一、 背景及需求分析

2022 年 3 月 1 日下午，上海市举行疫情防控工作新闻发布会，会上通报了 3 月 1 日新增一例本土病例。随后，疫情在上海大规模爆发，上海疫情备受关注。基于此背景下，我们爬取此次疫情以来，在微博平台上民众关于“上海疫情”的发言，并进行数据分析。通过分析，了解上海此次疫情期间民众的关注点和情绪。

由于我们想要分析在微博平台上关于上海疫情发生以来人们的发言，因此，采集发微博日期和微博正文内容。由于微博网页版提供高级搜索功能，可以指定时间段进行关键字搜索，因此我们确定使用微博搜索网页版获取相应的数据（https://s.weibo.com/）。

二、所用工具

2.1 运行环境

- Chrome 版本（32 位）
- Python 3.8.2

2.2 前置库

- requests
- datetime
- BeautifulSoup
- jieba
- sklearn
- snownlp

三、数据采集

3.1 分析网页

由于微博高级搜索功能只能获取 50 页指定时间段的内容，50 页后为实时发布内容，为了获取足够的数据，我们指定搜索时间段单位为天，爬取 3 月 1 日以来关于“上海疫情”的每天发布内容。

借助 Chrome 开发者工具(F12)来分析网页 ，在 Elements 下找到需要的数据位置。

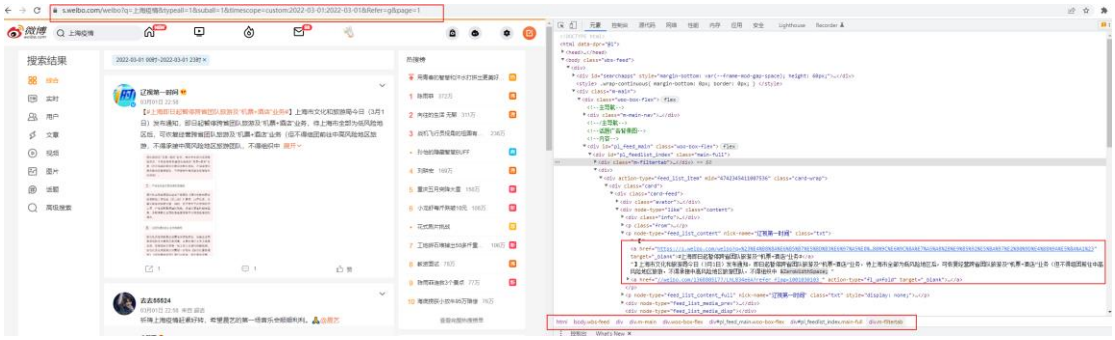


图 1 使用开发者工具(F12)打开的微博内容界面

由上面的网址(https://s.weibo.com/weibo?q=%E4%B8%8A%E6%B5%B7%E7%96%AB%E6%83%85&typeall=1&suball=1&timescope=custom:2022-03-01:2022-03-01&Refer=g&page=1)可知，我们只需要修改爬取时间段(timescope=custom)和页数(page)就可以获取指定时间段内指定页数的微博内容。

3.2 爬取网页

首次使用 requests 来爬取，但是微博有反爬虫机制，不能获取到我们需要的数据。查阅资料后，采用 selenium 直接登录并获取 cookie 来模拟登录，同时进行会话保持操作。

```
def login_account():
    '''账户登录，初始化 cookie'''
    driver = webdriver.Chrome()
    driver.get("https://login.sina.com.cn/signup/signin.php")
    wait = WebDriverWait(driver, 5)
    # 暂停 1 分钟进行预登陆，填写账号密码及验证
```

```
    time.sleep(60)
    cookies = driver.get_cookies()
    s = requests.Session()
    c = requests.cookies.RequestsCookieJar()
    for item in cookies:
        c.set(item["name"], item["value"])
    s.cookies.update(c)  # 载入 cookie
    return s
```

登录后使用 requests 来返回微博内容并转换为 utf-8 编码形式：

```
''' 爬取网页，返回微博内容 '''
html = session.get(url)
html.encoding = 'utf-8'
```

3.3 解析内容

由图 1 我们可以看到微博内容正文位于以下位置：

```
#pl_feedlist_index > div >div.card-wrap >div.card >div.card-feed>div.content>p.txt
```

BeautifulSoup 是是一个 HTML/XML 的解析器，主要的功能是解析和提取 HTML/XML 数据，它使用简单，支持 CSS 样式选择器，因此我们使用 BeautifulSoup 的 select()函数通过 CSS 样式选择器来进行元素查找，该函数返回一个包含所有指定元素的列表。

```
soup = BeautifulSoup(html.text, 'lxml')
txt_list = soup.select('#pl_feedlist_index > div >div.card-wrap >div.card >div.card-feed>div.content>p.txt')
```

3.4 保存数据

获取到内容列表后，我们提取其中的文字并保存为 CSV 文件，以备后面使用。我们保存的 CSV 文件部分截图如图 2 所示，包含时间和微博正文两项：

	A	B
1	search_date	content
2	2022/3/1	【#上海即日起暂停跨省团队旅游及“机票+酒店”业务#】上海市文化和旅游局今日（3月1日）发布通知，即日起暂停跨省团队旅游及“机票+酒店”业务，待上海市全部为低风险地区后，可恢复经营跨省团队旅游及“机票+酒店”业务（但不得组团前往中高风险地区旅游，不得承接中高风险地区旅游团队，不得组织中高风险地区游客外出旅游），并要求各级文旅管理部门、各旅行社和在线旅游企业，切实贯彻落实防疫各项要求，继续毫不放松、科学精准做好旅游团队疫情防控工作。（乐游上海）0关于本市暂停跨省团队旅游及“机票+酒店”业务的通知
3	2022/3/1	祈祷上海疫情赶紧好转，希望晨艺的第一场音乐会顺顺利利。@晨艺
4	2022/3/1	【#上海新增1例本土确诊病例#】记者3月1日从上海市新冠肺炎疫情防控新闻发布会上获悉，上海新增1例新冠肺炎本土确诊病例，该病例2月28日下午因发热前往上海市同济医院发热门诊就诊，新冠病毒核酸检测结果为阳性。上海市普陀区石泉街道宁强路33号石泉社区文化活动中心被列为中风险地区，上海市其他区域风险等级不变。（@新华社 记者袁全、史依灵）#值阳生活# 2上海·普陀区
5	2022/3/1	刚刚社区打来电话，让去核酸，我只不过是2.23从江阴去了上海九院，在上海连头带尾才4小时，带＊了，这样得麻烦几天了！有的公司一刀切，根本不让你进门，哎！ 2无锡
6	2022/3/1	从二月二十几号就计划着去上海，去见你，去打卡你呆过的地方，可是因为这个b疫情，把我的一切计划全都打乱了！拿什么还我
7	2022/3/1	一心三用上海又带星了，武汉还在增。疫情反复不停@些许焦虑
8	2022/3/1	每次刚从上海回来，上海就疫情
9	2022/3/1	刚摘完苏州的@，又戴上了上海的@，这疫情啥时候结束啊！
10	2022/3/1	祈祷上海疫情赶紧好转，希望晨艺的第一场音乐会顺顺利利。@晨艺
11	2022/3/1	要去上海看病，年前上海疫情，单位不给放行！年后苏州疫情，人家上海医院嫌弃！金矿上海又爆出核酸阳性，妈呀，我什么时候才能去上海！体制内得悲哀！

图 2 爬取内容保存的 csv 文件部分截图

四、分词、数据清洗、词频统计

4.1 分词

中文分词指的是将一个汉字序列划分为一个一个单独的词。jieba 库是一款优秀的 Python 第三方中文分词库，它支持三种分词模式：精确模式、全模式和搜索引擎模式。其中精确模式试图将语句最精确的切分，不存在冗余数据，适合做文本分析；全模式将语句中所有可能是词的词语都切分出来，速度很快，但是存在冗余数据；搜索引擎模式是在精确模式的基础上，对长的词语再次进行划分，提高召回率，适合用于搜索引擎分词。

我们想要统计每条微博数据的关键信息以便对搜集到的所有数据做分析。因此使用 jieba 库的精确模式实现对爬取的微博文本进行分词。

4.2 数据清洗

停用词(stop words)是自然语言处理领域的一个重要工具，通常被用来提升文本特征的质量，或者降低文本特征的维度。通常情况下，停用词可以分为两类：1. 使用十分广泛，甚至是过于频繁的一些词，如“我”,“就”等词，2 文本中出现频率很高，但实际意义不大的词，主要包括语气助词、介词、连词等。文本中如果存在大量停用词容易对有效信息造成噪音干扰。

分词后的微博文本数据会存在很多没有意义的词语，如英文字符、数字、数字字符、标点符号及使用频率特别高的单汉字。因此进行后续数据分析之前，我们先进行数据清洗，删除停用词。主要方法为：若经过分词的微博文本数据在停词表中出现，则将这个词删掉。

4.3 词频统计

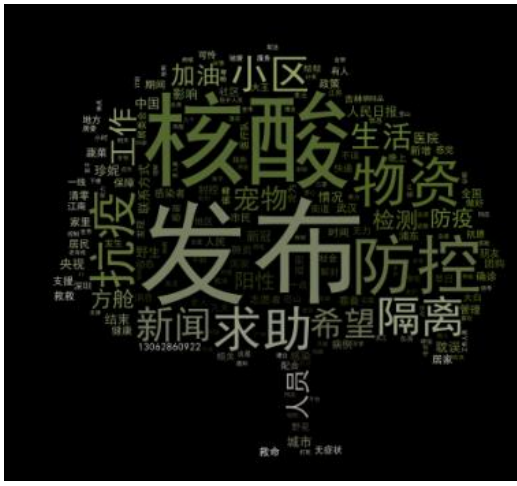
分词及数据清洗之后，就进行词频统计。通过词频统计可以看到大家关注点。我们以后为单位进行统计。统计结果以词云方式呈现：

(a)2022-3-1—2022-3-7 词云图

(b) 2022-3-8—2022-3-14 词云图

(c) 2022-3-15—2022-3-21 词云图

(d) 2022-3-22—2022-3-28 词云图



(e) 2022-3-29—2022-4-4 词云图

(f) 2022-4-5—2022-4-11 词云图

(g) 2022-4-12—2022-4-18 词云图

(h) 2022-4-19—2022-4-25 词云图

(i) 2022-4-26—2022-5-2 词云图

(j) 2022-5-3—2022-5-9 词云图

图 3 2022 年 3 月 1 日至 2022 年 5 月 9 号每周词云图

从图 3 可以看出，从 3 月 1 号上海本波疫情发生以来微博平台上的人们关注点的变化，刚开始的时候人们比较关注新增病例，防控，街道检测等。3 月 15 日那周开始，隔离和小区成为人们的关注点，4 月份开始，物资、希望、求助等成为高频词，体现了疫情之下，人们守望相助，同舟共济抗击疫情。

五、聚类分析

5.1 词频矩阵

将文本数据和词频统计得到的 50 个高频词分别作为词频矩阵的两个维度建立词频矩阵，即词频矩阵中的每行代表一条评论的内容，每列代表一个关键词。矩阵中的 1 表示评论中含有对应列的关键词，而 0 表示评论中不含对应列的关键词。表 1 展示 2022 年 3 月 1 日至 2022 年 3 月 7 日这一周内部分词频矩阵。

表 1 2022 年 3 月 1 日至 2022 年 3 月 7 日部分词频矩阵

	确诊	核酸	感染者	本土	隔离	无症状	地区
评论 1	0	0	0	0	0	0	0
评论 2	0	0	0	0	0	0	0
评论 3	0	0	0	0	0	0	1
评论 4	0	0	0	0	0	0	0
评论 5	0	0	0	1	0	0	0
评论 6	0	0	0	1	0	0	0
……	……	……	……	……	……	……	……

通过数据转化成词频矩阵的形式，可以将每条文本数据转换为 n 维 1/0 值向量的形式，可以用来对比分析每个向量之间的相似性进而对文本数据进行分类。

5.2 聚类分析

在对每周进行精准分类前，需要通过 SPSS 软件大致估计出有几个类别。

将 5.1 中计算得到的词频矩阵导入 SPSS 软件，进行分析-分类-系统聚类，经过多次试验，发现系统聚类的方法选择组间连接-欧式距离较好，该方法采用简单匹配系数度量评论之间的相似性，简单匹配系数是当两条评论在关键词上的数

值相同时出现的频率，频率越高说明两条评论越相似。根据对 4.3 高频词和 3.4 评论的分析，选择谱系图分类数量在 3 到 7 之间。具体每周的谱系图在附录文件中。

通过 SPSS 的估计后观察谱系图可以发现所有的数据大致被分为了 4 类，下面再进行具体的聚类。

六、自然语言处理、情感分析

6.1 TF-IDF 算法以及 K 均值聚类

精确聚类使用 TF-IDF 算法，TF-IDF 即 term frequency–inverse document frequency，是一种用于信息检索与数据挖掘的常用加权技术。其中 TF 是词频（Term Frequency），IDF 是逆文本频率指数（Inverse Document Frequency）。TF-IDF 的主要思想是：如果某个词或短语在一条数据中出现的频率 TF 高，并且在其他数据中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类。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。计算 TF-IDF 时还可以选择将出现频率高于一定值，以及出现次数非常低的词语删去，同时还可以借助正则表达式去除数字、符号等不利于分析的元素。

具体的实验过程如下：

1. 加载语料数据，并将每条数据保存为 list 的元素；
2. 计算 TF-IDF；

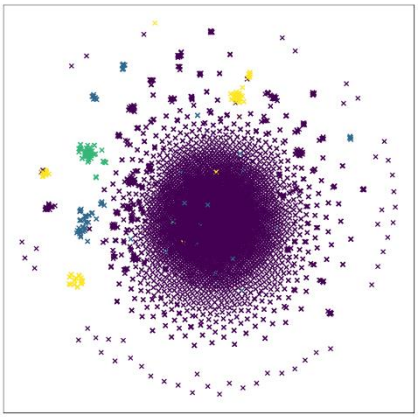
```
# 将文本向量化
vectorizer = CountVectorizer(
    max_df=0.8, min_df=2, token_pattern=u'(?u)\\b[^\\d\\W]\\w+\\b')
transformer = TfidfTransformer()
# 拟合数据模型，返回词频矩阵
tfidf = transformer.fit_transform(vectorizer.fit_transform(corpus))
```

3. 获取模型中的所有字词特征（如果特征数量非常多的情况下可以按照权重降维），经过程序统计一般有 6000 个左右的特征词；
4. 导出权重矩阵，此时实现了将文字向量化的过程，矩阵中的每一行就是一条文本的向量表示。
5. 将文本数据的向量进行 K 均值聚类，K 均值聚类算法是在向量空间中分配 k 个随机点作为 k 个簇的初始虚拟均值，这里根据 5.2 聚类分析的结果指定分成 4 个类。然后，将每个数据点分配给平均值最近的聚类。接下来，重新计算每个聚类的实际平均值。根据均值的偏移，重新分配数据点。重复此过程，直到集群的平均值停止移动；

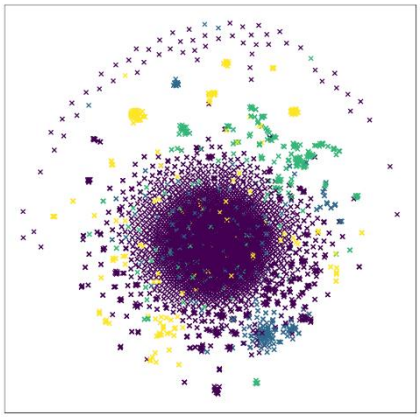
```
kmeans = KMeans(n_clusters=4)
kmeans.fit(tfidf_weight)
```

这里主要使用了 sklearn 机器学习库，可以方便地实现聚类。

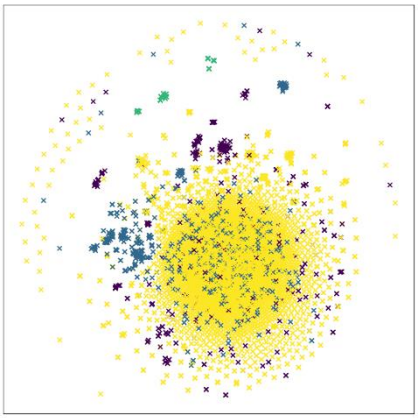
6. 得出聚类结果。将聚类后的文本数据进行词频统计，可以得到精确分类后的数据。



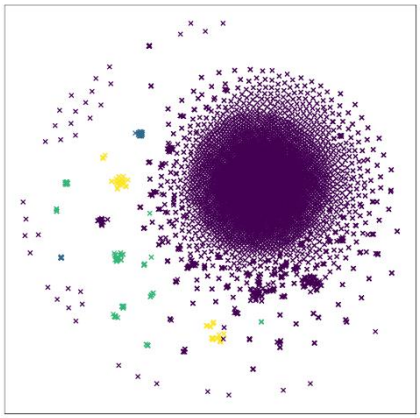
(a)2022-3-1—2022-3-7 聚类结果



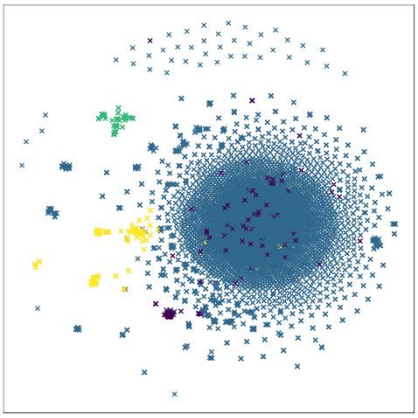
(b) 2022-3-8—2022-3-14 聚类结果



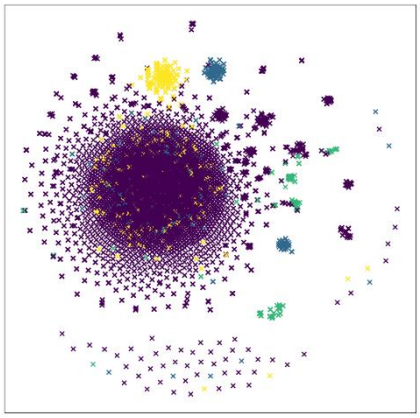
(c) 2022-3-15—2022-3-21 聚类结果



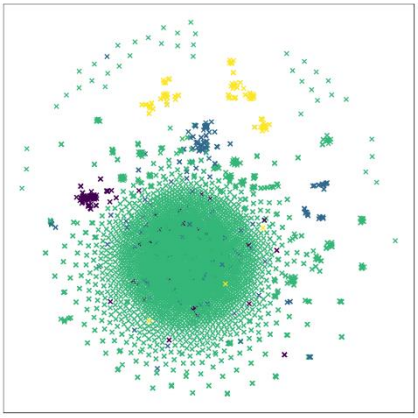
(d) 2022-3-22—2022-3-28 聚类结果



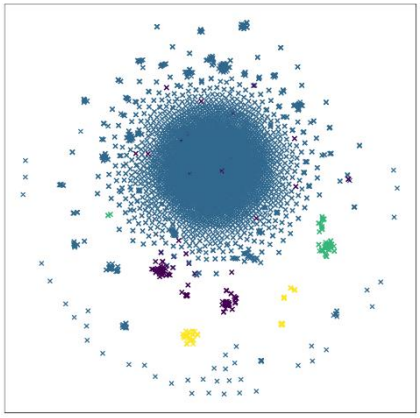
(e) 2022-3-29—2022-4-4 聚类结果



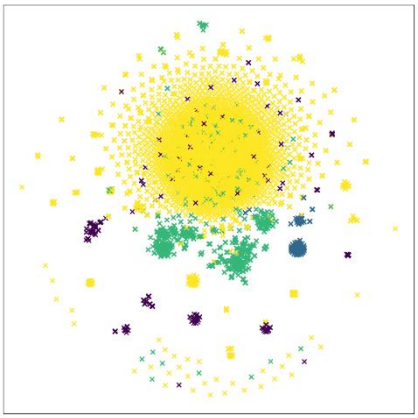
(f) 2022-4-5—2022-4-11 聚类结果



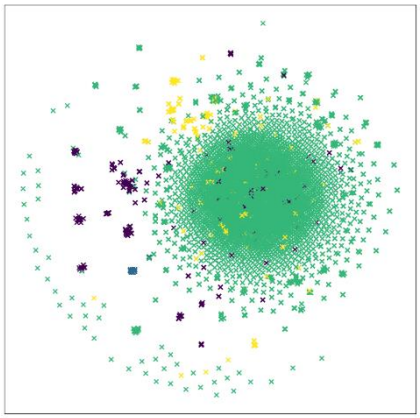
(g) 2022-4-12—2022-4-18 聚类结果



(h) 2022-4-19—2022-4-25 聚类结果



(i) 2022-4-26—2022-5-2 聚类结果



(j) 2022-5-3—2022-5-9 聚类结果

图 4 2022 年 3 月 1 日至 2022 年 5 月 9 日每周聚类结果

用不同颜色代表不同的类别，从聚类的结果可以发现，分成四种类别，有一类的数据样本比较大，其余的三类数据样本的占比比较小。观察每种类别的聚类效果，发现有些周单独类别可以很好的聚类，较为集中；而某些周，如 2022-4-26—2022-5-2 这周绿色和紫色类别都比较分散。

6.2 情感分析

我们想要通过情感分析来判断这段时间内舆论的走向。文字内蕴情感，这里借助使用 snownlp 库，对爬虫获取的未分词评论进行情感分析。使用 SnowNLP()方法，里面调用了通用的情感分析模型，目前的训练集来自在线商城平台的评价，如果需要对特定领域进行情感分析，可以导入数据进行训练，来提高模型在应用场景下的准确性。

举例如下：

```
s1 = "如果27号我偷摸跑去上海回来是不是会全校通报"
s2 = '无语了祈祷太原没疫情'
s3 = "祈祷上海疫情赶紧好转，希望你的第一场音乐会顺顺利利。"
s4 = '每次刚从上海回来，上海就疫情'
s5 = '又被我妈整破防'
```

图 5 从第一周的评论中随机选取的 5 条

输出如下：

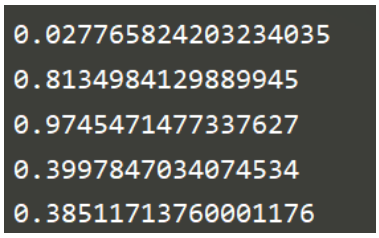


图 6 从第一周的评论中随机选取的 5 条蕴含积极情感概率输出

一般，我们认为概率大于 0.5 是积极的情感，低于则是消极的。观察结果，可以发现 s2 和 s3 被认定为积极情感。其余为消极情感。

接着我们对每一周的评论进行了一次平均，经过计算得到了一个蕴含积极情感文本百分比。结果如表 2 所示。

表 2 2022 年 3 月 1 日至 2022 年 5 月 9 日每周情感分析结果

通过数据的观察可以看出，在上海疫情爆发的这两个月内，微博内的评论 8 成都是消极的。绝大多数的人都是不冷静的，或许刚好也符合微博就是‘倒垃圾的垃圾场’的定位。总而言之，疫情发生算不上一件好事，上海对这次的疫情是始料未及，也发生了许多让人大跌眼镜的事。希望上海能够早期转好。

七、设计过程中存在的问题和解决过程

参考资料：

https://blog.csdn.net/s643494618/article/details/122075505

https://zhuanlan.zhihu.com/p/391608286