北京工业大学

2021-2022学年第一学期

信息学部（软件学院）

|  |  |  |  |
| --- | --- | --- | --- |
| 课程名称： | 大数据技术导论 | | |
| 报告性质： | ■作业报告 □实验报告 | | |
| 学号： | 20080320 | 姓名： | 吴祥曦 |
| 任课教师： | 李建强 | 课程性质： | 专业选修课 |
| 学分： | 2 | 学时： | 32 |
| 班级： | 200802  200803 | 成绩： |  |
| 小组成员： | 罗星悦 常兴阳 李天依 王可宣 | | |
| 教师评语： |  | | |

2022年11月28日

目录

[一、动机Motivation 4](#_Toc120551139)

[二、背景Background 4](#_Toc120551140)

[2.1 项目背景Project Background 4](#_Toc120551141)

[2.2 相关工作Related Work 4](#_Toc120551142)

[三、实现过程How 4](#_Toc120551143)

[3.1 爬虫部分 4](#_Toc120551144)

[3.1.1 顶置贴爬取 5](#_Toc120551145)

[3.1.2 非顶置贴爬取 5](#_Toc120551146)

[3.1.3 贴子回复爬取 5](#_Toc120551147)

[3.1.4 无用信息过滤 6](#_Toc120551148)

[3.2 数据处理部分 7](#_Toc120551149)

[3.2.1 数据预处理： 7](#_Toc120551150)

[3.2.2 使用数据时的处理 7](#_Toc120551151)

[3.3 词云 7](#_Toc120551152)

[3.3.1 获得制作词云的关键词 7](#_Toc120551153)

[3.3.2 制作词云的背景图与配色 8](#_Toc120551154)

[3.3.3 根据词频生成词云 8](#_Toc120551155)

[3.4 情感分析 8](#_Toc120551156)

[3.4.1 训练预测判断 8](#_Toc120551157)

[3.4.2 处理数据保存导出 9](#_Toc120551158)

[3.4.3 绘制情感分析饼状体 9](#_Toc120551159)

[3.5 语句相似度比较 10](#_Toc120551160)

[3.5.1 语句相似度数据处理 10](#_Toc120551161)

[3.5.2 贴子和评论间相似度比较 10](#_Toc120551162)

[3.5.3 计算并输出贴子与评论，评论与评论间的相似度 12](#_Toc120551163)

[3.6 单个贴吧内单独帖子回复数量，回复时间统计 13](#_Toc120551164)

[3.6.1 正则查找爬取下来网站的回复时间 13](#_Toc120551165)

[3.6.2 统计并做出发帖和回复时间的散点图 15](#_Toc120551166)

[3.7 图形界面设计 16](#_Toc120551167)

[3.7.1 pyqt5简介 16](#_Toc120551168)

[3.7.2 pyqt5的环境配置和使用流程 16](#_Toc120551169)

[3.7.3 Ui界面交互设计 17](#_Toc120551170)

[3.7.4 交互多线程调度 18](#_Toc120551171)

[3.7.5 程序运行效果图如下 22](#_Toc120551172)

[四、评估分析Evaluation 22](#_Toc120551173)

[4.1 数据分析精确度评估 22](#_Toc120551174)

[4.1.1 分词精确度评估 22](#_Toc120551175)

[4.1.2 情感分析精确度评估 23](#_Toc120551176)

[4.1.3 语句相似度分析精确度评估 24](#_Toc120551177)

[4.2 项目优缺点评估总结 25](#_Toc120551178)

[4.2.1 项目优点 25](#_Toc120551179)

[4.2.2 项目不足 25](#_Toc120551180)

[4.3 组员自我评价 26](#_Toc120551181)

[4.3.1 吴祥曦 26](#_Toc120551182)

[4.3.2 李天依 26](#_Toc120551183)

[4.3.3 罗星悦 26](#_Toc120551184)

[4.3.4 常兴阳 27](#_Toc120551185)

[4.3.5 王可宣 27](#_Toc120551186)

# 一、动机Motivation

平时逛贴吧也总是留意北工大吧的消息，很多最新资讯和通知也可以在贴吧里找到。与此同时也可以和小伙伴们一起畅谈学校里发生的趣事，找到自己青睐的小伙伴。但是在贴吧找信息不是很方便，因为同学们发的帖子很复杂，很可能一个话题中的不同评论聊的都不一样，有的时候在搜索栏搜索并不能找到自己想要的信息。这样一来，帖子的成分分析就至关重要了，因此我们希望运用大数据分析的理论知识，结合软件工程的优势，对贴吧的帖子和评论进行数据化分析以提高使用者的体验。

# 二、背景Background

## 2.1 项目背景Project Background

大数据作为时下最火热的IT行业的词汇，随之而来的数据仓库、数据安全、数据分析、数据挖掘等等围绕大数据的商业价值的利用逐渐成为行业人士争相追捧的利润焦点。随着大数据时代的来临，大数据分析也应运而生。结合大数据分析的方法，对贴吧数据进行爬取分析。

## 2.2 相关工作Related Work

对于贴吧数据抓取并分析这个话题，已有许多前人进行过类似的研究、相关程序开发以及分析，其中不乏有相当精细且深入的结论，本文参考前人经验并于其基础上进行了学习和创新。

[1]盛伟翔,扶齐彦,王迪.基于百度贴吧的当代大学生价值观评价系统设计与开发[J].电脑编程技巧与维护,2021(09):10-12.DOI:10.16184/j.cnki.comprg.2021.09.004.

[2]王颖. 面向高校贴吧发帖的文本分类及应用研究[D].扬州大学, 2021.DOI:10.27441/d.cnki.gyzdu.2021.001835.

[3]孙美凤,宋晨,王颖.基于PHP的百度贴吧数据爬取[J].软件,2020,41(11):23-26.

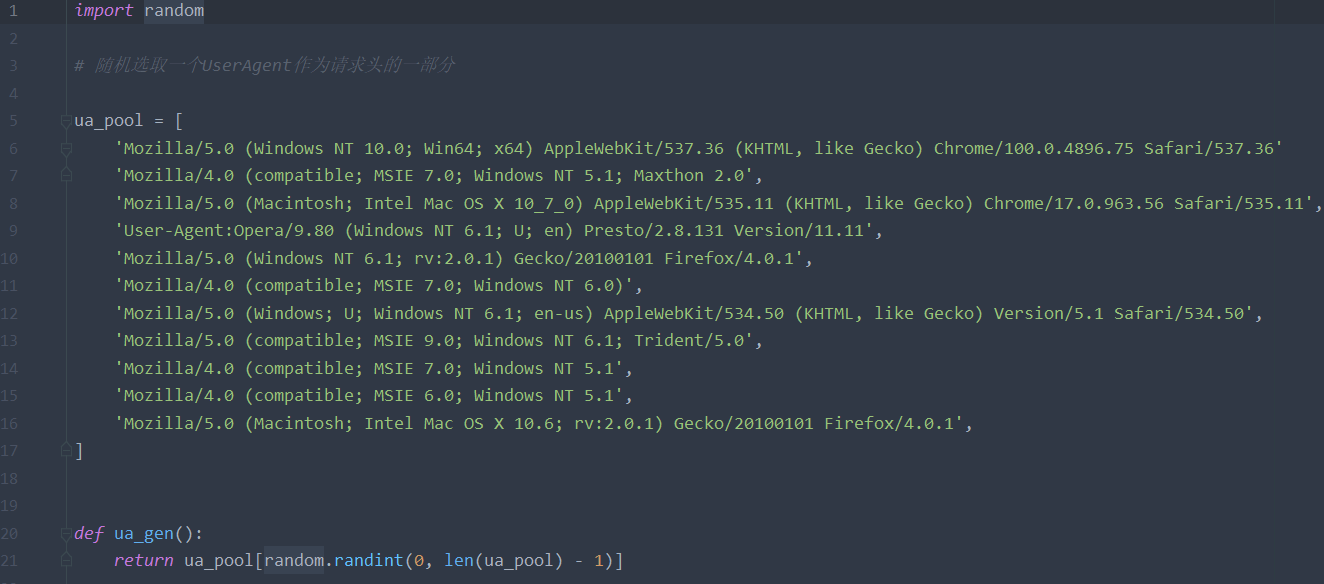
[4]李金海,胡旭.基于百度贴吧的高校网络舆情热点话题分析[J].计算机与现代化,2020(09):12-18.

[5]耿辉,马茂,张勇,尹小妹,徐安定,吕军.基于爬虫技术获取新型冠状病毒（2019-nCoV）贴吧数据的网络舆情分析及应对策略[J].医学新知,2020,30(02):107-110.

# 三、实现过程How

## 3.1 爬虫部分

本大作业中的爬虫部分利用了requests，urllib等库，通过request直接向网页发起请求并通过urllib将其转换为.html文件读入程序内存（用于分析）并保存至本地（用于观察网页特征，便于功能的拓展），并通过beautifulsoup4库对爬取下来的.html网页进行格式化，便于程序的调试。



为了防止触发贴吧的反爬机制，我们将爬取间隔设置为1秒，并每次使用一个随机的UserAgent池作为请求头中UserAgent的值。

通过对贴吧特定吧的首页观察，我们找到了帖子固定的格式满足如下两种正则：

### 3.1.1 顶置贴爬取

对于置顶帖，满足：

<i class="icon-top" alt="置顶" title="(.\*?)"></i>

其中title=后对应的组为帖子标题。

### 3.1.2 非顶置贴爬取

对于非置顶帖，满足：

r'<a rel="noopener" href="(/p/.\*?)" title="(.\*?)" target="\_blank" class="j\_th\_tit ">.\*?</a>

其中href=后对应的组为帖子唯一的pid，作为帖子的唯一标识，title=后的组对应帖子标题。

对于非置顶帖，我们还应该爬取对于帖子下第一条回帖（即一楼）内容的概述，满足：

<div class="threadlist\_abs.\*?">(.\*?)</div

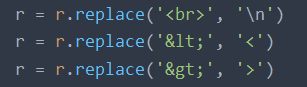
通过以上三条正则表达式，循环遍历当前页面下的所有内容，得到了当前页面所有帖子的集合。贴吧帖子的url设计较为简单直接，在https://tieba.baidu.com/后直接加p/与pid即刻访问至对应帖子，这也是上边要提取帖子唯一pid的原因。

### 3.1.3 贴子回复爬取

通过遍历访问所有帖子的pid，我们可以爬取到每条帖子下的全部回复。通过观察，贴吧帖子下的所有回复满足如下正则：

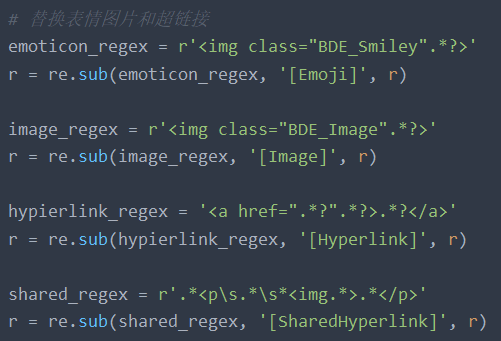
<div id="post\_content\_(.\*?)" class="d\_post\_content j\_d\_post\_content clearfix" style="display:;">(.\*?)</div>

其中post\_content\_后的组1是回复的唯一标识，最后一个后</div>中间的组2是回复内容。在回复的div中，由于是html的缘故，需要进行一些符号转换，例如回复中的<与>符号会被替换成&lt;和&gt;，于是需要进行如下的符号转换：



### 3.1.4 无用信息过滤

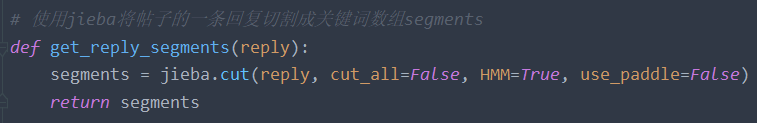
同时需要对回复中的图片、表情与超链接进行过滤（本大作业暂时不提取图片文件与超链接对应的网页内容），使用了如下的正则转换：

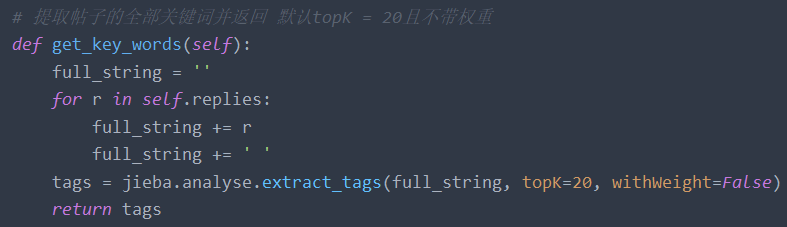


有些回复会具有背景，所以需要如下的正则表达式提取：

<div class="post\_bubble\_top".\*?div class="post\_bubble\_middle\_inner">(.\*?)</div>.\*?

其中组1的内容即为包含了html样式的回复内容，利用上述提取回复内容的正则表达式即可完成内容的提取。



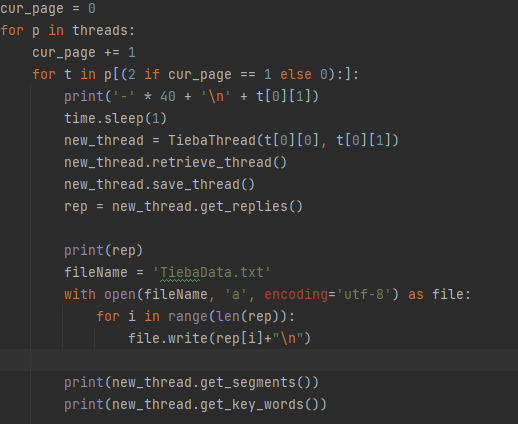


此时我们应用jieba切词库，在这里将帖子下每一条回复进行切割，得到一系列中文词语的集合，存于list中，再对其进行关键词提取，为后续的进一步分析与加工做准备。

## 3.2 数据处理部分

### 3.2.1 数据预处理：

基于我们前期完善的爬虫技术，获取了北京工业大学贴吧的前两页所有的贴子和贴子下的评论。但是由于网站反爬，爬取数据限制等因素，我们目前爬取获得的数据中图片image和表情emoji部分数据不能完全显示，所以我们决定对数据进行处理分为楼主和评论两大部分，并去除其中的image和emoji，将得到数据存储与.txt文本当中。



### 3.2.2 使用数据时的处理

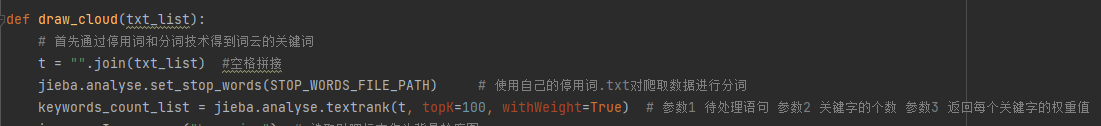
通过数据的预处理，我们已经得到了质量较高的文本数据。我们在使用数据时，通过去除回车和非中文符号的形式，最终得到可用于分析的完整数据。



## 3.3 词云

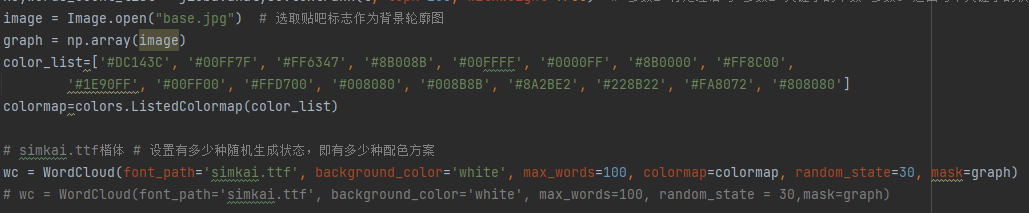
### 3.3.1 获得制作词云的关键词

根据前期爬虫，数据处理得到的数据，我们根据去停用词和jieba库分词，得到制作词云的关键词和其权重keywords\_count\_list，设定获取关键词中的前100个作为词云中的词汇



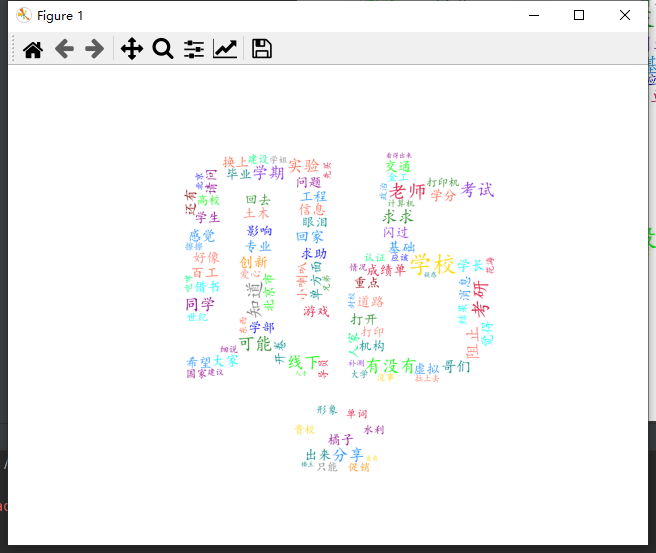
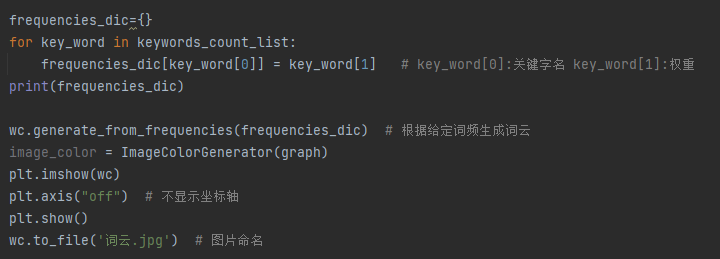
### 3.3.2 制作词云的背景图与配色

因为我们爬取的是北京工业大学吧的数据，所以我们选定词云的背景图为贴吧的“贴”字通过WordCloud库中的WordCloud函数，设定词云中字体为楷体，背景颜色为白色，词的个数为100，通过随机状态设定字体的颜色



### 3.3.3 根据词频生成词云

根据刚刚通过停用词和分词得到的的keywords\_count\_list，得到关键词的词频字典frequencies\_dic = {} 并根据frequencies\_dic生成词云



## 3.4 情感分析

### 3.4.1 训练预测判断

通过SnowNLP库，采用贝叶斯模型进行训练和预测，得出每条评论的情感值，我们设定若此数值大于0.53则判定其为积极情感，若小于0.47则判断为消极情感，在两值之间判定为中性情感



### 3.4.2 处理数据保存导出

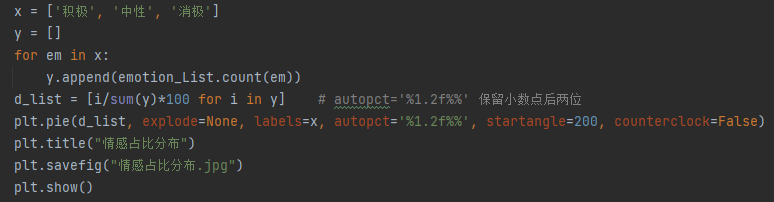
将每条评论得出的情感分析概率保存为.csv文件，通过pandas库将文本，情感指数，情感分为三列存储到.csv文件当中

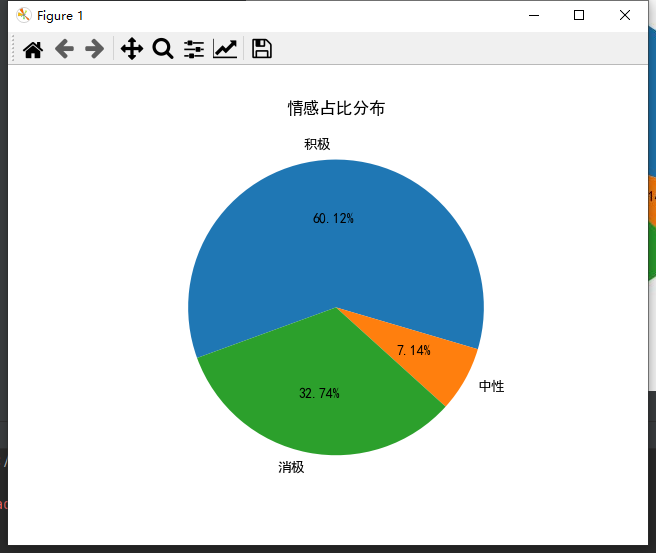




### 3.4.3 绘制情感分析饼状体

通过饼状图的形式，更直观的将情感分析结果进行展示

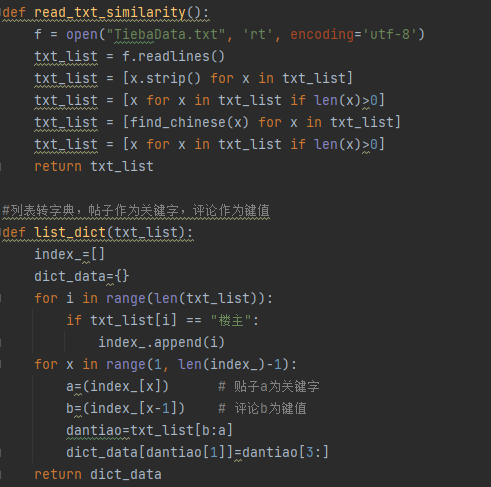




## 3.5 语句相似度比较

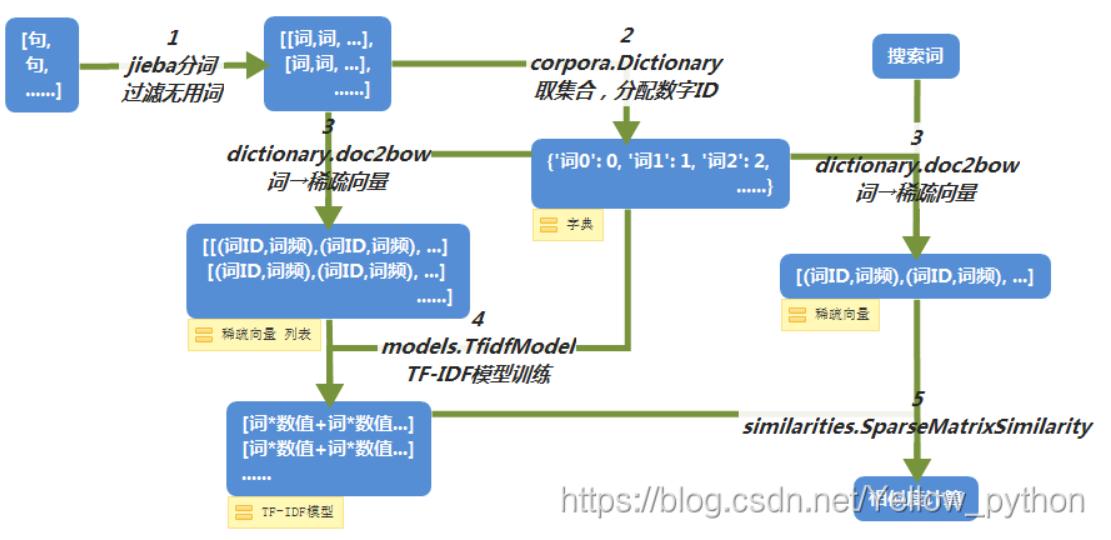
### 3.5.1 语句相似度数据处理

因为我们此次分析的是贴子和评论，评论和评论间的语句相似度。在分析相似度时，需要将每个贴子的楼主所发的语句作为基础语句，与下面的每条评论进行相似度分析，评论与评论间语句相似度分析同理。因此还需将数据的列表转换为字典，楼主的贴子作为关键字，评论作为键值

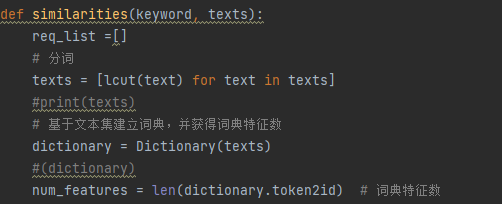


### 3.5.2 贴子和评论间相似度比较

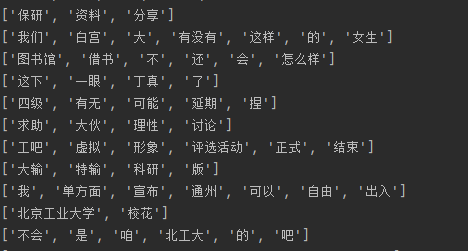
我们本次采用相似度分析使用的是gensim算法，gensim算法分析语句相似度的大体流程如下：



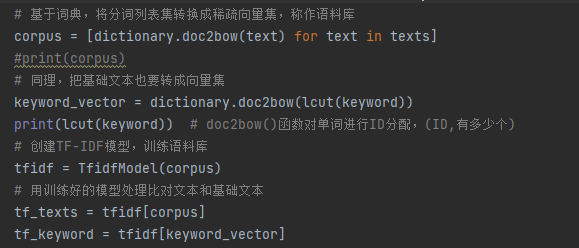
首先先将文本进行分词，并通过Dictionary函数基于文本建立词典



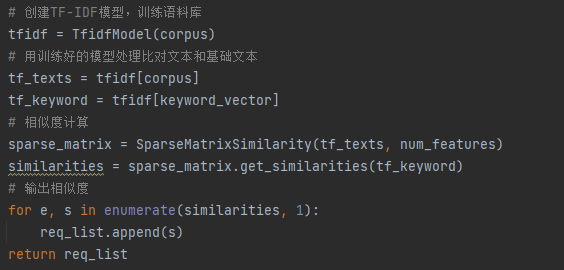
Gensim算法的核心是把文本进行主题（关键字）的提取，然后将主题转换为向量集，最终通过计算向量与向量间的余弦，从而得出相似度的值。将主题转换成向量集，用到的关键函数为doc2bow()次函数将对主题进行ID的分配，如下图所示，会将保研这一主题转换为【（0,1）】的向量用于后续余弦值的计算



稀疏向量集的转换如下：

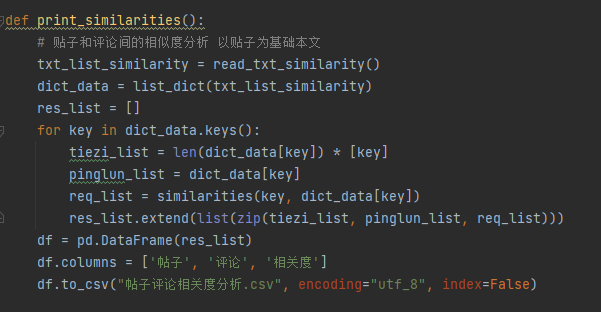


最后对两个稀疏向量合集进行TF-IDF模型的训练，在通过SparseMatrixSimilarity函数计算语句间的相似度



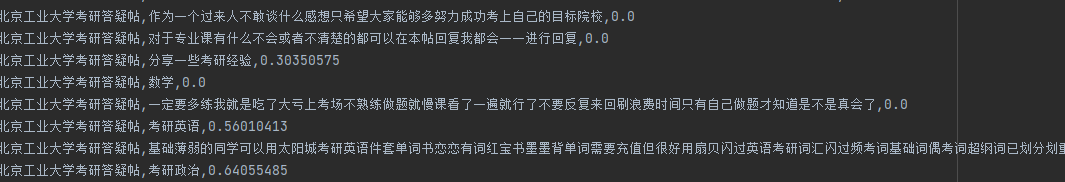
### 3.5.3 计算并输出贴子与评论，评论与评论间的相似度

每两条语句相似度的对比情况最终都保存在.csv文件当中



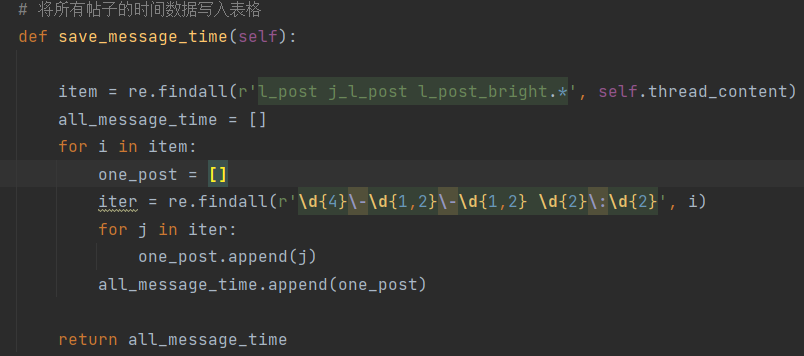


部分相似度分析如下：

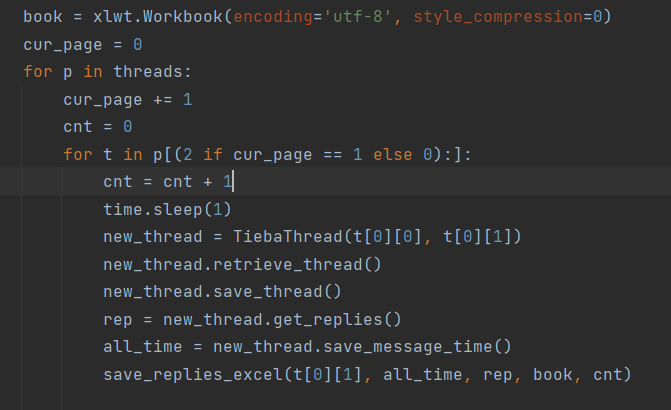


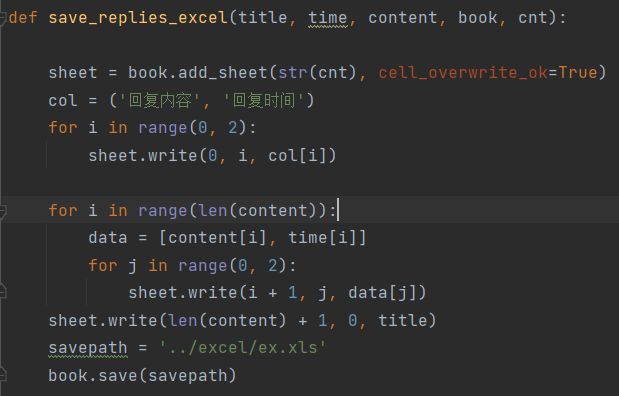
## 3.6 单个贴吧内单独帖子回复数量，回复时间统计

### 3.6.1 正则查找爬取下来网站的回复时间

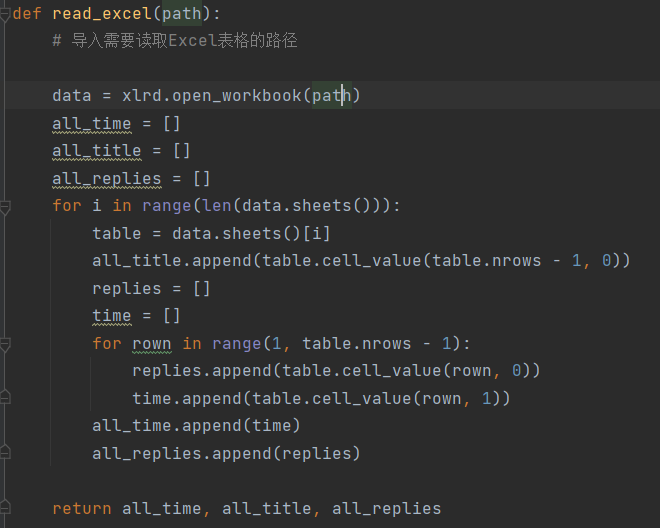


将数据储存到excel中



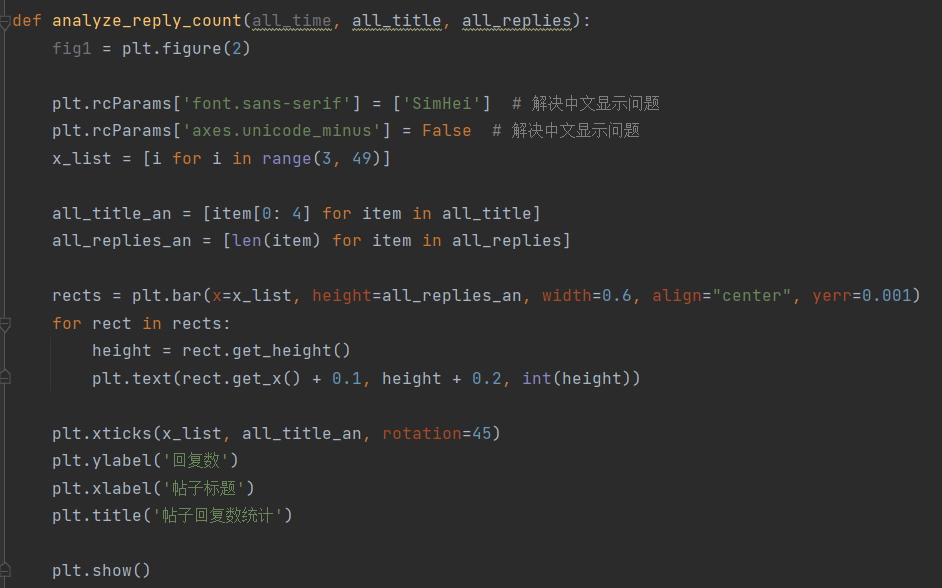


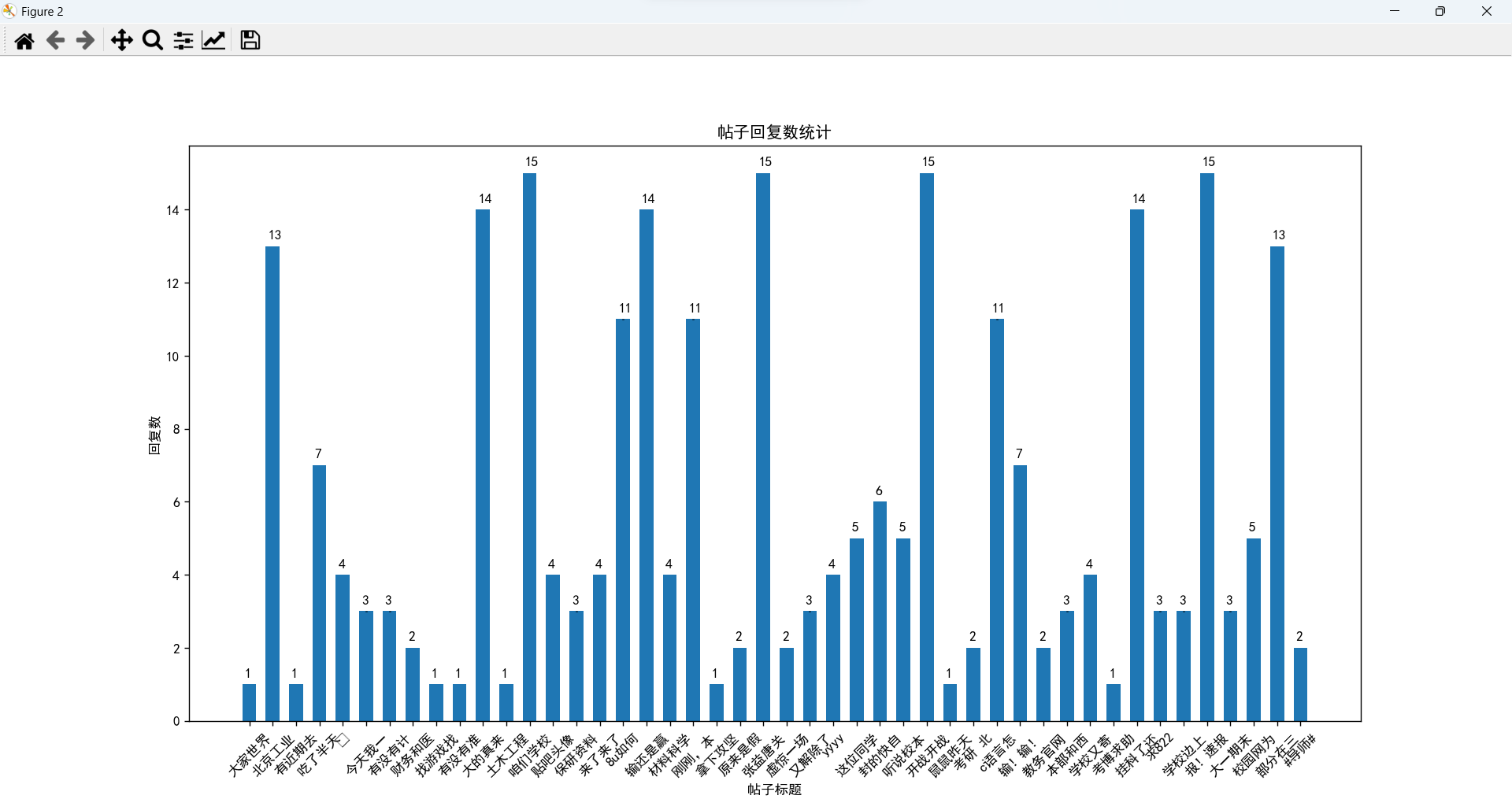
需要时从excel中读取数据



统计并做出回复数量的条形统计图

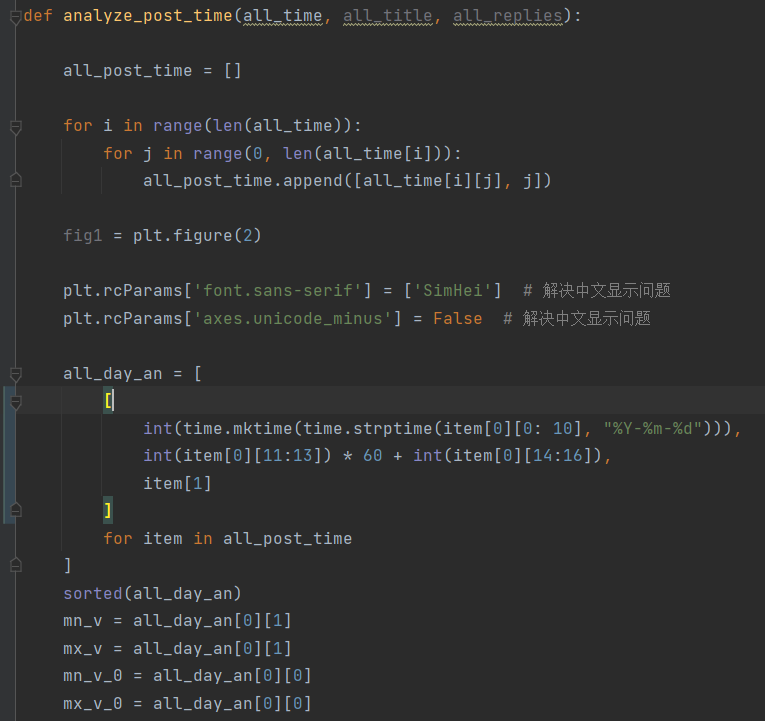
统计每个帖子下的回复数量，绘制统计图

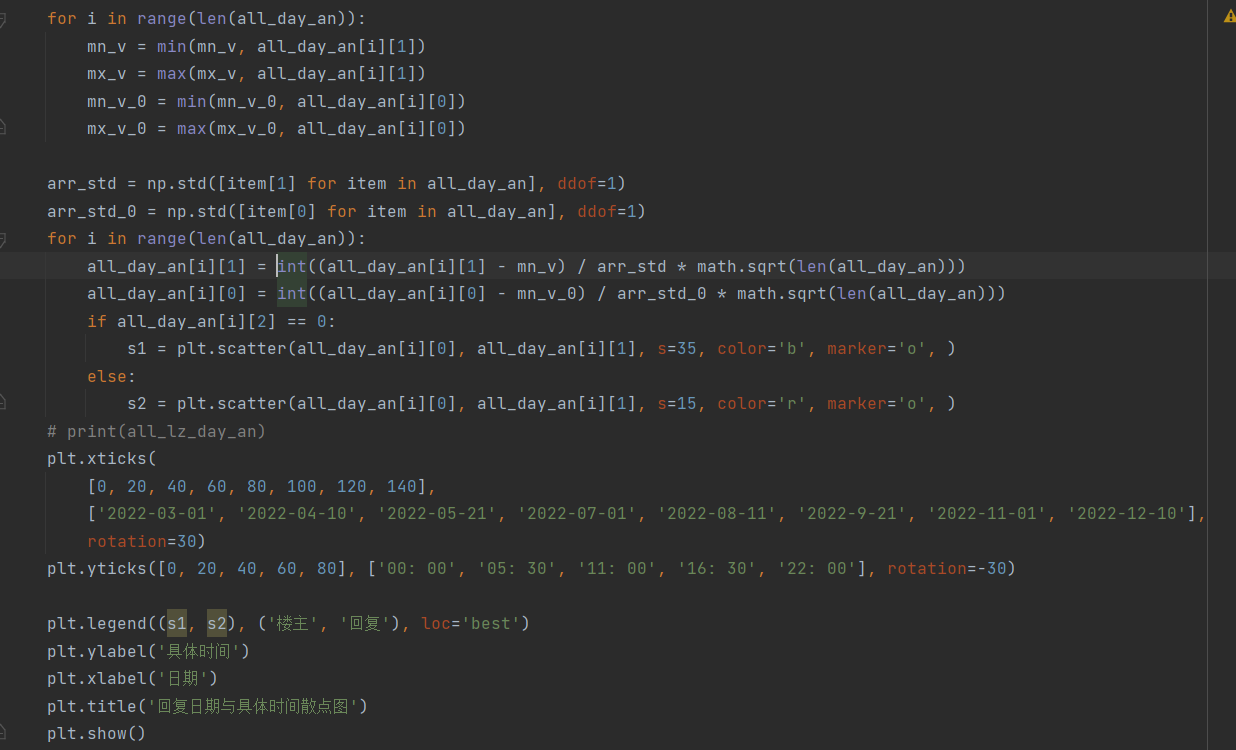


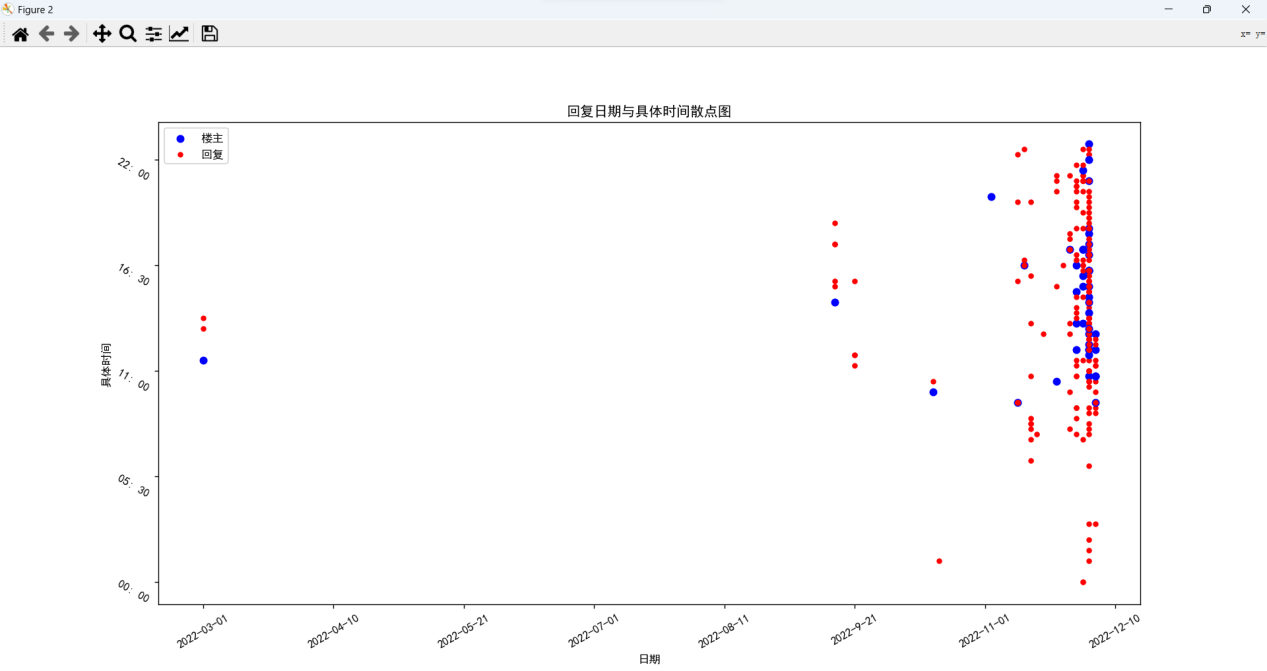


### 3.6.2 统计并做出发帖和回复时间的散点图

将时间分为”年-月-日“和“当日时间”两部分，并转为时间戳。将两个维度的整体移动，使得他们的最小值都是0，再除以他们的标准差缩小点与点之间的距离，最后画出散点图。







## 3.7 图形界面设计

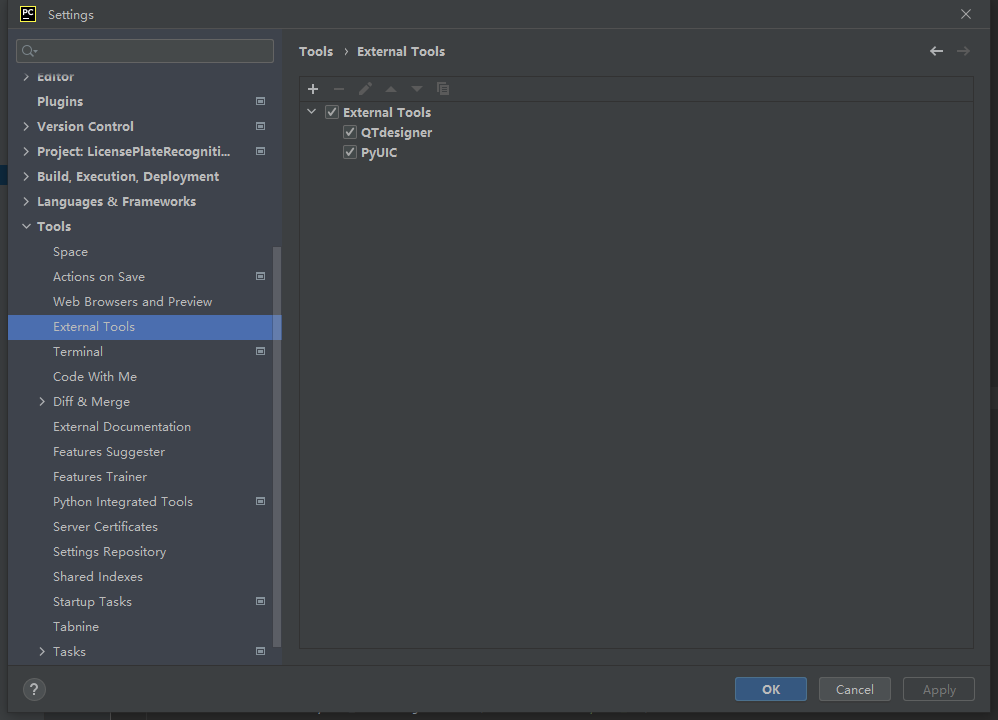
我们的程序主要分为数据爬取和数据分析两大部分，在数据分析部分我们会通过对数据的计算等处理方式最终得到贴子和评论，评论和评论间相似度和词云图，情感占比分布图，情感波动图，回复数量统计图和回复日期与具体回复时间图，通过图形的方式将我们的数据分析结果展示给大家。但是通过plt.show()的方法只能一张张显示图片，在结果的整体性上够，因此我们小组成员使用用pyqt5一起为我们的项目设计了图形界面，可以更加直观地展示我们的分析结果。

### 3.7.1 pyqt5简介

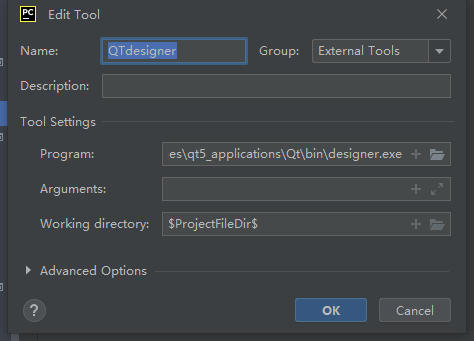
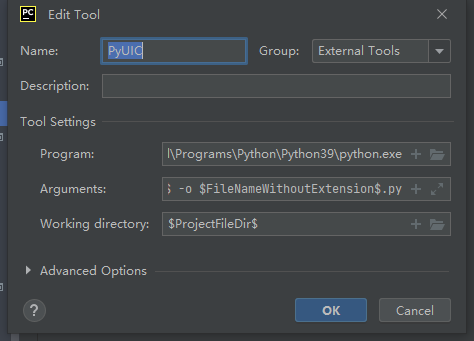
PyQt是Qt框架的Python语言实现，由Riverbank Computing开发，是最强大的GUI库之一。PyQt提供了一个设计良好的窗口控件集合，每一个PyQt控件都对应一个Qt控件，因此PyQt的API接口与Qt的API接口很接近，但PyQt不再使用QMake系统和Q\_OBJECT宏。

### 3.7.2 pyqt5的环境配置和使用流程

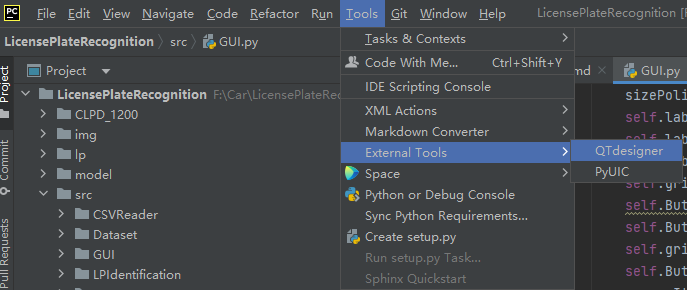
首先需要在pycharm中安装pyqt5的库，包括pyqt5和pyqt5-tools这 两大部分。然后在settings的tools中添加QT designer和PyUIC这两个工具，并且选择项目相应的designer.exe和python.exe路径



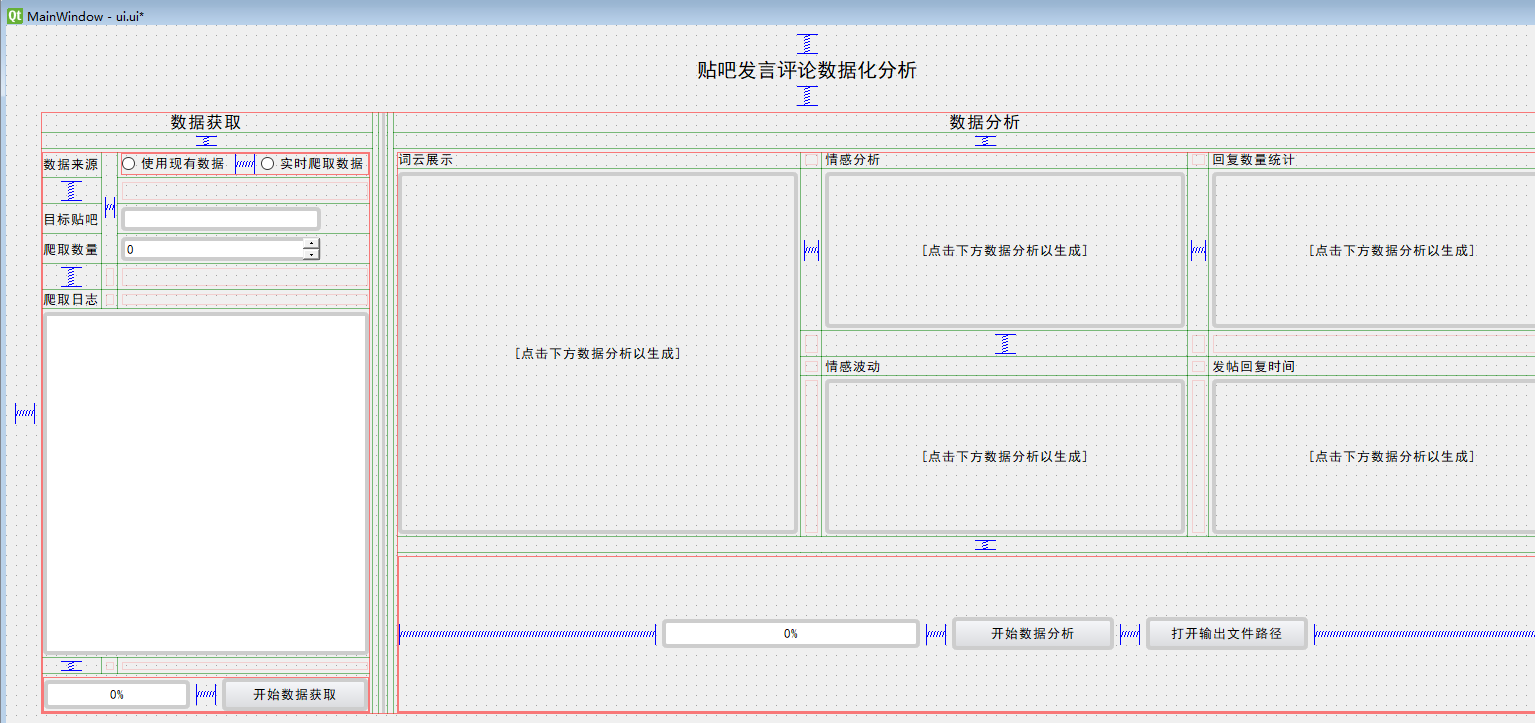
QT designer Tools路径配置以及PyUIC Tools 路径配置

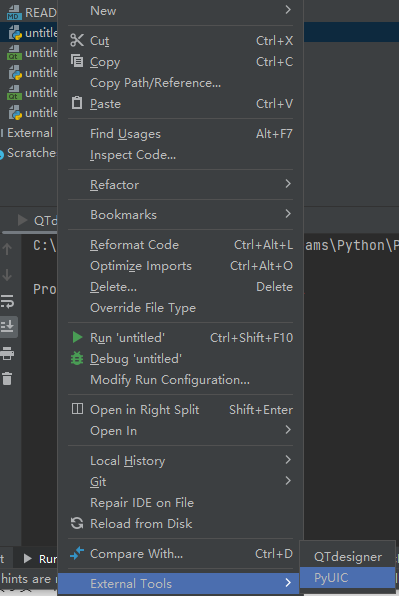
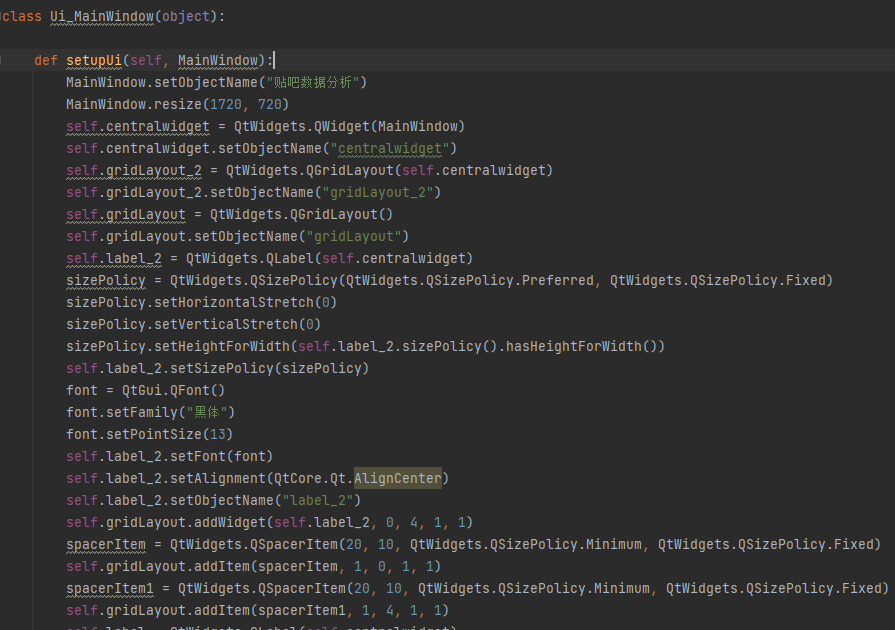
在上方导航栏中选择tools，选择External Tools进入QT designer中进行界面的UI设计



### 3.7.3 Ui界面交互设计



右键设计好的.ui文件选择External Tools，PyUIC将.ui文件转为.py文件

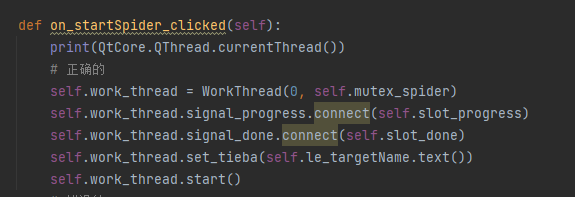
### 3.7.4 交互多线程调度

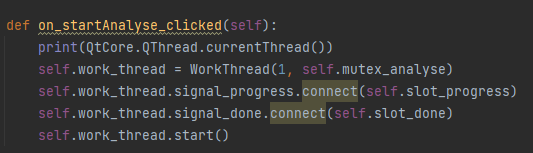
我们为图形界面中的“开始获取数据”按钮添加点击事件，若点击开始数据获取按钮，则调用on\_startSpider\_clicked函数，启动工作线程execute\_spider函数，通过预先设定的爬虫目标贴吧和爬取数量开始对贴吧数据进行爬取，并在爬取日志中实时显示爬取数据的情况。

工作线程如下：

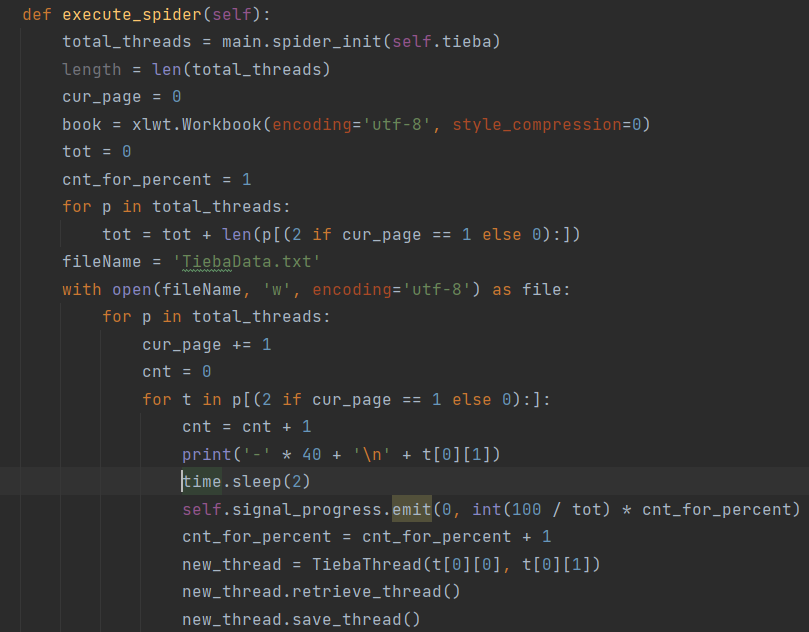


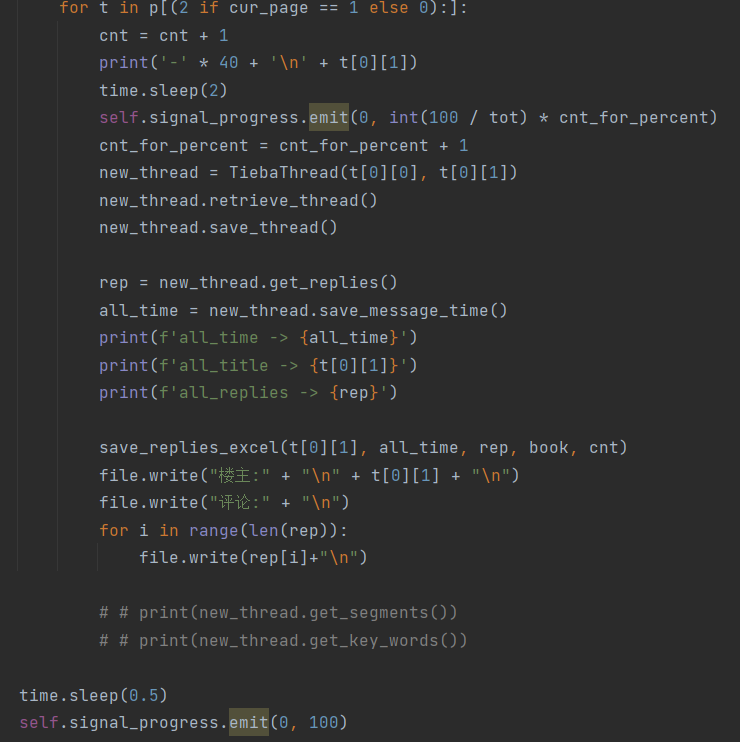
我们为图形界面中的“开始数据获取”“开始数据分析”按钮添加点击事件，若点击开始数据分析按钮，则调用on\_startAnalyse\_clicked函数，先启动工作线程execute\_analyse（）函数完成数据分析工作，数据分析完成并调用其中的slot\_done，load\_images函数，将程序运行后数据分析结束后的图片调整完大小后加载到图形界面当中

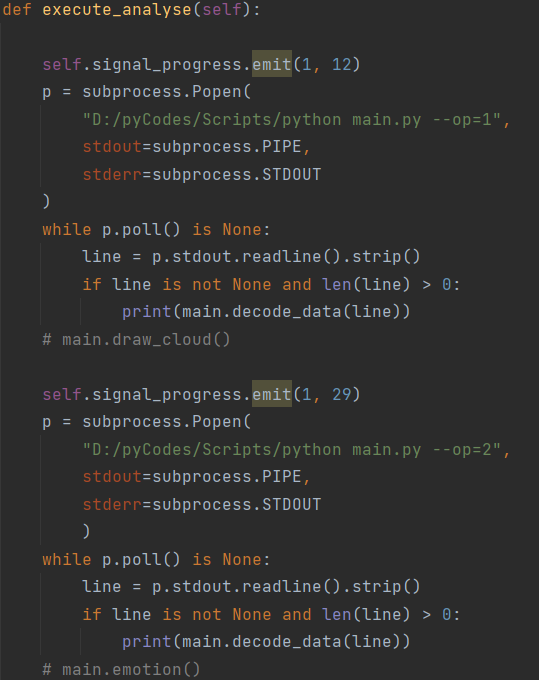


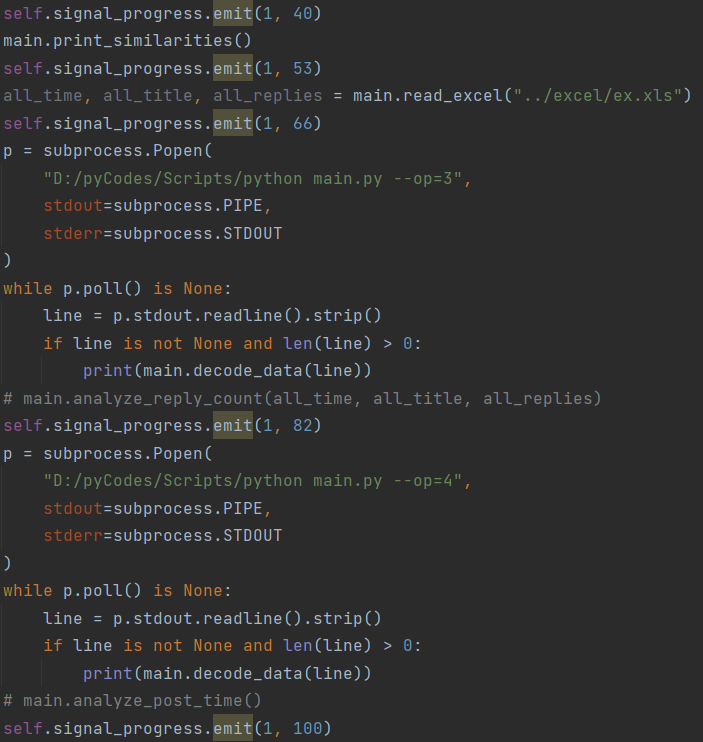


其中最重要的核心处理方法为execute\_analyse和execute\_spider，具体运行逻辑如下，其中通过发送signal\_progress与页面主线程同步运行进度

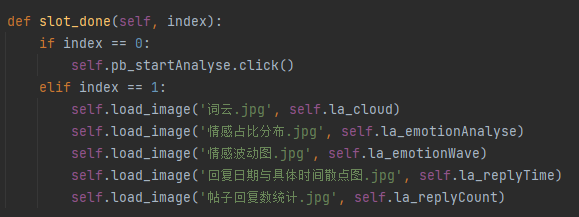




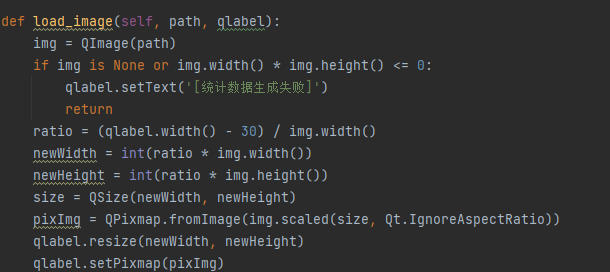




slot\_done()函数如下，负责在处理进程完成后加载生成的图片等行为：

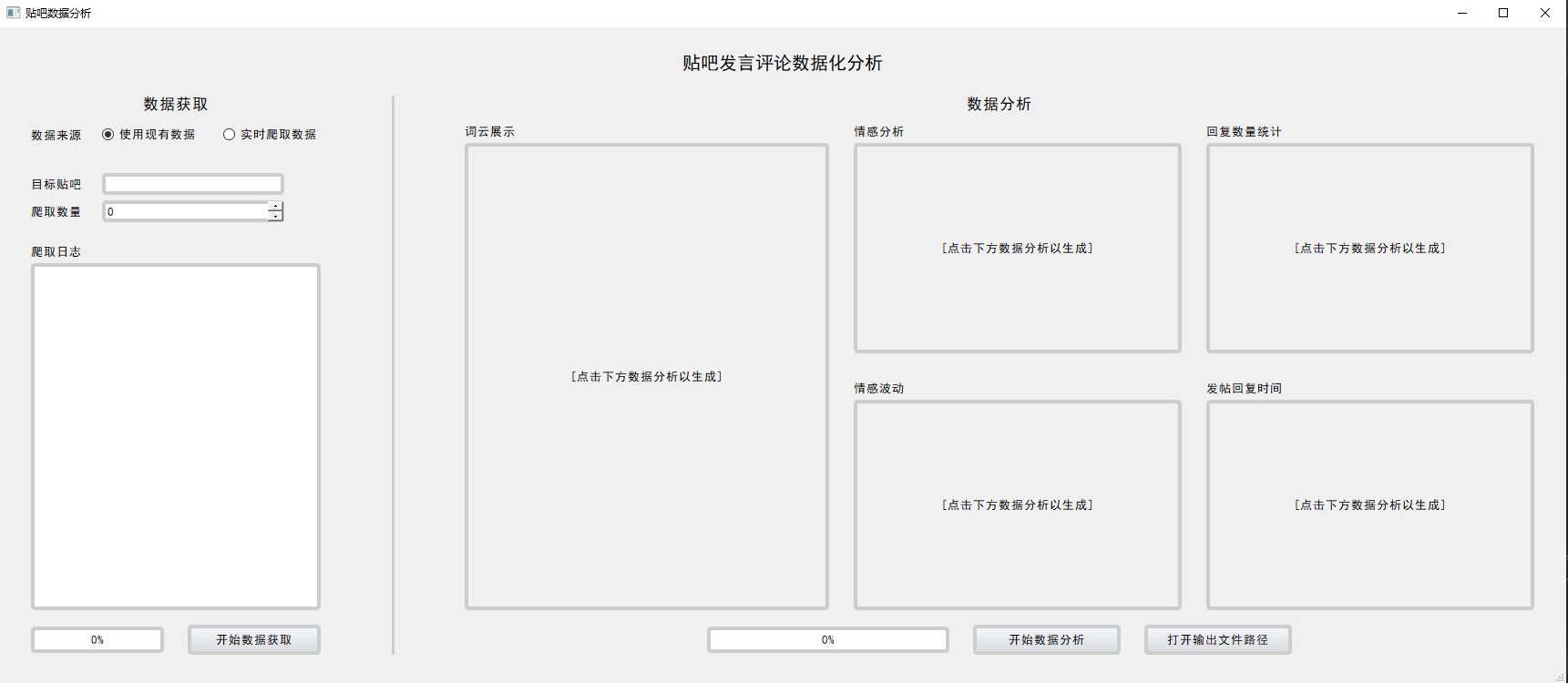


Load\_image()函数如下:



### 3.7.5 程序运行效果图如下

未选择操作时，进入应用界面效果如下：



使用现有数据进行数据分析完成后效果如下：

左侧显示部分数据分析过程中，如去停用词，分词后的结果；右侧显示数据分析完成后的词云图，情感占比图，情感波动图，回复数据统计图和发帖回复时间图。



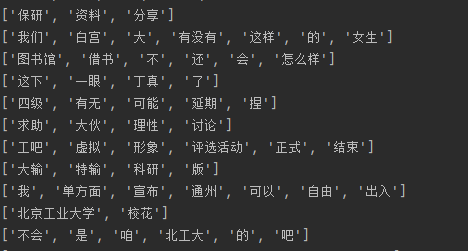
# 四、评估分析Evaluation

## 4.1 数据分析精确度评估

### 4.1.1 分词精确度评估

通过使用辅助文档停用词.txt和jieba库对北京工业大学贴吧的评论进行分词后，效果如下图所示；经过我们的观察可以将分词结果连成原文且动词和名词区分明显，所以认为我们的分词效果不错

例：[['鼠鼠', '我', '期末', '之前', '接', '了', '本书', '结果', '带回家', '了', '借书', '时间', '马上', '就要', '到期', '了', '也', '延长', '不了', '请问', '借书', '超期', '会', '让', '我', '赔', '吗']]



在分词精确度较高的前提下，我们才可以对关键词（分词结果）的进行词频统计，绘制出词云图，部分关键词词频统计如下

'考研': 0.6092572166811335

'老师': 0.5836153131787264

'知道': 0.5809243021719435

'可能': 0.502465460140125

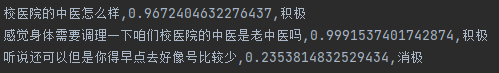
'考试': 0.477329848642965



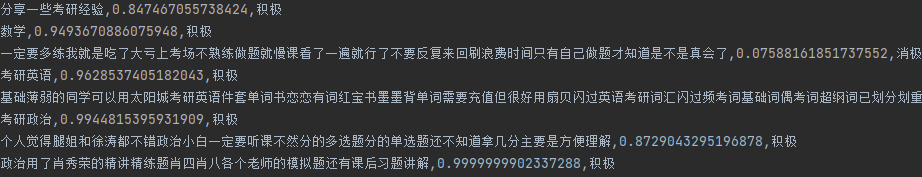
### 4.1.2 情感分析精确度评估

我们采用SnowNLP库对评论语句进行情感分析，将评论的情感分为积极，消极和中性，并将结果保存与.csv文件当中。

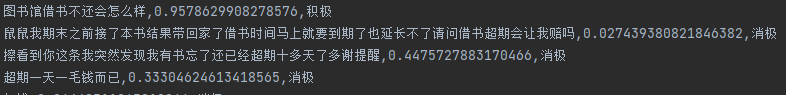
情感分析合理部分：在下图中我们可以看到对校医院讨论的评论中，询问校医院情况的语句被评判为积极情感，符合我们询问问题想得到他人帮助的积极情绪；第三句评论说到号比较少，说明就医有困难，被评判为消极情感也符合我们消极情绪。



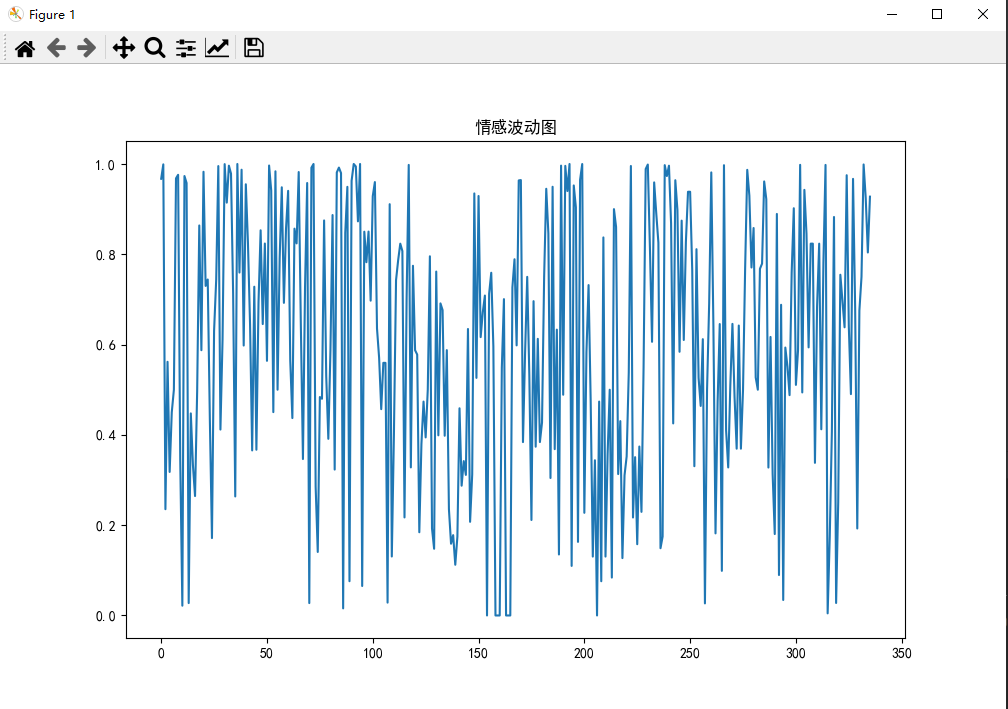
下面关于考研问题的评论，情感分析也很合理



情感分析不合理部分：在这则关于图书馆借书问题的评论中，我们小组认为倒数第二句，经过提醒意识到自己借的书还没有换，应该是积极情感而不是消极情感，对于这句话分情感分析我们认为是误判。

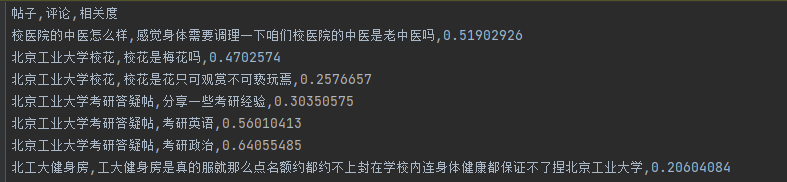


经过对于每条评论的情感值分析，我们得出情感分析波动图如下，可以看出北京工业大学贴吧评论的情感值波动较大，说明大家在发帖评论时都较为明确的表达了自己的情感。

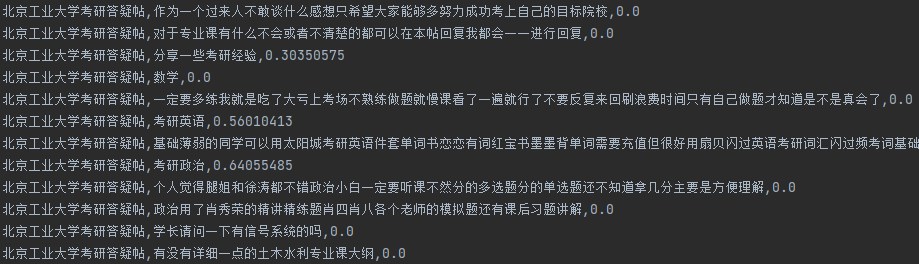


### 4.1.3 语句相似度分析精确度评估

我们使用gensim算法TF-IDF模型对语句的相似度进行了分析，选取了一些比较有代表性的结果如下，可以看出相似度比较符合我们的认知



但是由于语句相似度分析还是比较偏向于文本表面，分词后关键词是否一致的分析与对比，与语义相似度分析还差的很远，所以也出现了许多语句相似度为0的情况



## 4.2 项目优缺点评估总结

### 4.2.1 项目优点

1.选题现实意义强

我们小组此次大作业的主要研究方向为对北京工业大学吧数据的爬取和分析，这样的选题不仅贴近我们的生活，而且对于数据的爬取和分析也十分契合我们课上所学的内容，可以将上课所学应用于项目的功能实现当中。

2.数据分析全面

我们小组由浅至深的，通过去除停用词和分词，统计关键词的词频绘制了北京工业大学吧的词云；通过正则查找爬取下来网站的回复时间与每个贴子下的回复数，绘制了贴子回复数统计图，统计并做出发帖和回复时间的散点图；分析了贴吧语句的情感值，绘制了情感占比分布图和情感波动图；通过前期的分词和稀疏向量算法和TF-IDF模型的训练，分析计算了贴子楼主和评论，评论和评论间的语句相似度值。由此可见，我们很好地对于北京工业大学吧内的数据做了全面的分析。

3.编程技术点突出

我们小组在项目的编程实现过程中,不光使用了beautisoup4、jieba, WorldCloud、SnowNLP等经典的爬虫和数据分析库,我们还使用了较新的gensim算法和TF-IDF模型通过将关键词转化为稀疏向量集,计算向量间余弦值的方法进行语句相似的分析。除了调库，我们还通过正则表达式过滤,防止触发反爬机制等技术提高了爬取数据的全面性和精确性，使用自己设置的停用词文档辅助，更好的去除了停用词，得到了更为准确的分词结果。使得我们后续的数据分析展开的更加顺利。

### 4.2.2 项目不足

1. 数据分析的精确度有待提高

由于网站反爬和自身技术限制等原因，我们此次从北京工业大学贴吧上爬取下来的数据中没有图片和表情这两个元素。这导致了有许多的评论脱离了图片和表情，看起来会感觉到比较突兀，出现词不达意的情况。由此以来，我们对于数据的分析的准确度也会因此受到影响，出现准确度较低的情况，这在语句相似度分析中体现的尤为明显。这也是我们课下要研究和探讨的话题。

2. 时效性较差

我们此次的数据爬取并不是对于北京工业大学吧的实时爬取，所以我们的分析是基于已有数据的分析，不能涵盖到爬取数据后，北京工业大学吧更新的内容。

## 4.3 组员自我评价

### 4.3.1 吴祥曦

在这次我们的小组项目中，我主要负责对从贴吧上爬取下来数据进行处理和分析。在数据处理部分，我采用.txt文档的形式将实时爬取下来的数据保存下来，并通过去除特殊符号和留存中文的形式对数据进行了预处理，通过这样的处理后我们便得到了贴吧上的文本数据，方便用于后续分析。在数据分析过程中，通过去除预处理好的文本中的停用词，再经过jieba库的分词这两个步骤，我们会得到每条评论语句中的关键词和关键词词频。以此数据为支撑，绘制了词云图、情感占比分布饼状图和情感波动图。并且通过学习gensim算法中的稀疏向量集和TF-IDF模型，对贴吧贴子和评论间语句间相似度进行了分析。

最后和小组成员们一起对程序进行了调试，用多进程替代多线程的方式解决了plt函数必须在主进程调用和前端界面显示等问题，依靠我们小组互帮互助的精神和极强的个人能力，最终出色地完成了我们的项目。在此过程中我向小组成员们探讨和学了爬虫和界面设计等硬核知识，使得我的python编程水平和数据分析能力得到了很大的提高！感谢同组成员在我遇到困难时及时伸出援助之手！也感谢老师在课上为我们讲解许多有关大数据如NoSql等前沿技术，让我收获颇丰，老师您辛苦了！

### 4.3.2 李天依

我主要负责了本次大数据技术导论课程大作业的爬虫部分与总体功能设计，这一部分作为获取整个大作业所需的数据，可以说是至关重要的。我采用了比较传统的爬取流程，即获取网页转化为文本再用正则提取的模式。使用正则爬取数据的过程，也就是观察未经处理数据的过程。我通过浏览大量贴吧的网页源码，得到了所需的正则表达式模式，并解决了许多干扰html元素的干扰，完成了一个简单的百度贴吧爬虫，感受到了“爬虫的设计是至关重要而复杂的的”这一说法。在这学期的学习中，我也不仅仅学到了爬虫相关的知识。由于小组选题的约束，一些学到的但想要实践的内容并没有被整合进来，例如分类器、NoSQL技术等，但我有幸通过本课程了解这些前沿技术。希望在日后的编程与工作中进一步学习它们。

### 4.3.3 罗星悦

我在原有爬虫的基础上爬取了帖子回复时间，在正则匹配html页面时发现爬取下来的页面和在浏览器开发者模式中看到的界面内容不一样，经过查询有关资料可以使用selenium库进行处理，不过由于安装步骤较为繁琐、需要对代码进行重构以及关键信息依然可以被正则匹配得到等原因，最终没有选择使用selenium库。获取了这些信息后，我针对每个帖子进行了回复数量的统计以及发帖时间、回复时间的统计，分别绘制了条形统计图和散点图。针对散点图，我的作图依据是以日期为横轴，时间为纵轴作点。由于时间是采用时间戳的形式进行使用，不对其进行处理将会使得图像效果不好。因此我对数据的数值进行了平移，使得最小值为0。

最后，我们的前端使用了多线程的方式来进行爬取和绘图，这导致了plt报错，因为其只能运行在主线程中，最终我们采用了多进程的方法，使用了popen函数执行指定的cmd命令，从而避免plt类函数在子线程中执行。这个项目让我对一些分析方法的理解更加深刻，但是由于自身能力的问题没有将它们一一实现，还需要在日后的学习生活中继续努力

### 4.3.4 常兴阳

在本项目中，我负责了程序后端数据执行流程优化测试捉虫、前端界面的装配设计以及前端交互过程中涉及到的多线程相关设计，同时优化了程序和文档的整体视觉效果。在此过程中，我需要仔细阅读其他组员贡献的几乎所有代码，理解整体项目结构和模块运行逻辑，爬虫是以什么规则如何获取页面信息的、如何分析提取各个页面下有意义的文字信息、数据以什么格式存储并于什么时候读取转换为需要的格式、五张图是通过如何的统计手段生成等等诸类；当对代码内容有质疑或询问时需要频繁与其他小组成员交流、确认、排查修改，是一项比较费精力的工作。不过当第一遍捋清楚整个程序走向后，其中发生的小修小改便容易理解的多。

数据爬取、数据分析部分运行要花费相当长且不确定的时间，使用多线程将交互进程与处理进程分离可以保证交互页面处于可交互状态(非假死)情况下展示运行进度以及重定向命令台中内容至交互前端中，用户可实时以容易阅读的方式查看运行处理的内容，为了避免用户误操作导致线程竞争，我为子处理线程分别设置了进程同步锁。借此次机会，我得以在实践中深入了解数据的获取和处理应用，熟悉pyqt多线程中线程之间的信息传输控制。

此外，感谢您本学期的授课，为我们介绍构建了大数据技术的认知框架：从绪论中提到大数据用途于websearch、datamining、推荐算法、Ranking Approach、数据挖掘中的多种分类器的算法、机器学习中重要技术的概况。大数据时代数据的非结构化储存形式等等。课上受益匪浅，希望以后还能从您这里学习！

### 4.3.5 王可宣

经过本次的项目经历，我学会了熟练运用python编程，并结合人工智能相关的库进行数据集的创建、利用训练集进行训练，对数据分析常用的方法和工具有了进一步了解。我主要负责文献的收集、数据下载及处理，与吴祥曦同学一起对贴吧数据进行整理，包括表情包及部分特殊符号的截取，使得数据更加接近数据原型的基本格式，为之后的数据分析奠定了基础。此外由于项目投入时间原因且我本身有较好的英语基础，我还帮助各位组员进行论文翻译的校对与整理工作，帮助组员更好理解论文的详细内容，以便学习到我们可以运用与项目中的知识与方法。学会用图表的方式展示数据，学会在大数据的基础之上进行进一步的算法优化，将结果的准确率大大提升。

同时也看到了和同学之间的差距，在python编写方面也向他们学到了许多，优化自己的思维方式，优化代码提升效率和代码可读性，再次感谢我的组员们。也感谢老师一个学期的教导，经过您的讲解我对大数据有了更多的了解，也因此亲身体会了大数据的妙用，再次感谢您！