上海大学

SHANGHAIUNIVERSITY

毕业设计（论文）

UNDERGRADUATEPROJECT(THESIS)

题目:微博热搜及其评论分析

学院	计算机工程与科学学院
专业	软件工程
学号	18721802，18721803
学生姓名	李琦，李孝军
指导教师	刘炜
起讫日期	2018.09.3 – 2018.11.7

摘要

本项目主要着眼于人们当前社会上热点事件的看法以及其中的网民数据分析。使用 Python 语言通过对微博热门事件评论的数据获取，并针对得到的数据进行分析与可视化。

主要工作分为以下几步：

1）通过 python 爬虫爬取新浪微博焦点事件评论信息。

2）根据得到的评论内容进行分析与数据可视化得到：

① 评论关键字词云

② 评论内容情感分析折线图与柱状图

③ 评论用户所属地区热力图

关键词：微博 词云 情感分析 热力图

Abstract

This project focuses on the views of people's current hot events and the analysis of Internet users. Use the Python language to obtain data from Weibo's popular event comments and analyze and visualize the resulting data.

The main work is divided into the following steps:

1) Crawl the Sina Weibo focus event comment information through python crawler.

2) Analysis and data visualization based on the comments received:

①Comment keyword word cloud

②Comments content sentiment analysis line chart and histogram

③Comment on the heat map of the user's region

Keywords: Weibo, Word cloud, Sentiment analysis, Heat map

目录

第一章 绪论

随着社交网站的日益发展，网民对热点事件的关注度也随之上升。网民的发声散布于众多网络媒体社交网站中，而我们难以从其中某一条提取出重大舆情风向和网民对此的整体看法。基于此问题，本项目主要着眼于人们当前社会上热点事件的看法以及其中的网民数据分析。而众多社交网站中，微博的用户数量稳居首位，据统计截止 2018 年 6 月，中国网民规模达到 8.02 亿人，而其中微博用户数量达到 3.37 亿人，所以我们决定对微博数据进行爬取分析。

第二章 微博爬取

2.1 前期准备

2.1.1 站点分析

微博目前一共有三种类型的网站分别是：https://weibo.cn，https://m.weibo.com，https://weibo.com。三个网站的复杂程度依次提高，而其中第二个网站并未提供微博高级搜索功能，并且由于微博的反扒机制，我们选择了第一个网站。

2.1.2 微博登陆

要抓取到微博的数据，首先就是要登陆微博，否则就会重定向到登陆界面。而微博检测用户是不是登陆了微博，就是检查这次 Request 请求携带的 cookie。

而针对大量用户我们不能逐个使用复制 cookie 的方法,必须通过自动化登陆来实现。自动化登陆，就是通过代码来驱动浏览器，进行微博的登陆操作，具体通过自动化工具 selenium 来实现，它支持 Chrome，Firefox 和 PhantomJS 等多种浏览器。好处就是，不用再去分析登陆时候 js 加密，解密的过程，直接而且简单，坏处就是效率比较慢，但是我们只是用它来完成登陆并获取 cookie 的操作。

获取 cookie 以后，可以保存到数据库中。以后每次 request 请求，随机从数据库中选一个 cookie 加上，就免登录了。

2.1.3：环境搭建

1）数据库： mongodb

2）开发环境： Python2.7

3）一个新浪开发者账号：这里我们通过申请新浪合作 APP，来注册一个仅供学习的开发者账号。

4）需要的库：requests 和 csv(这些都可以在 Pycharm 中下载)

5）申请好开发者账号之后，我们打开自己的应用，获取应用的 APPkey 和 APPsecret。

图 1

6）设置回调页面

图 2

通过新浪微博的 API 接入网站，由于用户无需在网站上注册，就可以直接使用他/她在新浪微博的帐号和口令登录网站，这就需要确保网站在无需知道，也不能知道用户口令的情况下确认用户已经登录成功。由于用户的口令存储在新浪微博，因此，认证用户的过程只能由新浪微博完成，所以我们只能通过 OAuth(也就是我们自己的应用的回调网址)来确认用户登陆，OAuth 是一个标准的第三方登录协议，借助 OAuth，网站就可以安全地接入来自新浪微博登录成功的用户。

2.2 利用 APP 开发者授权进行数据获取

1）用户在网站上点击“使用新浪微博登录”，网站将用户重定向到新浪微博的 OAuth 认证页，重定向链接中包含 client_id 参数作为网站 ID，redirect_uri 参数告诉新浪微博当用户登录成功后，将浏览器重定向到网站。

2）用户在新浪微博的认证页输入帐号和口令。

3）新浪微博认证成功后，将浏览器重定向到您的网站，并附上 code 参数。

4）网站通过 code 参数向新浪微博请求用户的 access token。

5）网站拿到用户的 access token 后，用户登录完成。

OAuth 的 access token 是提供认证服务的网站（例如新浪微博）生成的令牌，代表一个用户认证信息。在随后的 API 调用中，传入该 access token 就代表这个登录用户，这样，通过 OAuth 协议，网站将验证用户的步骤交给新浪微博完成，并由新浪微博告知用户是否登录成功。

至此，我们的项目就成功完成了对于微博数据权限的获取。接下来就可以通过对微博 API 进行请求获取微博数据，以及进行数据提取。

本项目主要选取的微博内容是微博评论内容，具体有以下几个成员：

created_at（评论时间）

comment_id（评论者 ID）

text（评论内容）

followers（粉丝人数）

follow（关注人数）

province（所在省份代码）

为了与后面的数据分析同一接口，我们将获得的数据存储进 csv 文件。

第三章 数据分析

前面使用 Python 爬虫爬取了新浪微博的一些热门微博的评论与评论所属用户的个人信息，并生成 csv 文件。本章内容叙述对第二章所爬取的数据进行的处理与分析。主要包括三个部分的可视化显示：对评论进行分词与词频的统计，根据词频生成词云；对评论进行情感倾向判断与分析，生成情感人数分布图；统计各地区评论的人数，并生成地区评论热力图。使用到的 Python 模块：Wordcloud; Snownlp; Jieba; BaiduMapAPI。

3.1 词云生成

词云的生成主要为三个步骤：

1） 对爬取到的评论信息进行过滤

爬取到的评论原始格式中还包括一些 HTML 标签，表情，链接等，通过正则表达式匹配，过滤掉无用的信息，保留文字评论。

2） 分词与词频统计

对过滤后的评论字符串进行分词，并进行词频的统计。中文分词使用 Jieba 分词模块

3） 生成词云

通过分词得到的评论词库与其词频，使用 Wordcloud 模块生成词云。

效果图如下：

图 3

3.2 情感分析

通过前一步所过滤得到的纯净的微博评论字符串，使用 Snownlp 模块进行情感分析，获得每条评论的情感倾向值。并存储在自定义的用户数据结构中，用户构造函数如下：

```
def __init__(self, name, oricomment, comment, age='0', area='none', mood=-1, url='#'):
        self.name = name
        self.oricomment = oricomment
        self.comment = comment
        self.url = url
        self.area = area
        self.age = age
        self.mood = mood
```

根据得到的情感值数据进行频次的统计，生成情感倾向分布图表，如下（情感倾向从

0 到 10，积极程度逐渐增加）：

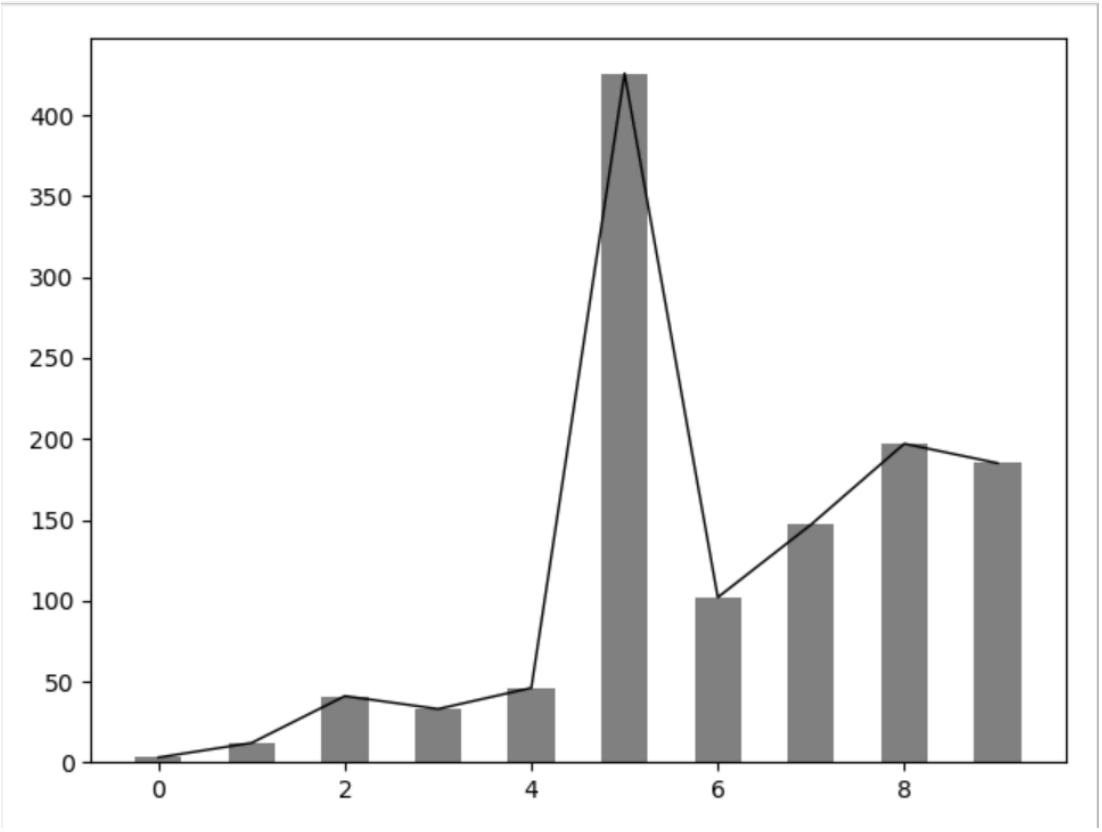

图 4

3.3 热力图

基于地图的数据可视化应用愈来愈广泛，目前，有很多方法来实现地图可视化，包括 excel 的 power map 包、各种数据分析软件的地图库以及在线交互地图可视化操作工具，如 Echarts、Tableau Public、polyMaps 等等。 还有一种手段就是通过软件调用百度、google 或者其他地图的 api，自己设计可视化地图。本文实现热力图的方式即为调用百度 api 生成基于省份的微博评论热度热力图。主要步骤如下：

1）获取经纬度

通过爬取得到的省份代码获取省份名称，调用经纬度获取 API 获取城市代码对应的经纬度信息。

2）统计人次

爬取的用户信息中包含用户的所属地。通过遍历自定义的用户数据结构，统计得到各地区评论的人次。

3）生成热力图

根据以上两步得到的数据，生成 json 文件，并使用 BaiduMapAPI，生成可视化的用户评论数量热力图。

图 5

总结

通过本次项目的实践，我们分析了几个当前的热门话题，如重庆公交车坠江事件以及金庸先生去世。从数据分析中我们不难看出，网民对于热点事件的关注程度非常之高，并且对于热点事件能够有足够的正面情感输出。这也间接反映了我国网民素质的普遍提高和网络环境的极大改善。

参考文献

[1] http://open.weibo.com/wiki/API

[2] https://www.cnblogs.com/FanLeiData/articles/7675890.html