

注：教师姓名后留有一个空格，后面填写教师职称。下面加下划线。

阅后删除此文本框。

**Python设计**

**基于微博评论的简单舆论分析**

学生姓名： 曾柏林 \_\_\_\_ \_

学 号：\_\_\_\_\_\_17550510004 \_

院（系）：\_\_\_\_\_ 数学系 \_\_\_\_\_\_\_\_\_

专 业： 信息与计算科学(17信计) \_

# 摘 要

我们现在处于21世纪，属于计算机的世界，生活处处离不开智能设备，例如手机、电脑、摄像头等。当然我们也处于一个生活隐私容易被曝光的时代，身边处处都充满着数据，当20世纪机器学习横空出世，智能世界更向前一步，也将我们的隐私推向于公众视野下。随着科学技术的进步，我们获得数据的来源越来越多，手段越来越简单，但是数据量实在太大了，数量级可达上百万不等。我们需要一种手段去获取数据的内在联系，这也是我们研究的意图。

本课题主要研究基于微博评论的简单舆论分析，首先微博作为现代中国人社交舆论最大的平台，其数据对我们研究很有帮助。每当每个圈子出现什么大事，微博平台是我们第一时间能接收的信息平台，所以微博是我们研究的第一方向。此课题主要使用python语言进行对微博用户的微博评论进行爬虫获取，并且使用python来对评论数据进行情感分析、作图，让我们更能认清事情的本质。

**关键词：**python；爬虫；情感分析；

# Abstract

We are now in the 21st century, belongs to the computer world, life is inseparable from smart devices everywhere, such as mobile phones, computers, cameras and so on. Of course, we are also in an era when the privacy of life is easy to be exposed, and we are full of data everywhere. When machine learning came out in the 20th century, the intelligent world will take a step forward, and also push our privacy to the public vision. With the progress of science and technology, we have more and more sources of data, and the means are more and more simple, but the amount of data is too large, the order of magnitude can reach millions. We need a means to get the inner connection of data, which is also the intention of our research.

This paper mainly studies the simple analysis of public opinion based on Weibo comments. Firstly, as the largest platform of social public opinion of modern Chinese, Weibo data is very helpful for our research. Every time something big happens in every circle, Weibo platform is the first information platform we can receive, so Weibo is the first direction of our research. This topic mainly uses Python language to crawl the comments of microblog users, and uses Python to analyze and map the comments data, so that we can better understand the essence of things.

**Key Words：**Python；Scary; Sentiment analysis

# 目录

[摘 要 1 -](#_Toc26476658)

[Abstract 2 -](#_Toc26476659)

[1. 前言 4 -](#_Toc26476660)

[**1.1研究背景** 4 -](#_Toc26476661)

[**1.2 研究的意义** 4 -](#_Toc26476662)

[**1.3 运行环境说明** 4 -](#_Toc26476663)

[**1.4 论文章节安排** 5 -](#_Toc26476664)

[2. 微博爬虫概述 5 -](#_Toc26476665)

[**2.1 微博平台全面分析和选择** 5 -](#_Toc26476666)

[**2.2 微博评论分析** 7 -](#_Toc26476667)

[**2.3 爬虫代码实现与分析** 8 -](#_Toc26476668)

[**2.4 小结** 12 -](#_Toc26476669)

[3. 中文数据的情感分析 13 -](#_Toc26476670)

[**3.1 基于情感词典的方法分析** 13 -](#_Toc26476671)

[**3.2 基于机器学习的方法分析** 13 -](#_Toc26476672)

[**3.5 小结** 14 -](#_Toc26476673)

[4. 建立中文情感分析模型 14 -](#_Toc26476674)

[**4.1 基于情感词典的建模** 15 -](#_Toc26476675)

[**4.2 基于机器学习的方法** 17 -](#_Toc26476676)

[**4.3 分析结果** 19 -](#_Toc26476677)

[**4.4 小结** 21 -](#_Toc26476678)

[5. 总结与展望 21 -](#_Toc26476679)

[参考文献 23 -](#_Toc26476680)

**基于微博评论的简单舆论分析**

在我们浏览网页,浏览器会渲染输出HTML、JS、CSS等信息；通过这些元素，我们就可以看到我们想要查看的新闻,图片,电影,评论,商品等等。一般情况下我们看到自己需要的内容，图片可能会复制文字并且下载图片保存，但是如果面对大量的文字和图片，我们人工是处理不过来的，同时比如类似百度需要每天定时获取大量网站最新文章并且收录，这些大量数据与每天的定时的工作我们是无法通过人工去处理的，这时候我们可以使用爬虫手段来抓取数据，并且可以使用python大量第三方优秀内库来对数据进行剖析。

# 1. 前言

## **1.1研究背景**

当今社会，微博作为一种新兴的社交网络和信息分享、传播方式，在最近几年变得尤为流行。常用的微博服务、比如国外的推特、国内的新浪微博、腾讯微博等，都有上亿的注册用户。微博最初的理念就是让用户能够更加简单的与好友、家人分享自己正在做什么。实际中，用户不仅用微博发布正在做什么的信息，还用它来与好友交流，转发一些突发的新闻事件，对某件事发表自己的观点，甚至在危险的时候作为一种求救方式。

如今，随着微博等包含大量短文本的社交媒体兴起，研究人员越来越多的把精力放在针对微博上自媒体、社交媒体等研究。社交媒体数据挖掘领域越来越受关注，例如最近的华为251、网易裁员等大新闻。

我们的研究就是基于此微博中的大新闻，我们想要去了解大事背后民众的思想，想看看人民对于此的态度。研究课题的对象是澎湃新闻中有关于华为251事件的报道中微博的评论，我们想去分析这件事情的背后。

## **1.2 研究的意义**

微博更像是传统博客与社交网络的一个结合体。微博服务的一个重要特性就是它的实时性。比如，传统的微博用户可能每过几天会更新一次微博的内容。而微博用户经常会在一天内发布多条微博内容。而且用户发布的微博内容大部分都是跟用户当天的活动有关。因此，微博内容可以被看作是基于人的信息聚合种子。微博的这些特性让研究者有了新的途径去获取用户的实时兴趣。与传统获取用户数据的方法相比，微博提供了更海量、更实时的数据。

对微博评论的研究具有重要意义，并且可以加强我们python编程的基础，能让我们能做更多很酷的事情。

## **1.3 运行环境说明**

开发语言：Python 3.7.3

开发环境：64位Windows10系统，12G运行内存，AMD 2500U处理器。

数据库：MySQL5.56

开发工具:Python编辑器：PyCharm 2019.1；

MySQL管理工具：SQLyog12.9

使用说明: 1）本机电脑要安装有MySQL，并且账号密码皆为root。若密码不是root，请在源代码改变MySQL中的账号与密码。2）本机内存需要大于8GB(最低要求为8G)，因为需要训练模型，将大容量数据装载进内存，若不过8GB，请不要训练模型，直接使用提供好的模型。3）在运行前，确保python版本>3，安装代码所需要的所有依赖库。执行pip install -r requirements.txt

## **1.4 论文章节安排**

本论文主要是介绍基于微博评论的简单舆论分析的整个过程，主体分别五个章节：

（1）前言，介绍本课题相关的研究背景和意义。

（2）微博爬虫概述，主要详细讲述python爬虫每部分代码的撰写，分析其思路的实现。

（3）针对基础微博数据，选取适合的第三方内库进行建模。

（4）针对抓取后的微博数据，使用以上的建立好的模型对评论数据进行分析，画图。

（5）总结与展望，阐述事件背后民意的方向。结合自己所作作品，总结出存在的不足之处与改进方向。

# 2. 微博爬虫概述

为了能够更好地实现微博爬虫的进行，对微博反爬虫策略进行对抗，我们要先对微博平台进行全面仔细的分析，然后选择采用哪种手段进行爬虫数据获取。

## **2.1 微博平台全面分析和选择**

微博作为社交大平台，自然有许多个可用的站点，是为了让更多设备所支持。目前可知的微博站点，一共有三个，分别是网页端、手机端以及微端，这三个站点的抓取数据的复杂程度是逐渐降低的，很显然，若要抓取更华丽更庞大的数据量，肯定要在更复杂的站点取爬取数据。下面是三个站点的界面介绍。

微博网页端: **https://weibo.com**



微博手机端: **https://m.weibo.com**



微博微端: **https://weibo.cn**



针对以上三个站点，我们的需求是抓取博主的每条微博下的评论，具体到事件，昵称，评论内容等。微博的网页端是最复杂的，但却是最能满足我们的需求，可以爬取各种各类的数据，还能进行高级搜索等操作。对于m站(微博移动端)，爬取微博评论可以使用Ajax异步加载请求获取JSON格式数据解码获得数据，但是很遗憾，微博评论中的JSON地址不可访问，且构造页数时无法入手。针对微博微端，搜索功能极其有限，且只能访问用户部分数据。最后我们选择了最复杂最难上手的站点微博网页端进行数据的挖掘和分析。

## **2.2 微博评论分析**

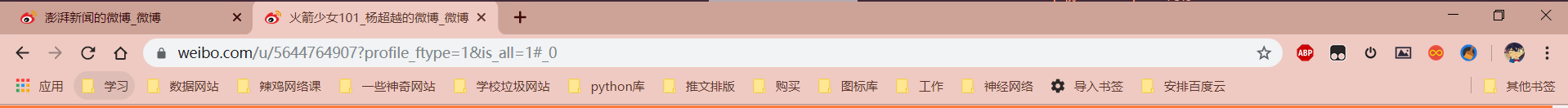
以“澎湃新闻”博主的微博为例子，要去看微博中的评论如何构建和要获取微博评论中哪些数据。从以下微博评论截图中，我们可以看到，网页评论绑定在“评论”的这个按钮中，且继续往下滑动时，地址栏没有发生变化，可知道网页的加载的数据是JSON数据，且通过Ajax异步加载数据。

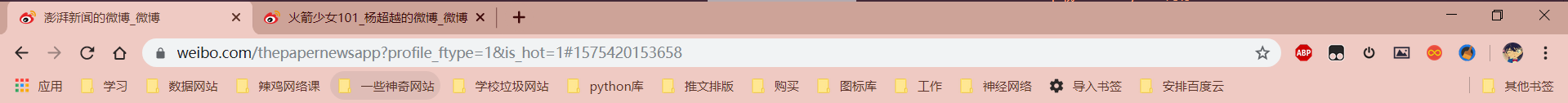


了解了评论区的网页构建方法，我们的需求是获取微博评论中有价值的信息。所以最后我们选取评论区中微博评论内容、评论人、评论头像以及评论时间进行爬虫获取。知晓了要抓取的目标和网页的构成，我们就可以开始构建爬虫代码了。

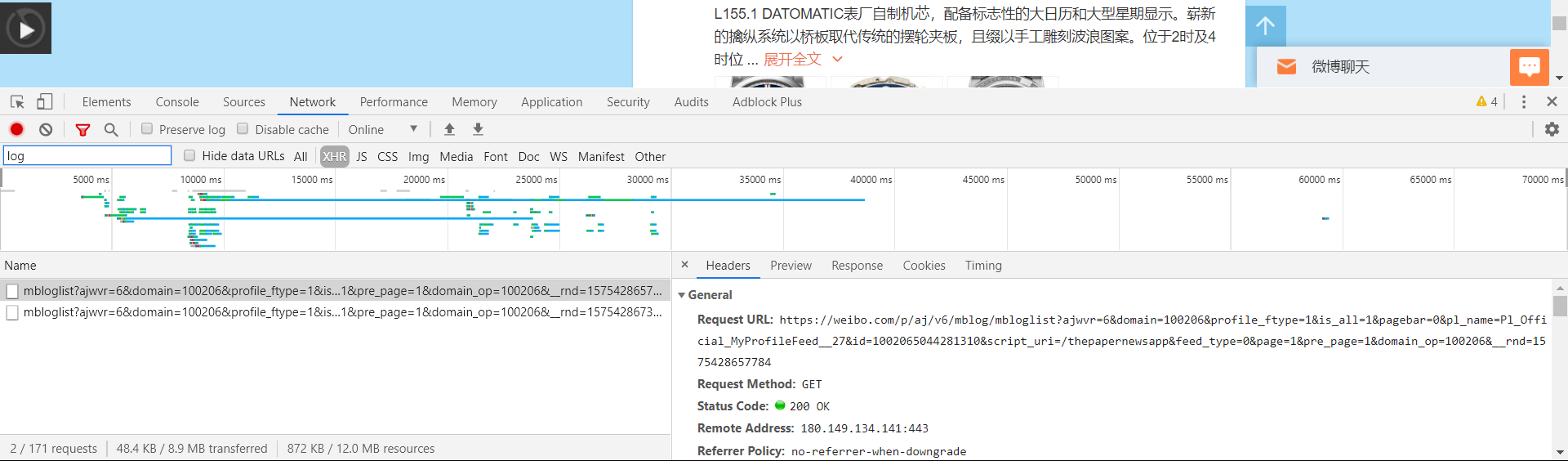
## **2.3 爬虫代码实现与分析**

首先查看各个博主的主页地址，发现有两处不一样。以下以“澎湃新闻”和“杨超越”的微博作为例子。不难发现，地址栏有明显的不同，为什么同是微博用户，地址构成不一样呢？仔细观察，发现微博id号可以修改的，若被用户修改后，地址栏优先显示用户所设置的昵称。且微博在他们设定特定的id时，会去修改一个特定一个值，为domain值，以此用来区分用户。





接下来我们可以查看微博评论的网页加载规律。首先明确，数据由json加密，加载每页评论由Ajax异步加载所得。打开浏览器中的“检查”功能，查看其“网络”项，选择“XHR”项，因为此项是所有Ajax请求的网页。



可以看到网页构成，并且以page为变量，抓取一直加载后博主微博，网页为: <https://weibo.com/p/aj/v6/mblog/mbloglist?ajwvr=6&domain=100406&is_all=1&page=1&pagebar=0&pl_name=Pl_Official_MyProfileFeed__20&id=1004065644764907&script_uri=/u/5644764907&pre_page=1>，我们用此网页进行访问，并且进行JSON数据解析，可以得到博主微博详情。(解析JSON数据使用在线解析: www.json.cn)



网址结构由domain、uid和id组成，所以我们要先去构建爬取这三个值，将其进行字符串的拼接，得到网址，并解析其JSON数据，利用正则表达式获取微博文本内容。评论区也是如此构建爬虫代码进行抓取。以下我们来详细解释代码的实现和过程。

考虑每个用户所要抓取的用户微博不同，并且需求的页数不同，我们设置了程序入口，用户可以输入自定义的微博名字和爬取的页数。首先根据网址的规律，我们首先要去获取用户特定的uid和domain，在微博搜索中搜索用户名可以在源代码中找到相应的uid。以搜索“澎湃新闻”为例，打开开发者工具，查看源代码，搜索发现出现其uid，这样我们可以获取其源代码转文本，用正则表达式筛选文本，找到所需要的内容。



图1 源代码出现uid信息



图2 代码构造正则表达式筛选

但是在此页面无法获得domain值，所以我们猜想domain的值只有博主主页才会有此信息。此又会遇到问题，如果使用没有携带用户的uid的headers无法进入主页从而无法获取源码的文本内容。避免代码的重复，我将代码写入爬虫类中，以此共享headers值，获取domain的值和上面方法一样。

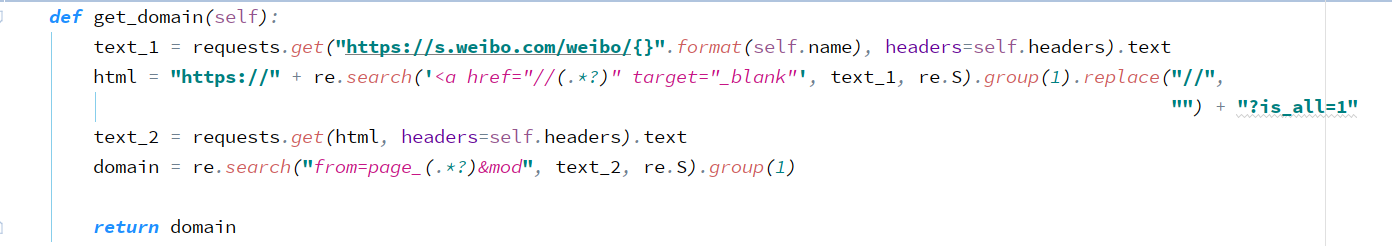


图 构造正则表达式筛选domain值

接下来，我们来分析首页中微博以及微博评论的地址构成。由上述的Ajax请求分析到，首页由id等构成，那我们可以在页面获取id值，并填入Ajax请求中，获取博主更多微博内容。我们用home\_url来存储所有首页能获取的微博地址。



图 通过源代码获得name\_id并填充地址



图 解析首页获得微博博主所有微博地址

我们来观察微博评论的Ajax构成，和首页的构成差不多，但是要携带的信息更多，不止需要特定的uid，还需要domain值和id值，所幸id的值是domain值和uid值直接拼接，容易获取。查看Ajax请求预览中，可以明显看到评论数据就夹杂在html中，此时我们就可以使用xpath语法选择评论文本内容，并写入文本文件中。此时我们所要做的第一步工作完成了，获取数据的时间有点长，建议获取对自己有益的数据。

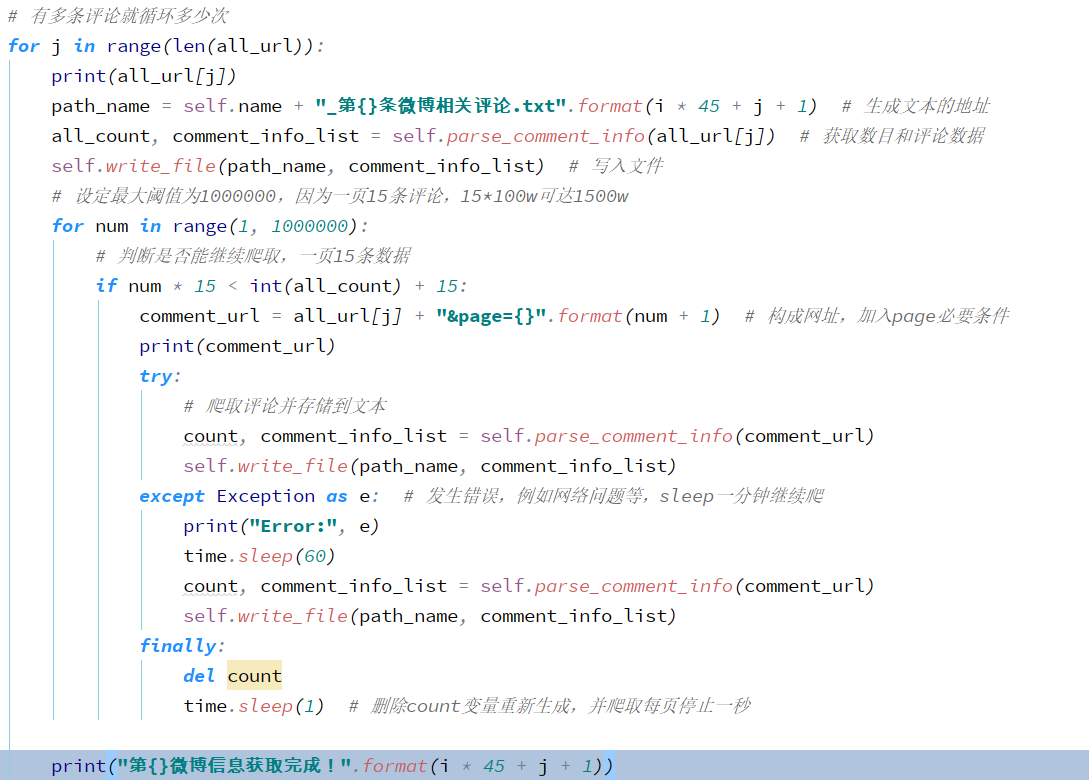


图 确定评论条数并爬取到文本文件中

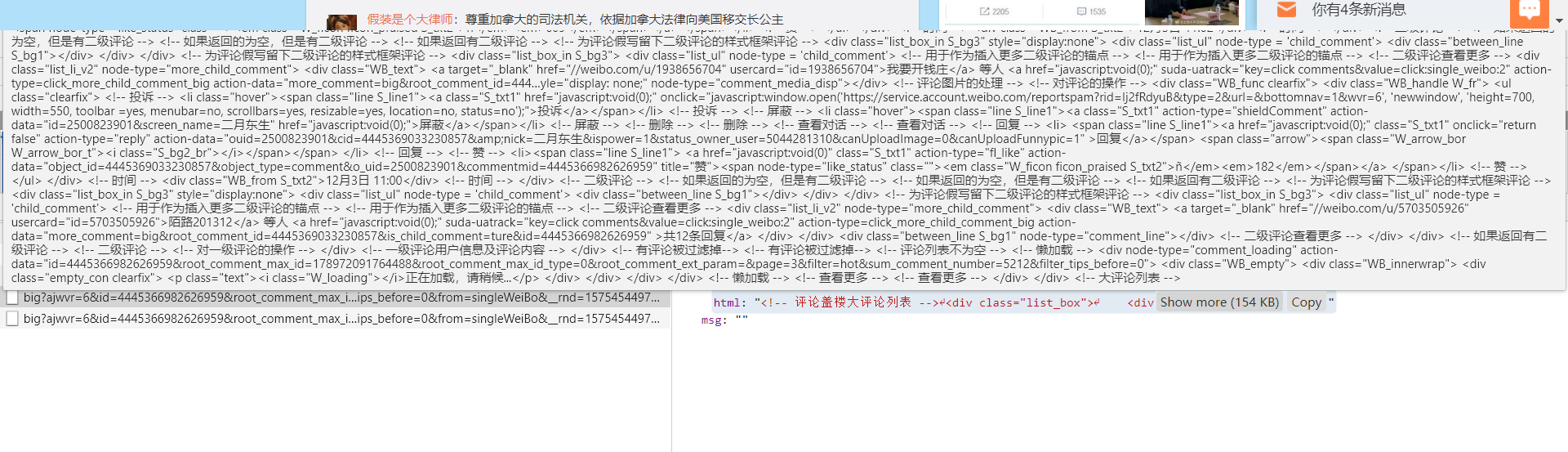


图 在preview中可以看到评论数据隐藏在html中



图 利用xpath语法筛选评论内容

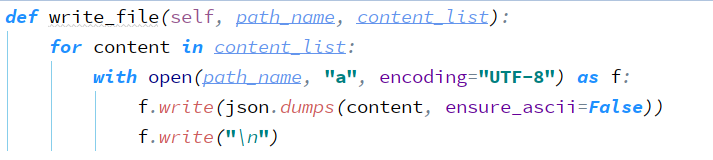


图 直接将获取的文本内容写入到文件中

## **2.4 小结**

通过对微博评论的爬虫代码的实现与分析，我们可以通过此来知道Ajax异步加载的网页是怎么获取数据，同时还要对网站的反爬机制做出一定的措施，例如爬取一定数量停止爬取几秒，或者设定多几个代理。只针对微博平台，它只封ip，不封账号，只要代理多，可以使用自己的微博cookie来测试代码都没有问题。

# 3. 中文数据的情感分析

情感极性分析是对带有感情色彩的主观性文本进行分析、处理、归纳和推理的过程。按照处理文本的类别不同，可分为基于新闻评论的情感分析和基于产品评论的情感分析。其中，前者多用于舆情监控和信息预测，后者可帮助用户了解某一产品在大众心目中的口碑。

目前常见的情感极性分析方法主要是两种：基于情感词典的方法和基于机器学习的方法。第一种方法是使用第三方库来训练模型，通过情感打分的方式进行文本情感极性判断，score > 0.5判断为正向，score < 0.5判断为负向。第二种方法是机器学习方法，使用分词切分数据集进行词向量训练模型，达到情感分析的目的。

## **3.1 基于情感词典的方法分析**

基于情感词典的方法是基于第三方库来对特定数据的训练得到特定的模型处理特定的数据，现在处理英文语言有很多库，因为英文的分隔符是空格，能够以空格为分隔符来判断词性。但是中文的语句的分隔符是没有的，只能通过分词来单独判断词性。

Python的第三方处理中文语言的库很少，以下以SnowNLP优秀库为例，来实现基于情感词典的方法处理中文语言。SnowNLP是一个python写的类库，可以方便的处理中文文本内容，是受到了TextBlob的启发而写的，由于现在大部分的自然语言处理库基本都是针对英文的，于是写了一个方便处理中文的类库，并且和TextBlob不同的是，这里没有用NLTK，所有的算法都是作者实现的，并且自带了一些训练好的字典。

因为snownlp这个库首先是以购物评论为基础数据来训练模型，不适合我们对微博评论的情感分析，所以我们需要以微博评论为基础数据进行模型的训练，并且替换原来的模型，使用新的模型来判断评论的情感。

## **3.2 基于机器学习的方法分析**

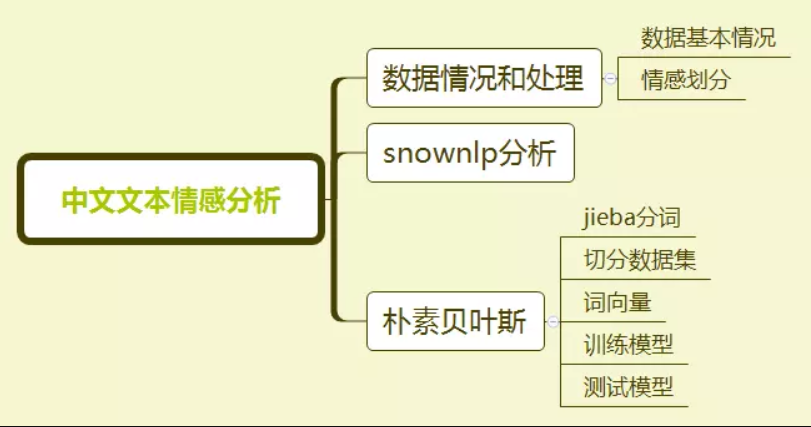
机器学习的方法和内库在现如今很多，包括sklearn、tensorflow等，使用python来训练深度学习的模型很简单。

针对论题中的微博评论，我们实现朴素贝叶斯模型用来分类微博评论的情感，以中文分词，对基础数据进行划分情感值，积极的值为1，消极的值为0。以下使用sklearn库实现机器学习中的朴素贝叶斯模型，进行对数据的分析。

朴素贝叶斯分类（NBC）是以贝叶斯定理为基础并且假设特征条件之间相互独立的方法，先通过已给定的训练集，以特征词之间独立作为前提假设，学习从输入到输出的联合概率分布，再基于学习到的模型，输入x求出使得后验概率最大的输出y。

## **3.5 小结**

本章介绍了实现中文数据的情感分析的两种方法，从而我们了解到具体的算法的实现步骤，能对中文的数据进一步剖析。以下为具体代码实现的流程。



# 4. 建立中文情感分析模型

根据建模的基本流程，并对各个实现步骤进行设计，则可开始实现各个模型对文本的情感分析。以下文本数据采用第二章中爬虫所获得“澎湃新闻”的四条有关于华为251事件的报道中的评论为例子，构建模型。



图 文本数据展示

## **4.1 基于情感词典的建模**

首先我们要去获取基础微博评论数据，snownlp这个库是专门为电商所设计的库，预测我们的微博评论准确率不高。利用第二章所得的爬虫技术可以得到大量数据，若将其存入mysql数据库，也可得到大量基础评论数据。考虑时间成本和微博防爬机制，我们要站在前人的肩膀上，直接获取前人所获得的数据。数据来源于: [百度网盘](https://pan.baidu.com/share/init?surl=eSeXh5K) => 密码：tvdo。

这个500w数据是存储在mysql数据库中，导入请按照数据说明导入，也可以使用mysql导入: mysql –u用户名 –p密码 <weibo500w.sql。对微博语料进行情感分类，是基于原有SnowNLP进行积极和消极情感分类和训练。得到新的模型数据。基准为>0.8为积极评论，<0.3为消极评论，进行二分类。

得到模型

预测文本情感

基于500w微博评论数据

训练

snownlp模型

图 SnowNLP模型训练过程

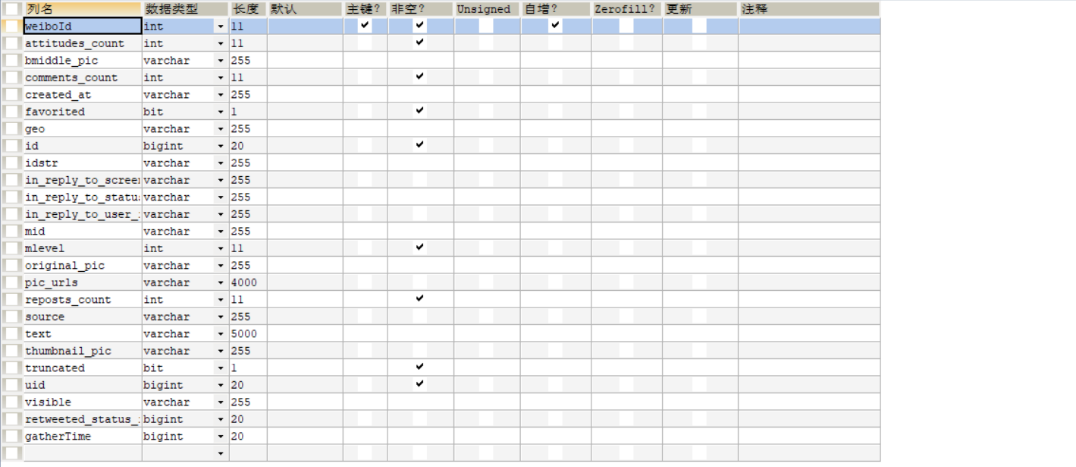


图 mysql微博语料中表结构

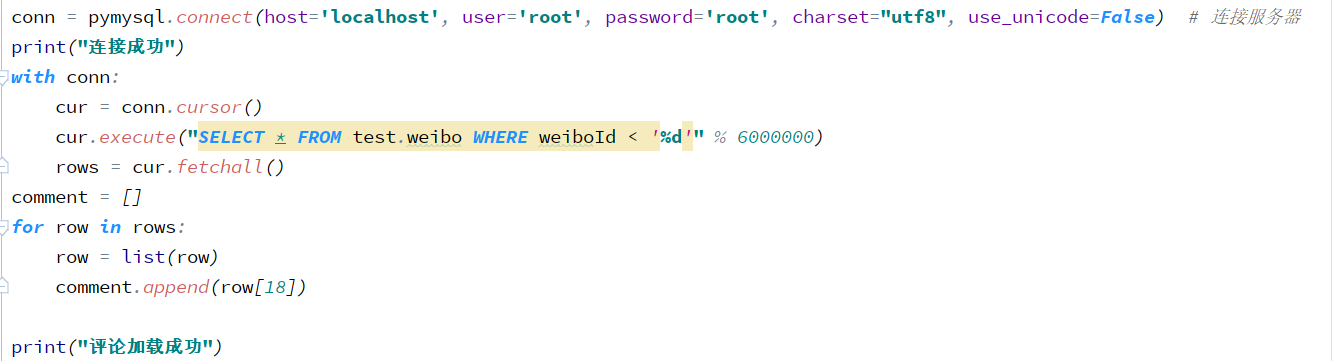


图 连接数据库，将19列数据获取



图 将数据导入SnowNLP模型进行训练并获得新模型

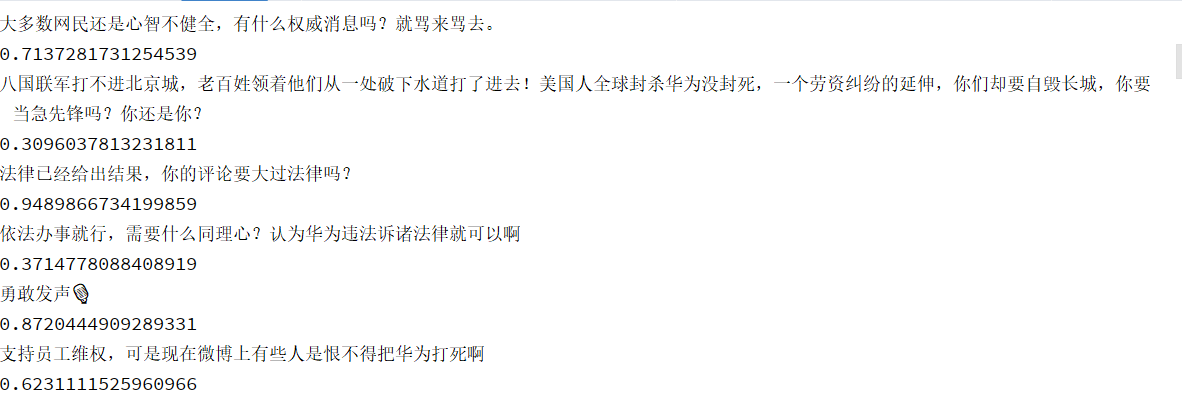


图 以文本中的数据做例子，模型给出的每句话的分数

训练完成后会生成sentiment.marshal.3模型文件，将python位置/Lib/site-packages/snownlp/sentiment/中sentiment.marshal.3直接替换，训练可以进行多轮训练，精度会更好(但是不建议)。使用模型的方法也很简单，直接将文本文件导入模型中，就会自动分析每句话的情感积极度。注意，运行内存小于8G的计算机最好不用训练模型，因为此方法是将数据全部倾倒在内存里，500w数据和程序运行所要的内存至少6GB，而且训练模型的时间成本实在太高了。可以直接使用附件中已经训练好的模型

## **4.2 基于机器学习的方法**

相较于以上的方法，机器学习的方法更加准确更加有针对性，因为可以用不同的数据，只要有置信度，就一定可以以此作为目标训练，以此达到最高的精度。以下我们以50w消极及50w积极.csv(csv文件最大容纳100w左右数据)中的数据作为参考数据，里面一共两列，第一列为评论内容，第一列为评论所获得的评分，1为积极，0为消极。数据的来源于上节所生成的两个区分好的文本文件，分别将此打分，生成csv文件

首先，我们对评论数据分词。为什么要分词了？中文和英文不一样，例如：i love python，就是通过空格来分词的；我们中文不一样，例如：我喜欢编程，我们要分成我/喜欢/编程（通过空格隔开），这个主要是为了后面词向量做准备。没有进行词性分词，就会识别不出每个词的置信度，从而判断不出来整句话的置信度。

接下来我们就可以切分数据集，以第一列评论数据作为特征，以第二列分数作为标签，按照8：2切分训练集和测试集。然后就用sklearn中自带的朴素贝叶斯算法训练，得到最后我们的机器学习模型。

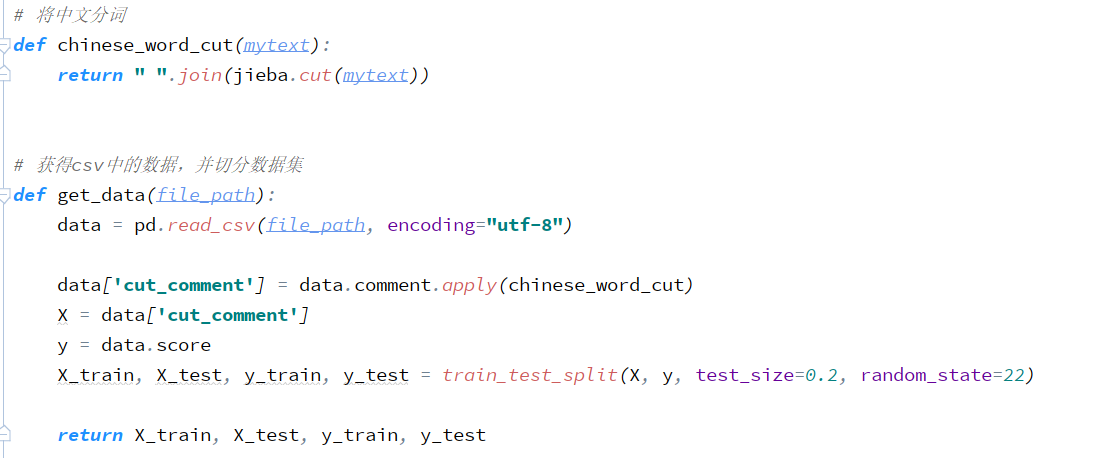


图 将数据集中文分词以及划分数据集

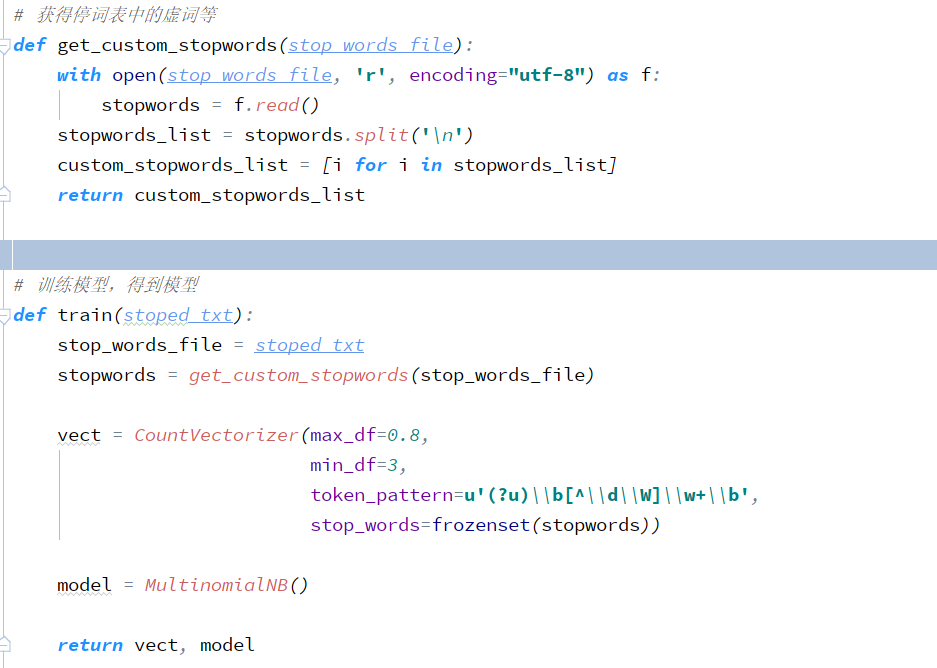


图 利用停词表去除无用虚词等，训练模型

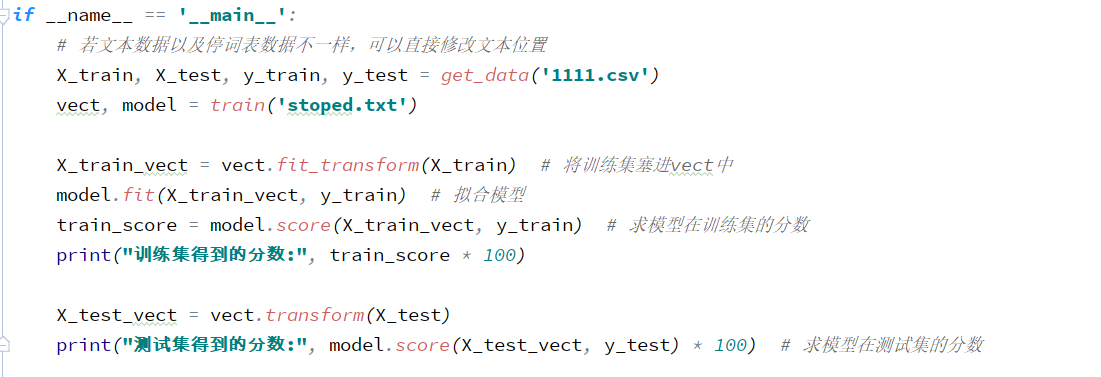


图 将数据加载模型中并测试模型的分数



图 最后模型跑出来的结果

## **4.3 分析结果**

我们拥有四条微博评论的数据，首先手动将它合并到同一个txt，方便余下处理，一共1w多条。用人眼去识别每句语句的情感度数值，显然不合理，所以我们采用条形图的形式，这样更加直观更加容易分析。



图 获得文本数据得出情感度分数进行条形图绘制

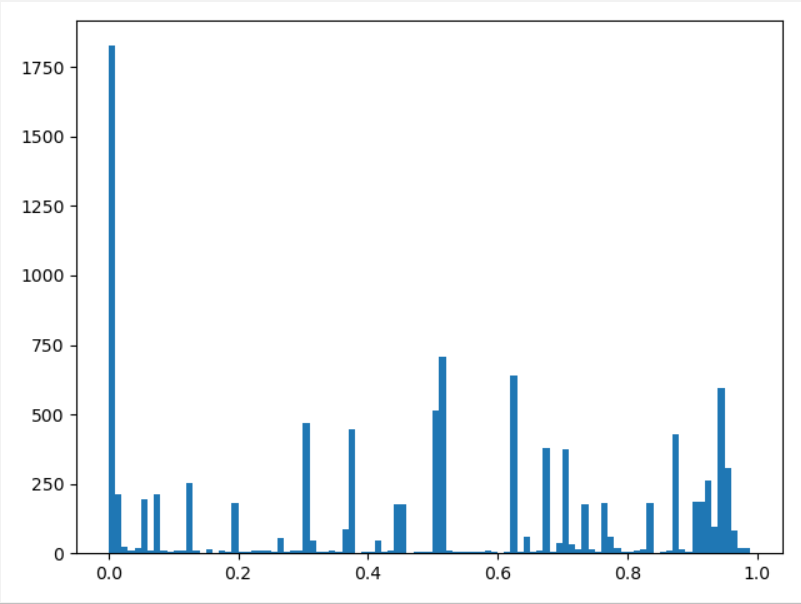


图 根据所得到的1w多个评分所做统计图

当评分>0.5时，语句被认为是积极的，当评分<0.5时，语句被认为是消极的。X轴为分数，y轴为所得分数的人数。在此统计图中可以明显看出消极态度比积极态度人数多。

由此联系事实，华为251事件引起社会轰动，谁也没有想到一家和爱国联系在一起的企业会因为这起事情而遭到社会谴责。事因华为员工被辞退拿到赔偿，无缘无故被华为控告遭到251天拘禁。就此微博对这事的报道以及民众对此事的态度，无疑都是消极的声音更多。除了部分华为水军灌水之外，如今的舆论都偏向于华为道歉这边。

接下来，我们来看看民众对这事的具体评论出现频率最多的词。可以使用jieba分词中的词频统计，由于去掉虚词等无用词后，统计的结果不尽如意，分词的数目实在太多，无法一一统计。我们采用词云这种方式，让最重要的词出现在图案中，以及是我们一目了然。



图 生成词云的详细参数设置



图 词云最后生成的词图

在词图中，我们可以明显看到“员工”，“法律制度”，“可笑”，“华为”等词。在这里，我们可以更加明显看到民众对于华为251事件的态度，消极态度占了半个图，当然中立态度也有。大部分的矛头指向于华为公司的不作为不道歉不解释，这很令人心寒，员工为公司辛勤劳作12年，最后被公司控告落狱251天。虽然华为可以在一定程度代表中国企业的成功，但是这不能意味着公司要丢掉最基本的人性道德，最后希望华为公司与当事人能给予我们最后的真相吧。

## **4.4 小结**

本章实现了中文情感分析两种方法，其中建立的两个模型，通用性很强。只要将数据改成自己所要分析的数据，模型就能给予总体的分数。若换为淘宝京东等电商的购物评论，依旧可以使用第二章所用到的爬虫技术抓取相对应的数据，然后对数据使用这一章知识进行建模，得到专属于购物评论的模型。总体而言，效果基本达到预期。

# 5. 总结与展望

回顾本课题的研究过程，通过参考相关的文献资料，了解本课题的python知识点，由此找到研究方向。通过归纳python爬虫技术与python机器学习知识，了解到数据的来源与建模的方向，使用现有的中文库来对各条微博评论进行情感性分析。程序最后生成的模型可以用来预测微博用户的下一条评论，得到综合分数。最后，基于微博评论的简单舆论分析，效果基本达到预期。

在这个实验中，主要使用了两个算法实现建模，数据来源使用爬虫技术获得。获取微博评论技术，主要是网络与网页加载的知识，从微博网页中找到其加载方式以及数据的存储位置，使用第三方库进行抓包解密解码等操作。建模使用两个算法--贝叶斯算法和朴素贝叶斯算法，两者分别被封装在snownlp和sklearn库中，后者算法是前者的进一步深化，比前者更稳定更有效，所以机器学习建模要比snownlp要更准确一点。以华为251事件来引出这次实验的目的，结合现实，代码更有说服力。

虽然基于微博评论简单舆论分析效果基本达到预期，但代码仍存在许多不足之处。例如，未能针对微博的反爬机制做出很好的策略，未能真正获取所有的微博评论，未能将模型做到更加通用，未能多使用算法进行建模，未能提高模型的精确度到95%。

在此实验中，出现了很多错误和不足。例如爬虫爬取的数据中，不能获取所有的数据，而且还爬取很多空值。这些空值直接影响到建模的进行，中文分词无法对空值进行切分。基于机器学习得到的模型，无法对一句话进行情感度分析。500w微博评论数据太落后于时代了，且在这份语料中，竟然没有置信度分数。最后只能徒手将snownlp所构建的模型得到的文本提取100w数据进行置信度区分，用来进行机器学习训练模型。

总体而言，本课题研究的基于微博评论的简单舆论分析虽然仍存在不足的小细节，但主要功能都已经得到了实现。接下来就可以基于这个小模型开发一个小型推荐系统，针对数据做出更大的文本数据挖掘，使用更为先进的深度学习进行神经网络训练更加复杂更加有效的模型。

# 参考文献

1. SnowNLP：处理中文文本内容 - a123456ei的专栏 - CSDN博客
2. 基于微博信息的关键词库爬虫策略，林红静;黄梦醒 - 《海南大学学报：自然科学版》- 2016
3. 分段卷积神经网络在文本情感分析中的应用 , 杜昌顺;黄磊 - 《计算机工程与科学》- 2017.
4. 中文微博情感分析方法研究，王银;吴新玲 - 《广东技术师范学院学报 》- 2014
5. 情感分析——深入snownlp原理和实践，https://blog.csdn.net/google19890102/article/details/80091502.
6. Python有趣|中文文本情感分析. <https://www.jianshu.com/p/b775f978b004>