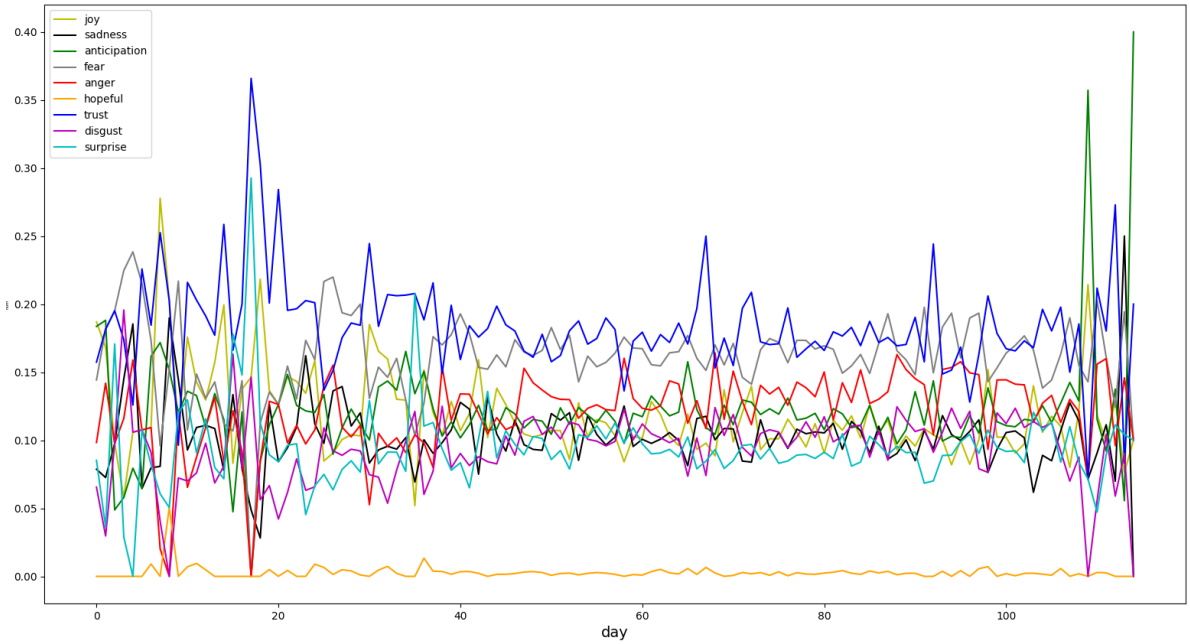
除了总体情绪分析外，我们还对每种心态在某个阶段的全国分布进行了可视化，例如第三阶段的 anger 心态（其余各阶段的各心态图片位于项目中的”..\map\各阶段各情绪“内）



我们也分析出了每天的代表情绪在所有日期中的占比情况




我们也统计了每天各种情绪权重的逐渐变化过程，绘制成折线图



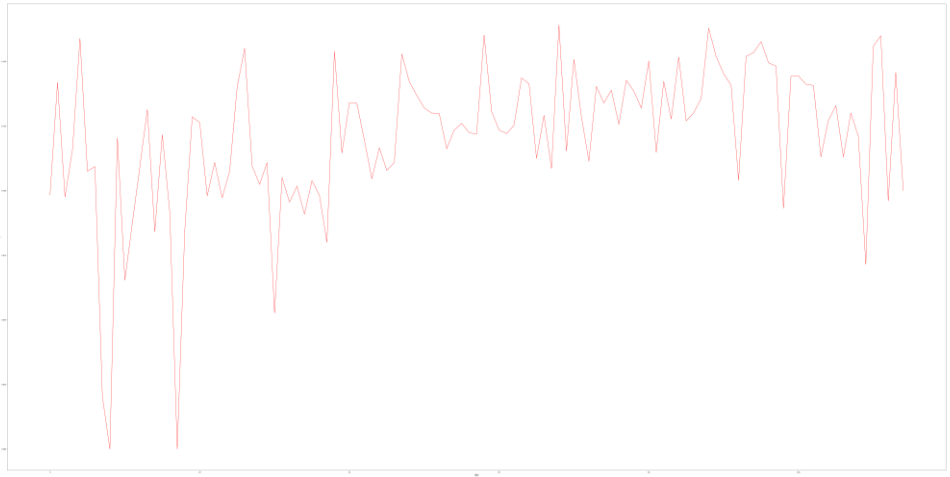

剩余单独情绪的折线图位于项目的"map\每天情绪变化"目录下

## 4. 案例分析

可视化结果已加入研究方法中，本部分将直接在可视化结果的基础上进行分析

**大的方向上：**

1. 从一阶段到四阶段，民众心态整体呈现了由信任（蓝）到恐慌（灰）再到信任（蓝）的趋势，这恰好贴合了事件从不重视与无奈扩赛阶段的盲目自信到资源缺乏阶段的过度恐慌，最后在中国切实有效的防疫步伐下重铸稳固信心的历史进程，也反映大众心态在重大事件中存在的**滞后性**
2. 不难发现，在总体的情绪占比中，trust 和 fear 的占比是最大的，分别达到 65.8% 和 28.1%，其实也不难理解，在整体疫情的恐慌情绪当中，随着其他国家的新闻频出和社会上坏新闻不断出现，同时全国同胞万众一心互相鼓励，在信息茧房的聚焦作用之下，网民在网络上接收到的新闻不断将网民内心的害怕和焦虑的情绪放大，同时互不认识的网友们互相抱团取暖，互相鼓励，在疫情不断变化的背景下传递着信心和希望
3. 折线图则与饼图联合体现出了大众心态以信任和恐慌为主体，夹杂各种**复杂**的情绪变化，以愤怒为例，在疫情较为严重的前半段大众反倒没有太多愤怒的心态，而在渐渐好转的后期反倒开始愤怒起来，这种**反常识**的心态变化有待进一步的研究



**小的方向上：**

1. 可以发现疫情最轻的西藏青海云南等边缘地区情绪一直比较积极，**证明大众的心态**

**确实与疫情态势有一定的相关性，**

2. 同时值得注意的是边缘地区（如新疆、甘肃）因疫情较轻相对不易消极，但同时也更难从消极的心态中恢复（一直处于恐惧情绪直到第四阶段），可能因为信息相对闭塞、及疫情对传统地区冲击更大

**5. 对本课程意见与建议**

大作业的评分要求变动有点频繁非常不友好，还有就是大作业讲解 PPT 的信息内容杂糅，非常容易混淆，作为辅助资料的使用感不佳。其实这些的困扰实际上都不算很大，只是细碎的不便叠加起来就会让人担心助教和老师对这门课的教学定位