计算社会学

小组信息：蒋承霖 191250059
季子昂 191250057
李智强 191250078

一．研究问题，研究方向：

在本次两个课题中，我们小组选择了计算社会学这个课题。因为疫情可以说是我们这短短人生中难得经历的重大事件，因此，我们组的成员都对其非常感兴趣，因此我们选择了计算社会学这门课题。

我们小组认为，想要分析疫情期间社会大众心理变化就是需要利用心态词典，对我们获得的数据进行碰撞分析，从而宏观上获得疫情期间社会大众的心态变化。

在助教学姐的指导下，我们很快确立了我们的研究思路：首先，我们将从微博上爬取并筛选数据，然后，在获得了有用的数据之后，我们将利用创建的心态字典，对收集到的数据进行碰撞分析，将每条数据对应上一个或者多个心态关键词，从而可以借此宏观分析得到结果。最后我们将对数据进行构图等分析，并且对获得的数据进行分析，看是否能找到规律。以上就是我们对这个课题的基本思路

二．代码开源地址：

https://github.com/jclx0/-.git

‘爬虫.java’为爬虫部分的代码。’main.py’为数据分析部分的代码，’Main.java’为数据总结部分的代码。’main.cpp’为无效代码，无需关注

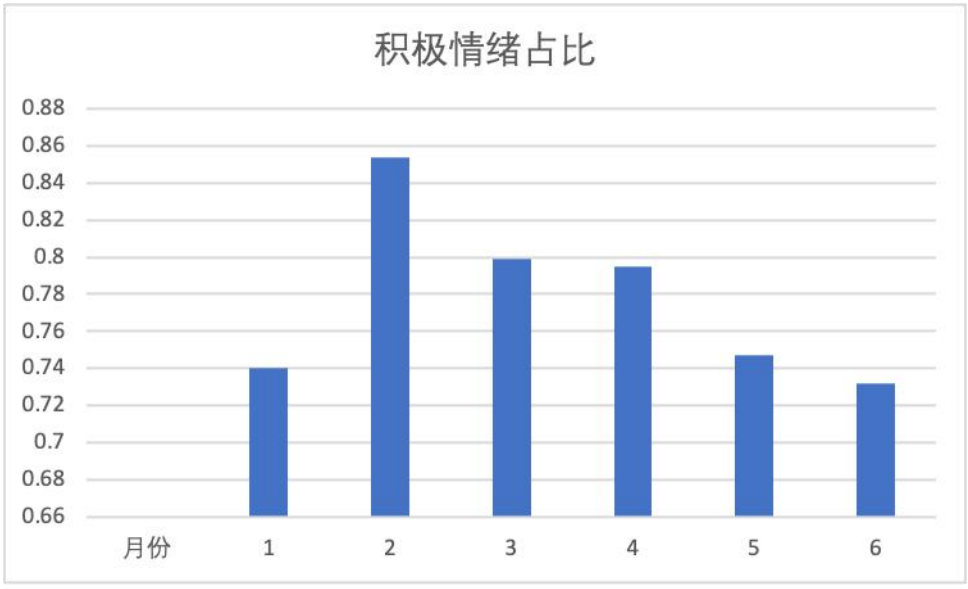
三．第一阶段：数据的爬取和数据的筛选：数据的爬取我们组使用的是 java 的爬虫代码，具体代码见以上代码开源地址。爬虫代码逻辑这一部分，我们首先是确立了我们要爬取的内容的时间，然后我们获取第一页的内容，然后将第 2 到 200 页的内容加载到任务队列中来，然后不停爬取，其中有个注意点就是每执行完一个任务，就让机器休息 5ms 左右，避免被爬取过程出现问题。而在数据的筛选上，我们组并没有采用先大量爬取然后再进行筛选，而是从问题本身出发，有针对性的选择目标数据，避免了再次筛选的麻烦，将筛选的思想贯彻到爬数据的过程中。在数据分析前，我们先用爬虫获取了 2019 年 12 月 8 日至 2020 年 6 月 15 日的与新冠肺炎疫情有关的微博上的新闻。在用爬虫爬微博数据时，我们没有爬重点新闻下的评论，而是用个体化用户发的微博来代替。我们认为采用了这种方法是一箭双雕：1.几乎直接就消除了水军的影响，从而直接代替了之后的数据筛选的作用。水军的中坚力量（僵尸粉、微博大 V 之类的不算，疫情问题他们对数据分析也无很大影响）做的事情是在热门新闻下面刷评论刷赞刷点击量，因此做数据分析时需要筛除掉这些水军的评论，而我们通过不爬评论，而用个体化用户发的微博来代替评论则直接避开了水军们 2.个体化用户的发的微博似乎更能反映大众的心态，因为评论中的词汇很大一部分是针对文章内容的，比如“文章写的真烂”、“大爱作者，期待更新”，这里的“烂”是不是分析的时候就是消极词汇呢，然而其实评论人并不一定是对疫情消极，“爱”是不是分析的时候就是积极词汇呢，然而评论人或许对疫情消极，只是赞同这篇文章作者的观点，因而通过新闻下的评论来分析大众的心态存在着一定的偏差，而用个体化用户发的微博则可以消除这个影响。因此，通过这样的逻辑，我们获得了我们的目标数据（见数据集。。。。。。）。

第二阶段：数据的分析：这一阶段是我们组进行问题分析的重点阶段，我们选用了心态字典这一方法。首先，我们先根据需求，定义了相关心态词（乐观，高兴，赞赏，担心，不满，怀疑，愤怒，后悔，失望，冷漠），然后将心态分成两个大类积极与消极，然后尽力建立了较为完善的心态词典，即关键词到心态的对应集合。建立好了心态词典之后，我们对我们之前获得数据集中的每一条数据进行分词，根据日期对其进行分组。然后将每一条数据分词后的结果与我们建立的心态词典进行映射碰撞，选择关键心态词并尽可能的全面建立与之相关的词的集合，对于得到的结果按照日期进行进一步的整合或者分析，然后输出出我们需要的数据，比如说每个月的积极心态占比，每天的积极心态占比，每月每天的各种心态占比等等。这部分代码逻辑是：对每个数据提取内容和时间，并且对内容去除停用词并按日期放入对应的 list 中，由两个 list 实现，一个 list 存储日期，一个 list 存储日期对应的数据，两者下标保持一致。而上面我们已经建立了有十个关键情绪词的心态词典。因此，我们遍历每一个词组，查询其对应的关键情绪词，如果有的话，就存到每个关键词对应的 list 中去，将各 list 长度与查询到的总情绪词 list 长度进行比对，便可以得到每个心态或者积极消极心态的占比，结果存入 txt 文档中。（根据不同的输出，整体的代码也有所调整）

第三阶段：数据的总结：目前来说我们所所获得数据中，最简洁明了的便是以下这组数据

20-01active:0.7404030461586282
20-02active:0.8539055322653882
20-03active:0.7992545794892365
20-04active:0.794905847010551
20-05active:0.747124824684432
20-06active:0.7317663459118279

（active 是指积极这一类心态在情绪词中的占比）

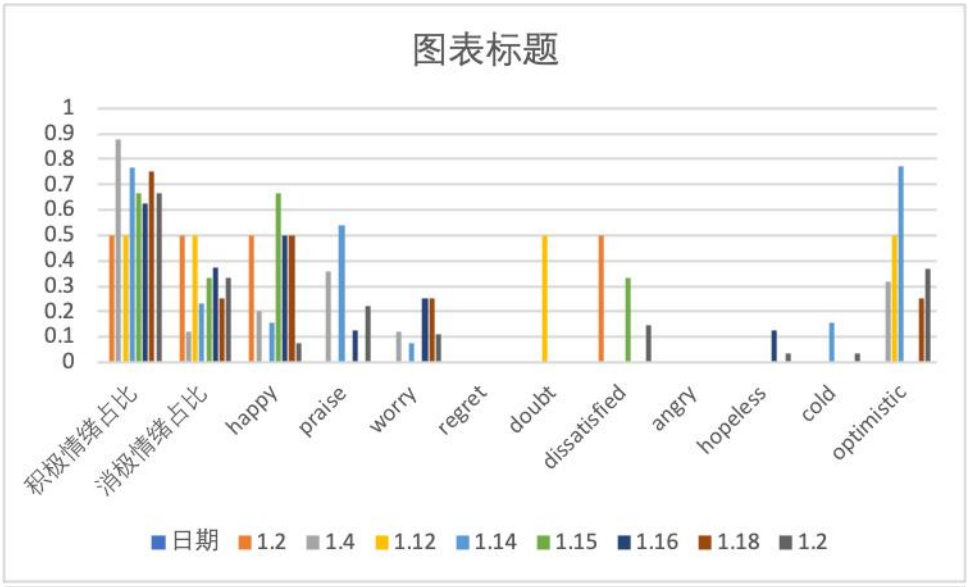


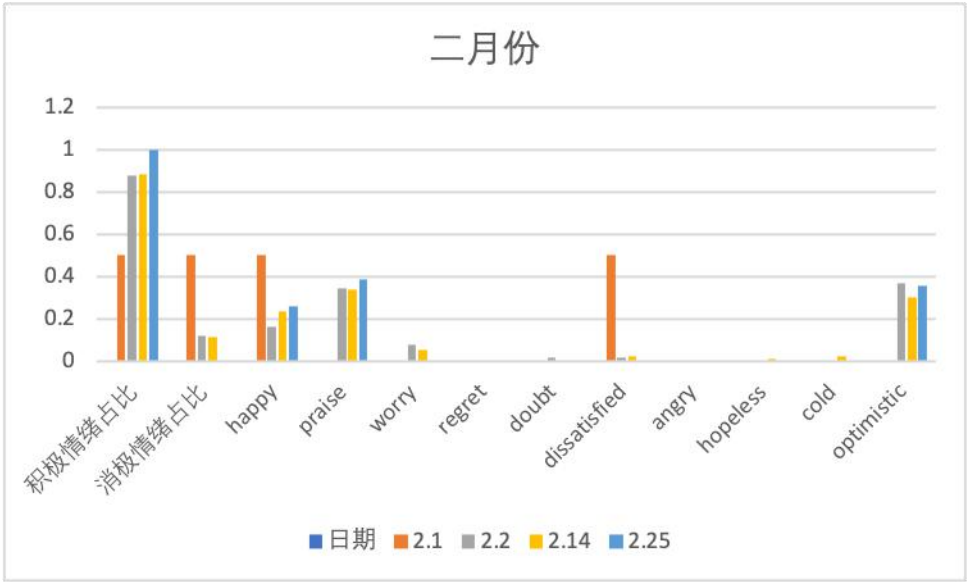
而不论是单独日期，还是像这样一个月的分割，其实所有数据最大的特点就是，跟我们所预想的其实并不一样，我们本来预想的是我们群众的心态变化应该是随着疫情的严重程度加深从而变低，又随着疫情的好转从而上升，但是我们可以看到，在疫情最可怕最人心惶惶的时候，这时候的积极情绪比例反而越来越高，而随着疫情在控制下逐渐好转下积极情绪占比却有所下降。我们经过对数据的取样研究认为，造成这样的原因是因为积极情绪我们对应的都

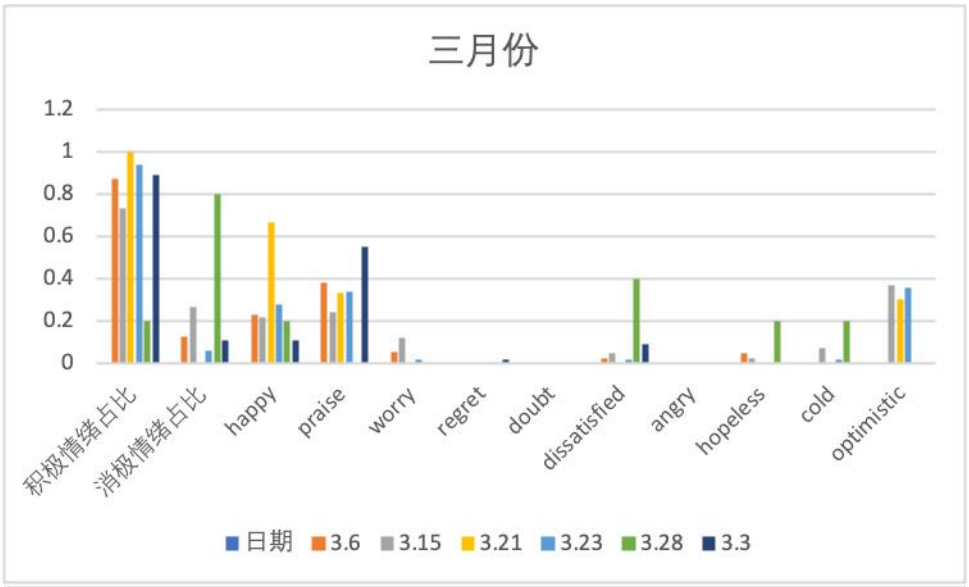
是情绪热烈的词语，因此在疫情最困难的时候，人们倾向于在网络中互相鼓励，用一些积极鼓励的话语来鼓励彼此，而随着疫情的逐渐好转，人们对疫情的信心越来越强，更倾向于用一些中性情感词语或者对国外疫情抱有一种担忧的态度，这也一定程度上影响的积极情绪的占比。

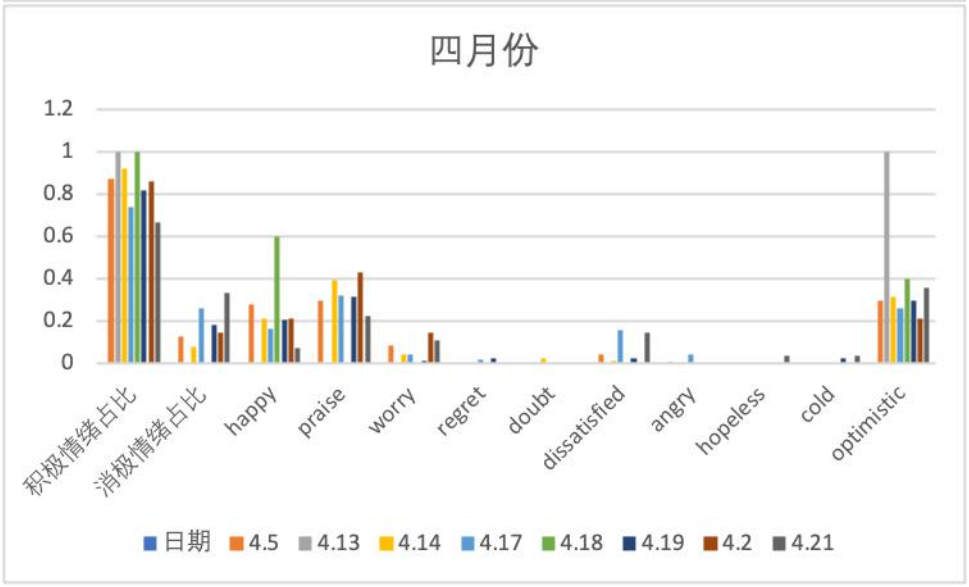
附：得到这样的数据之后，我们小组猜想，可能积极情绪的占比根据时间的变化呈现一种正态分布的趋势，因此我们利用估计法（见源代码），得到积极情绪（每天天）服从于期望为 0.8059777777777777776 和方差为 0.028889895506178387 的正态分布。可以见到即使服从于正态分布，数据所构成的正态分布曲线仍然是比较平缓的。

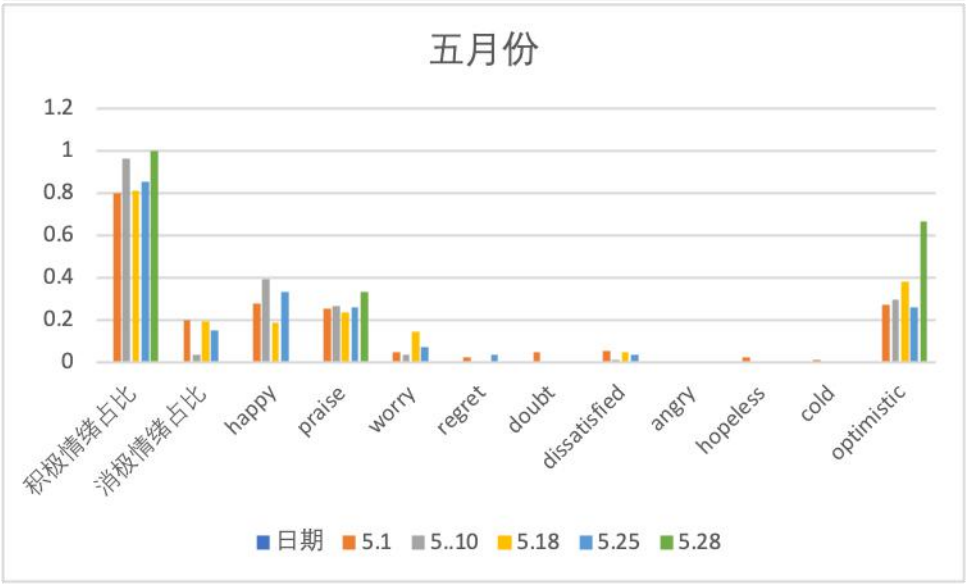
一月份

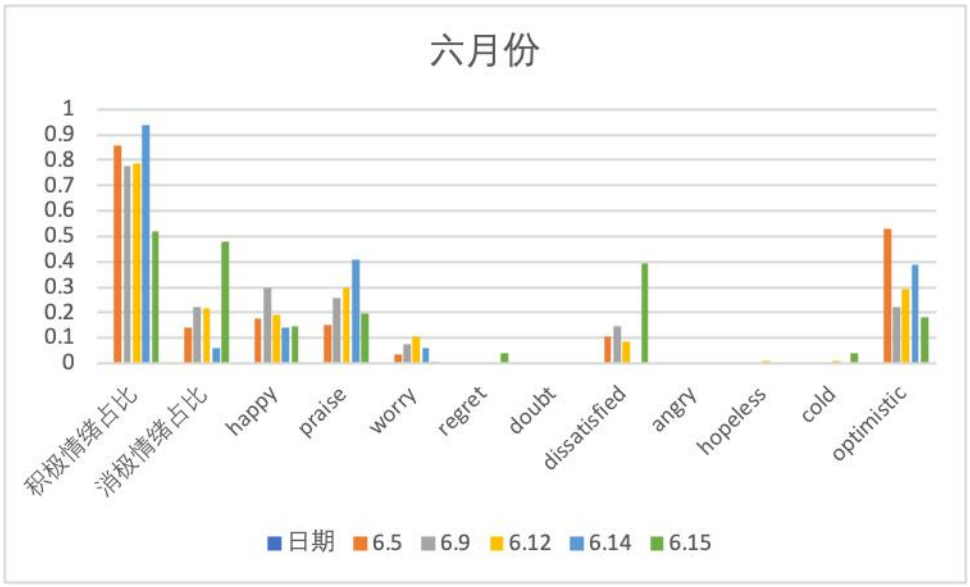





三月份
1.2
1
0.8
0.6
0.4
0.2
0
积极情绪占比
消极情绪占比
happy
praise
worry
regret
doubt
dissatisfied
angry
hopeless
cold
optimistic
日期 3.6 3.15 3.21 3.23 3.28 3.3


四月份
1.2
1
0.8
0.6
0.4
0.2
0
积极情绪占比
消极情绪占比
happy
praise
worry
regret
doubt
dissatisfied
angry
hopeless
cold
optimistic
日期 4.5 4.13 4.14 4.17 4.18 4.19 4.2 4.21


五月份
1.2
1
0.8
0.6
0.4
0.2
0
积极情绪占比
消极情绪占比
happy
praise
worry
regret
doubt
dissatisfied
angry
hopeless
cold
optimistic
日期 5.1 5..10 5.18 5.25 5.28


六月份
1
0.9
0.8
0.7
0.6
0.5
0.4
0.3
0.2
0.1
0
积极情绪占比
消极情绪占比
happy
praise
worry
regret
doubt
dissatisfied
angry
hopeless
cold
optimistic
日期
6.5
6.9
6.12
6.14
6.15