

数据科学大作业报告——计算社会学篇

1. 小组信息

组长：陈益超 191250012

邮箱：1780485093@qq.com

小组人数：1

2. 概要：

本次大作业，我选择了计算社会学作为研究课题。主要参考了助教所给的思路，自主完成了数据爬取、心态词典的建立、数据标注和分析以及简单的数据可视化。由于小组仅有我一个成员，以上功能的实现皆由我独立完成。

3. 研究过程和技术路线

3.1 数据爬取

数据来源：新浪微博知名账号（人民网、央视新闻、半月谈等）的微博文章及文章评论。

本次大作业的课题要求是对2019年12月8号到2020年6月这段时间内，新冠疫情（COVID-19）这一大公共卫生事件情景下的中国大众的网络社会心态进行研究与描绘。而微博作为发布信息快速，信息传播迅速，用户数量庞大的网络平台，在相当程度上可以反映网络社会心态的变化。在信息来源方面，选择了拥有较大粉丝数量的官方账号，原因主要有：

1) 新闻数量和质量占优

2) 评论数量较多，更能反映大众心态。

技术实现：分析新浪微博网站源码，实现针对新浪微博的爬虫程序，自动爬取所需数据。

新浪微博的PC版网页，即www.weibo.com下的页面都比较复杂，爬取难度很大，于是我选择针对新浪微博的移动端网页，即www.m.weibo.cn进行分析和爬取。

以人民网微博主页为例，通过使用浏览器的开发者工具，我们可以轻易地看到，如果对新浪微博发送一个访问请求，浏览器需要提供的request url是这样的类型：

```
https://m.weibo.cn/api/container/getIndex?
uid=2286908003&t=0&luicode=10000011&lfid=100103type%3D1%26q%3D%E4%BA%BA%E6%B0%91%E7%BD%91&type=uid&value=2286908003&containerid=1076032286908003&since_id=4597207359557657
```

通过下滑翻页发现url的变动在于参数since_id，这个参数是当前页的第一条微博文章的id编号，基本上这个编号会按照从小到大的顺序分配，但却不是连续的，所以通过改变url里since_id来实现翻页操作不太现实。我通过查找资料（<https://www.bilibili.com/video/BV15b411p7i6?from=search&seid=2996445705626806906>）发现2019年时新浪用的url与现在不同，是如下类型：

```
https://m.weibo.cn/api/container/getIndex?containerid=1076032318265821&page=2000
```

每个微博账号对应一个containerid，而现在版本的url中也携带containerid参数。那么这样就简单了，从所需要爬取的微博账号主页上分析源码，得到containerid，将其代入2019年版的url，同料想的一样，虽然不能直接显示主页内容，但可以得到当前页面的一个XHR文件，可以通过这个文件获取到页面下的微博文章和相关信息。至于url里的page参数，这显然是用来选择页数的。到这里，我就基本上完成了对网页的分析，接着的就是编写爬虫程序。

爬虫的代码应该都是大同小异，这里我主要说明一下初次编写爬虫程序遇到的一些问题和解决方案：

1) 数据的初步处理。我的目标数据是微博正文和评论，以及相关的点赞数，评论数，转发数和发布时间。其中微博正文和评论中常常会带有网络链接，这并不是我所需要的。因此，我根据链接的格式，通过调re库来初步筛选爬取下来的文本信息

```
pat = re.compile("<span.*?span>|<a.*?a>")
```

2) 数据的存储。考虑到接下来可能要做的数据处理，我选择用调用xlwt库，将数据以xls文件格式来存储，后来的数据处理过程证明了用xls存储有一定的便捷之处，但效率较低。

```
weibo = (  
    text,  
    scheme,  
    created_at,  
    comments_count,  
    attitudes_count,  
    reposts_count  
)  
for col in range(0, 6):  
    sheet1.write(row, col, weibo[col])  
row += 1
```

3) 新浪微博的反爬。我在爬取数据的过程中发现新浪微博具有一定程度的反爬取功能，主要体现于如果一个账号在较短时间内向服务器发送多次访问请求，服务器将接下来的一段时间内不回应此账号的访问请求，或回复空响应。我的解决方案包括两方面，一是设置异常处理（try—expect），二是利用循环更换request的头部信息（主要是cookie）和time.sleep()方法，实现文本的持续爬取。

3.2数据筛选和处理

前面的爬虫软件在爬取的过程中是针对微博文章和评论的发布时间进行的筛选，以及对一些难以使用的信息（主要是网络连接）进行初步排查。而课题要求获取到的数据是疫情相关的新闻和评论，所以如何判断新闻文章主体是否与疫情相关，以及进行筛选，是接下来要做的工作。

1) 建立疫情相关关键词表：

初步设想为，从一部分文章主体中提取出频数最高的一部分词语，再人工手动标注其中哪些是与疫情相关，建立起疫情相关关键词表。在根据词表，与所有数据进行碰撞，筛选出与疫情相关的新闻。

初步建立的词表部分如下：

```
v1 = ['疫情', '病例', '新冠', '肺炎', '确诊', '防控', '武汉', '新增', '输入', '检测',  
      '病毒', '湖北', '医院', '死亡', '医疗队', '患者', '隔离', '专家', '医疗', '抗疫']
```

在实际运用词表v1的时候发现筛选出的新闻仍然有相当一部分与疫情无关，这说明了单纯的人为判断会有很大的误差。于是，我采取了TF-IDF文本分析方法，对词表进行优化和迭代。

$$TF = \frac{\text{关键词的出现次数}}{\text{文档中所有词的数量}}$$

$$IDF_w = \log \frac{N}{\sum_{i=1}^N I(w, D_i)}$$

得到的新的词表部分如下：

```
v2 = ['疫情', '病例', '新冠', '肺炎', '确诊', '防控', '病毒', '隔离', '抗疫', '感染',  
      '核酸', '口罩', '复工', '防疫', '疑似病例', '复产', '感染者', '疫苗', '钟南山']
```

通过抽样检验，新得到的数据疫情相关率达到84.86%

2) 接着对数据按照时间进行分组，依据课题讲解PPT上提示的按照标志大事件进行阶段性划分，这部分比较简单，就此略过。值得一提的是下一阶段的数据分析时候发现第三阶段——2020.2.10-2.13由于时间太短，相关数据较少，难以进行有效分析，于是将时间范围扩大到了2.10-3.09。

3.3心态词典的建立

心态词典的建立涉及几个方面的问题。

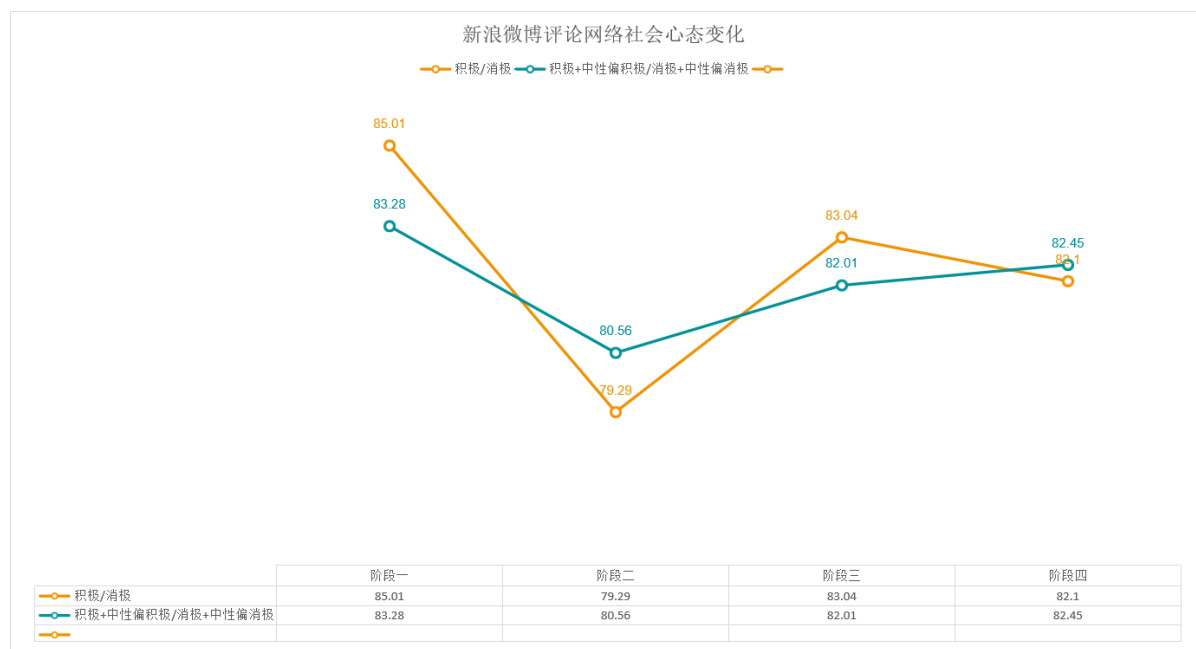
一个是情绪关键词来源。另一个是心态分类。一般而言，新闻本体，特别是官方媒体发布的新闻正文文风是十分中立的，从中能够反映情绪或心态的词十分的少，即便是有，也是积极，或中性偏积极的，分布十分不均匀，基本上不能从中分析出网络社会心态。而相较之下，新闻下面的评论则显得含有较高的情绪化因素。于是，我选择用新闻的评论文本进行分词，作为情绪关键词来源。依然采取TF-IDF的文本分析方法，辅以人工标注，在4000条关键词中筛选出能反映情绪的关键词并进行归类。先是初步分为积极、中性、消极三个大类，在情绪词归类的过程挨个添加小类，最后形成的是这样的心态分类表和心态词典：

```
积极：0：    [感激00高兴01乐观02鼓励03赞美04其他05]  
中性：1：    [怀疑10冷漠11懊悔12期待13感动14其他15]  
消极：2：    [悲痛20愤怒21忧虑22其他23]
```

```
mindkeywords = (
    ['加油03\n', '希望02\n', '致敬00\n', '感谢00\n', '好消息01\n', '挺住03\n', '相信02\n', '支持03\n', '可爱04\n', '喜欢04\n', '必胜02\n', '战胜02\n', '稳住03\n', '胜利02\n', '没事02\n', '自豪01\n', '棒棒04\n', '好看04\n', '开心01\n', '好吃04\n', '好听04\n', '欣赏04\n', '坚强03\n', '真棒04\n', '了不起04\n', '撑住03\n', '庆祝05\n', '顶上去04\n', '满意05\n', '好帅04\n', '太帅04\n', '盼望着03\n', '点赞04\n', '帅气04\n', '冲冲04\n', '热烈祝贺05\n', '敬意04\n', '鼓掌04\n', '最帅04\n', '伟大祖国05\n', '笑容03\n', '胜利在望02\n', '微笑03\n'],
    ['期待13\n', '平平安安13\n', '感动14\n', '祝愿13\n', '保佑13\n', '祈祷13\n', '羡慕13\n', '不信10\n', '前程似锦13\n', '泪目14\n', '佩服13\n', '预祝13\n', '反思15\n', '考必过13\n', '质疑10\n', '缅怀13\n', '期盼13\n', '冷漠11\n', '看哭14\n', '震撼15\n', '祝愿13\n', '震惊15\n', '走好15\n', '感人14\n', '牢记15\n', '无忧13\n', '抱歉15\n', '疑问15\n', '同情15\n', '丢脸15\n', '感激14\n', '落泪14\n', '恭祝13\n', '泪点14\n', '淡定15\n', '恳求15\n', '反省15\n', '后悔15\n'],
    ['难过20\n', '眼泪20\n', '可怜20\n', '发抖22\n', '抱怨23\n', '咋办22\n', '活该21\n', '迷惑22\n', '吓人23\n', '失望23\n', '心酸20\n', '绝望23\n', '崩溃22\n', '可恶23\n', '无情23\n', '讨厌23\n', '该死23\n', '心寒23\n', '痛心20\n', '委屈20\n', '无助23\n', '嘲笑23\n', '不服23\n', '累垮23\n', '寒心20\n', '嫌弃23\n', '伤心20\n', '自作孽23\n', '消极23\n', '耻辱23\n', '嫉妒23\n', '可笑23\n', '担忧22\n', '心碎20\n', '顶不住22\n', '气愤23\n', '生怕22\n', '报应23\n', '吓死23\n', '哭泣20\n', '煎熬22\n', '愤怒21\n', '心理压力22\n', '恫吓23\n', '嘲讽23\n']
)
```

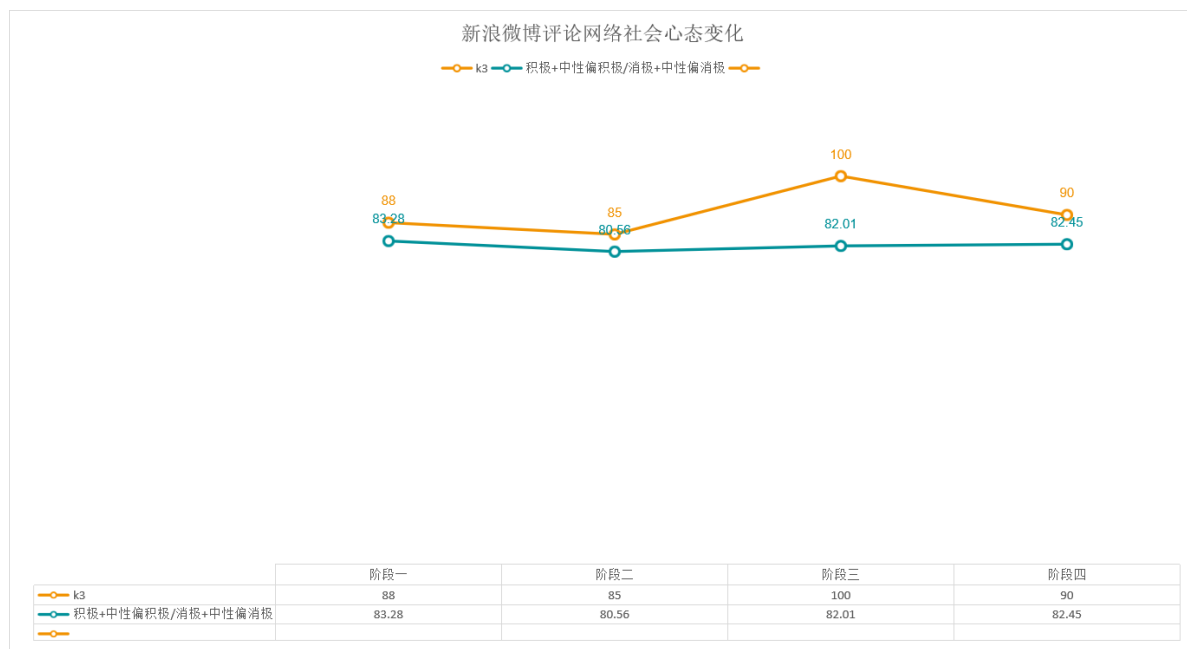
3.4数据分析

由于每个阶段划分的时间长度不同和数据量不同，不太可能用出现的心态词频数描述阶段性心态。于是我通过对出现的心态词按类划分并计算比例系数 $k_1=a_1/b_1$ ， $k_2=a_2/b_2$ （ a_1 ：积极心态词频数； b_1 ：消极心态词频数； a_2 ：积极和中性偏积极心态词频数； b_2 ：消极和中性偏消极心态词频），制作了以下折线图。（单位：百分比）




可以看到，将中性词加进统计范围得到的曲线较为平滑。基本上可以用来说明评论网友的心态变化。

前面提到过的新闻主体由于文风中立，且能提取出的关键词基本上偏向积极，并不能反映大众的心态变化，但一定程度上，可以反映官方媒体想要引导大众心态走向哪个方面。基于此，我还统计了各阶段新闻正文出现的偏积极心态关键词统计量 $k_3=1000*a_2/s-40$ （ s ：文本词数），并做了以下折线图：



4.结论

由于时间有限，没能提取出更多的统计信息。目前来看，可以得出一些粗显的推论：

- 1) 从第一到第二阶段，由于疫情爆发后的初期信息资源和物质资源都十分匮乏，公众对疫情没有一个全面的了解和把控，消极心态如忧虑、恐惧等相关心态词比例增大，使得心态曲线有一个明显的滑坡。
- 2) 从第三阶段开始，国家基本上建立起了防疫系统，各项措施开始落实，公众对疫情也有了一定程度的了解，消极心态词比例降低。这段时间出现了以下比重较大的偏积极的心态词，如“加油”“挺住”等。
- 3) 在国家落实各项防疫措施的同时，官方媒体也在积极将公众的心态引向乐观。新闻正文中常出现的关键词有“英雄”“加油”等。

5.相关链接

项目地址：<https://github.com/cycsir/software>