# 热门电影影评数据爬取与分析项目报告

目录

[热门电影影评数据爬取与分析项目报告 1](#_Toc16301716)

[一、 项目背景 2](#_Toc16301717)

[二、 项目目标 2](#_Toc16301718)

[三、 项目内容 2](#_Toc16301719)

[3.1 爬取数据并保存 2](#_Toc16301720)

[3.2 短评词频分析 2](#_Toc16301721)

[3.3 时间分析 2](#_Toc16301722)

[3.4 城市分析 3](#_Toc16301723)

[3.5 用户会龄分析 3](#_Toc16301724)

[四、 总结 3](#_Toc16301725)

## 项目背景

在电影行业飞速发展的当今，电影已经成了非常普遍的娱乐选择，中国电影最近几年也是突飞猛进，越来越多的人走进电影院。而信息媒体的发展也使得所有人都拥有了随时随地与世界各地的人讨论电影的己会，电影无疑已经进入了全民关注，全民讨论时代。

但是随着越来越多的电影上映，越来越大的阅片量，大家对电影的质量和制作水平的要求也越来越高，电影出品放想要斩获高票房，就要把握好消费者的喜好，制作出符合市场期待的电影。

在国内，豆瓣、淘票票、猫眼三家网站都有自己所属平台的评分功能，俨然形成了国内的“电影评分三巨头”。电影评分网站上蕴含了大量的电影评论、排行、评分等信息，是消费者择片困难时的重要参考依据，也是电影出品方了解用户喜好的有效途径。

一份科学的电影消费者数据分析报告，是预测票房的重要依据之一，也能够帮助出品方规划未来电影拍摄计划，对出品方规划未来电影拍摄计划，对出品方投资有着重要意义。所以，如何获取、分析消费者喜好信息是电影出品方非常关注的事情。

## 项目目标

本项目通过python网络爬虫获取主流电影评分网站豆瓣网上关于《复仇者联盟4》评论用户的基本信息、影评内容等数据，并对获取到的影评数据进行简单的文本分析，帮助出品方了解用户偏好。

## 项目内容

### 爬取数据并保存

爬取数据主要分为两个部分，一是获取网页内容，而是对网页内容进行解析，之后存储数据即可。

而获取网页内容有两种，一种是使用requests库对豆瓣服务器发送请求获得，从网络获取原始的网页；一种是使用Selenium库模拟浏览器请求，带有动态js代码执行后的结果。开始使用requests库爬取，并使用cookie表单登陆，很容易被豆瓣封掉ip，后期对cookie值每次随机，并使用time.sleep()功能保证较慢的爬取速率，才勉强解决这个问题。这个时候爬取到578个数据，并且存入了本地数据库，具体代码见【TASK1\_1.py】（爬取评论页数据）、【TASK1\_2.py】（根据用户url爬取用户信息）。网页的解析使用BeautifulSoup。

后来在尝试完成TASK3的时候，发现需要评论的时间具体到时刻，而开始用requests库爬取的评论时间只爬到评论的日期。于是又进行了一次数据的爬取。

第二次爬取尝试了Selenium库，模拟真实用户操作浏览器。该过程中确实没有遇到封ip的状况，就是登陆次数多了再次登陆的时候会要求输入验证码。另外，Selenium库爬取确实要相较而言要慢很多。这次由于时间关系没有爬取最新数据，所以减去重复的评论只有488个数据，用此数据完成Task3、Task4、Task5。这次存储数据到excel文件中，具体代码见【TASK1\_3.py】。网页的解析用BeautifulSoup。

结果在【all\_comments.xlsx】中。

### 短评词频分析

主要任务——关键信息展示，包含以下子任务：

* 分词并去除停用词
* 词频统计
* 绘制词云
* 分别绘制好评 中评 差评的词云

主要用到jieba库对短评信息进行分词，然后对得到的分词结果通过wordcloud库制作成词云展示，频数越高的词汇，在词云中展示得就越大。

具体代码见【TASK2.py】

结果如下：

3.2.1 所有评论的词云



3.2.2 好评的词云



3.2.3 中评的词云



3.2.4 差评的词云



### 时间分析

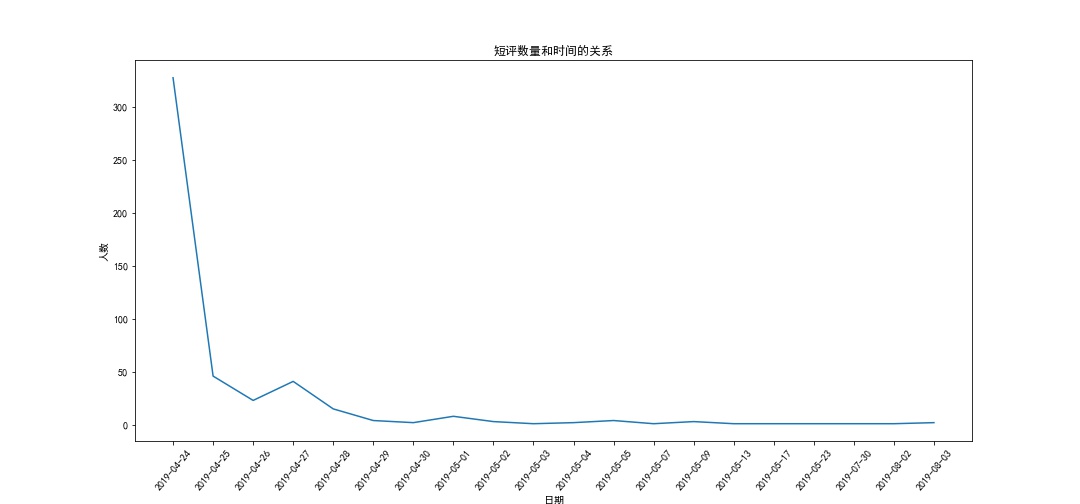
主要任务：

* 分析短评数量随日期的变化情况
* 分析短评数量随时刻的变化情况
* 分析随日期变化，评分变化情况

主要使用pandas来对数据进行操作，将数据转化为timestamp类型；对数据进行聚合操作。并使用matplotlib.pyplot对处理后的数据绘制达到可视化目的。

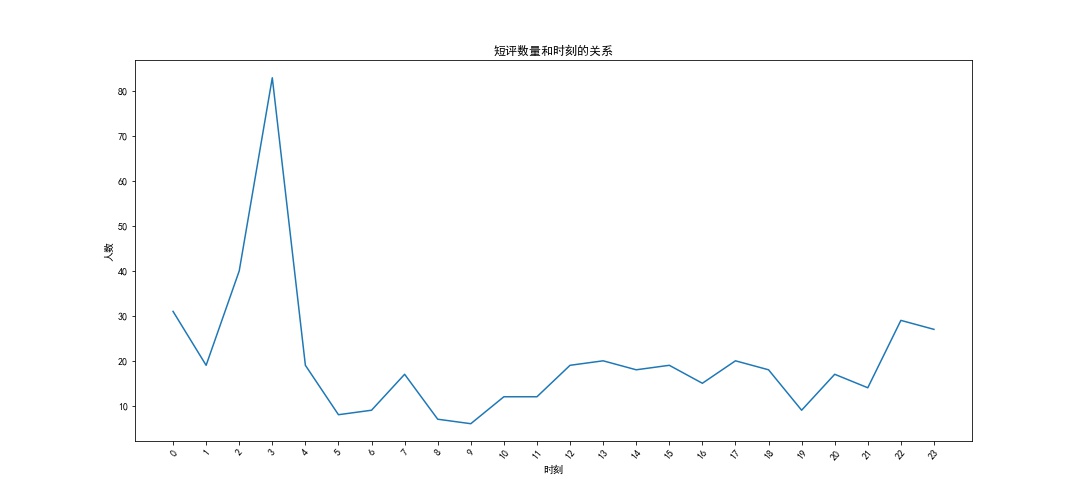
具体代码见【TASK3.py】

结果如下：  
3.3.1 短评数量随日期的变化



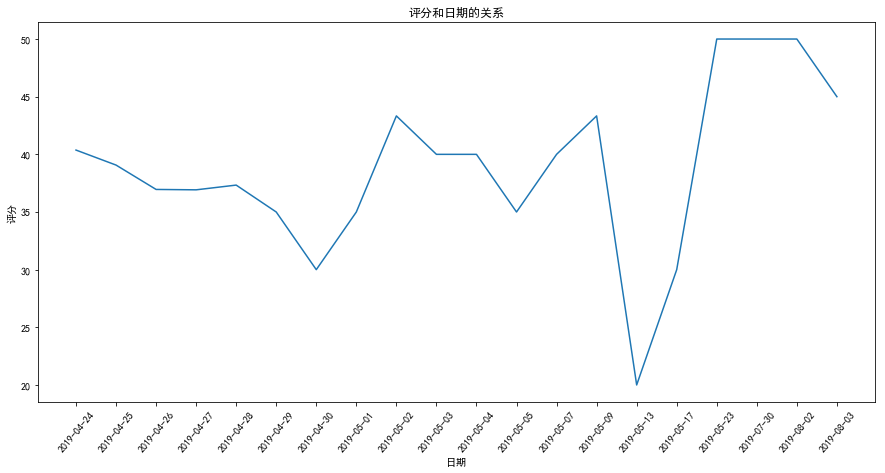
分析：可以看到首映当日的短评数量是最多的，其后4天也维持在几十人评论的水平，后期每天基本就没有多少人写评论了。

3.3.2 短评数量随时刻的变化



分析：可以看到，首映结束后的两点半是最多人发短评的时刻，八十几人。可见还是有相当多人看半夜场的首映的，还有一个高峰是早上7点，而12点到18点人数维持在一般水平。另外一个高峰出现在22点到24点。大多数人是看完电影之后评论，因此可以推测，较多人看八点到十点开始的场次。

3.3.3 随着日期变化，评分的变化情况



分析:首映当天评分还是相对较高，40左右，首映当天后6天评分一直呈现下降趋势，并在4.30到30分低谷，而后评分开始一路攀升，于5.2到达接近43分的高分。而后又是一波下降到5.5到达又一个低谷35分，又一波上升到5.9又是43分，到了5.13号居然是断崖式降分到20分，5.23又到达50的高分。可能评分降低之后就有粉丝将评论分刷高。

### 城市分析

主要任务：

* 分析用户常居城市分布情况
* 分析不同城市的评分情况 不同城市对应的评分情况用户会龄分析

主要使用pyecharts库进行地理图的绘制。通过pandas统计每个城市的用户数量，以及每个城市的用户的评分的均值作为数据进行绘制。

具体代码见【TASK4.py】

结果如下：

3.4.1 具体结果见*《复联4》短评用户常居城市分布.html*

有小部分用户是在国外的，但是目前只绘制了位于国内的用户分布情况。可以看到北京评论用户是最多的。可能是因为豆瓣在北京最为流行，可能是复联在北京最受欢迎，可能是因为北京用户评论的质量比较高。其次是上海共52名。接着是广州、深圳、成都、武汉、南京、杭州。可以看到一线城市和准一线城市对复联的关注比较多。

3.4.2 具体结果见*《复联4》短评评分与城市分布.html*

北京评分均分41.2，上海36.15，广州43，深圳39，成都38，武汉40，南京36，杭州37。可能北京和广州的文化底蕴更加深厚，更加给情怀买账，而上海、深圳、南京、杭州等城市用户更加理性。

### 用户会龄分析

主要任务：

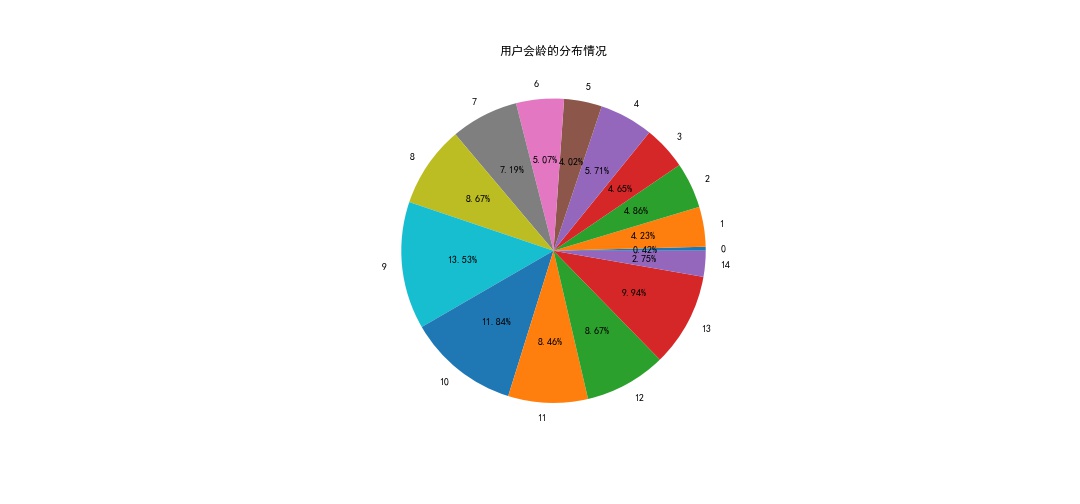
* 分析复联4发布短评用户的会龄分布
* 分析会龄对评分的影响 分析出受众情况怎么样

通过用户加入豆瓣的时间，采用正则表达式进行处理，并用时间戳类型的运算得到会龄，通过计算各个会龄的用户数量，以及会龄对应的用户的评论均值进行数据的可视化。

具体代码见【TASK5.py】

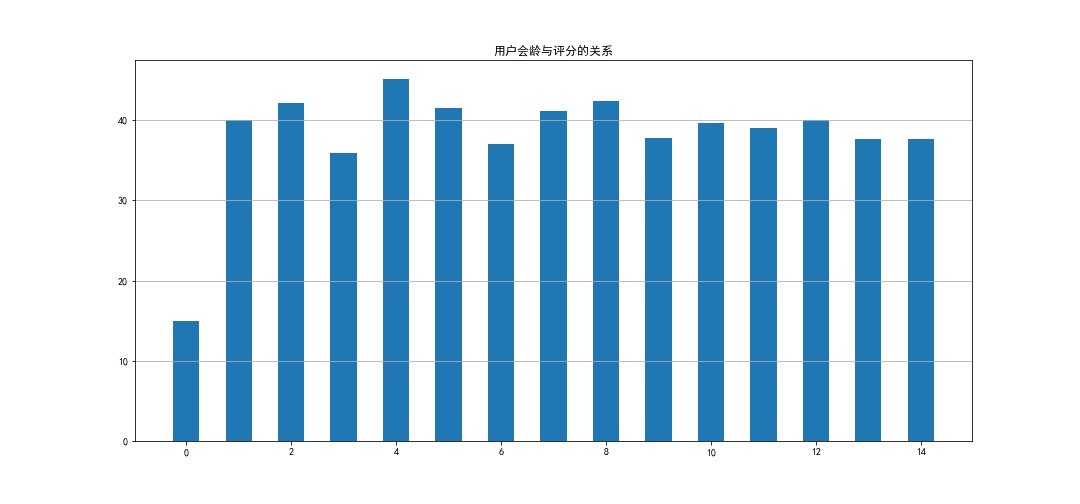
结果如下：

3.5.1 短评用户的会龄分布



分析：可见评价复联4这部电影的人中，会龄5年以内的只占百分之二十左右，而5年-14年会龄的占百分之八十，可能是因为漫威宇宙是从10年前开始的。会龄和年龄有相关性，就是说会龄可以粗略作为年龄的一种映射，假设一个人在青少年时期注册豆瓣，那么可以得知，观看复联4并评论的人中，他们并不是非常年轻的青少年，而很可能是距离青少年时期几年时间的人。其中9年前，也就是2010年注册豆瓣的用户数量最多，而那时就是《钢铁侠2》上映的时候，这么看，漫威宇宙的故事很可能覆盖了这些用户的青少年时期。

3.5.2 会龄对评分的影响



分析：可以看到，最低分属于今年刚刚申请豆瓣账号的用户，这些用户有两种可能，第一种是刷差评的水军，第二种是对漫威宇宙并没有深刻的感情而只是看到了这一部影片而已。8年以及比8年短的会龄的用户的评分普遍比8年及八年以上会龄的用户的评分高。可能是年龄偏大的用户对情怀不再那么买账而更加理性。

## 总结

爬取数据并不是一件容易的事，因为有很多网站设立了反爬虫机制导致爬虫并不顺利。处理数据时，对pandas的应用有了更多了解，但是距离灵活应用还有差距，比如对dataframe的聚合操作后如何生成字典等等很多细节上的问题。通过这次项目，对爬虫、NLP、python数据处理以及可视化都有了一定的了解。