《复仇者联盟4》短评爬取与分析报告

一、引言

在电影行业飞速发展的当今，电影成为了非常普遍的娱乐选择。但随着越来越多的电影上映，越来越大的阅片量，大家对电影质量和制作水平的要求也越来越高。电影出品方想要获得高票房，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，豆瓣作为被广泛认可的电影网站之一，包含了大量电影评论、排行、评分等信息，是消费者选择影片的参考，也是电影出品方了解用户喜好的有效途径。

《复仇者联盟4》是今年具有超高话题度的一部电影，本文运用爬虫工具从豆瓣网站上爬取500条对于《复仇者联盟4》的短评，以及短评的时间，对电影的评分和相关用户信息。并从文本、时间、地理位置和用户会龄四个方面进行分析。

二、数据获取

使用selenium库爬取豆瓣短评，并使用beautifulsoup获取短评、发表时间及用户信息，循环获取每一页的内容，最后把信息转化到数据框并保存到Excel中。

三、分析短评关键思想

将获取的500条短评读取到Python中，利用jieba库对所有短评进行分词，并去除无实际意义的停用词。然后进行词频统计，绘制出所有短评的词云，如下所示：



由上图可以知道，用户短评中提到最多的是“漫威”、“英雄”、“情怀”等词，这反映了用户对复联4这部电影的感受：这是漫威系列的一部英雄电影，作为一个系列的结局，它代表了多数人对漫威英雄的情怀。

在所有短评中筛选出评价为“力荐”和“推荐”的短评，绘制词云如下：



从上图可以看出，认为这部电影值得一看的用户在短评中提到最多的是“复联”、“结局”、“宇宙”、“角色”、“侠”等词，说明用户比较认可电影里面的角色，认为角色能体现他们心中的侠义。

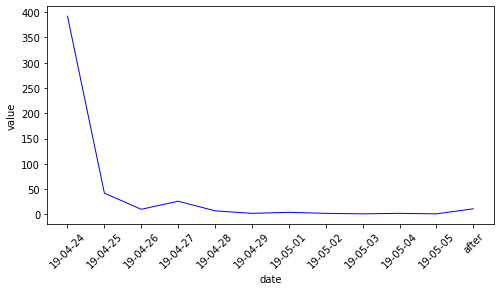
在所有短评中筛选出评价为“较差”和“很差”的短评，绘制词云如下：



由评价为差的用户短评绘制出的词语中，可以看出提到最多的是“时间”、“剧情”、“故事”、“情怀”等词，结合大概浏览这部分评论可以了解到用户的不满主要在于电影本身的情节，但仍会为了情怀而观看。

四、时间分析

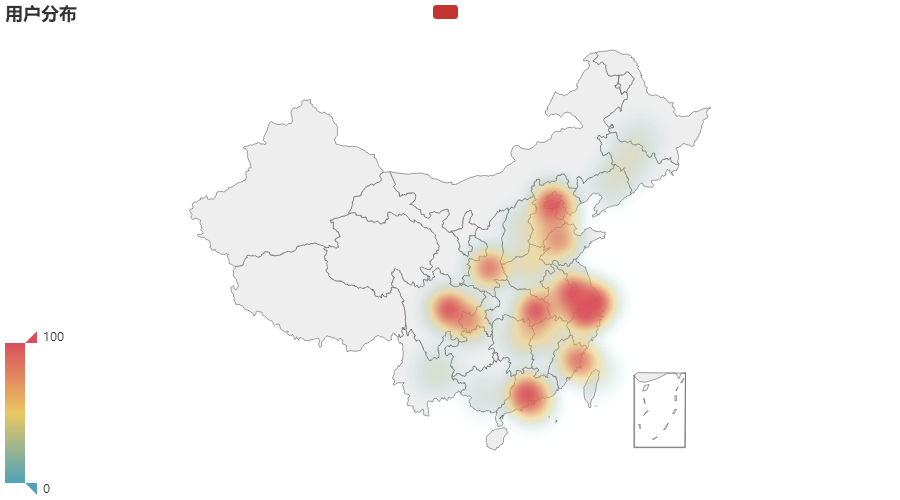
统计短评发表的日期，使用matplotlib绘制折线图，如下所示：



从上图可知，评论数量在首映当天爆发式出现，随后迅速下滑，在三天之后出现回潮的一个小高峰，随后再滑落，稳定在一两条的水平。说明电影在首映当天拥有最高话题度，在之后第三天出现小高峰，电影出品方应把握这两个点。

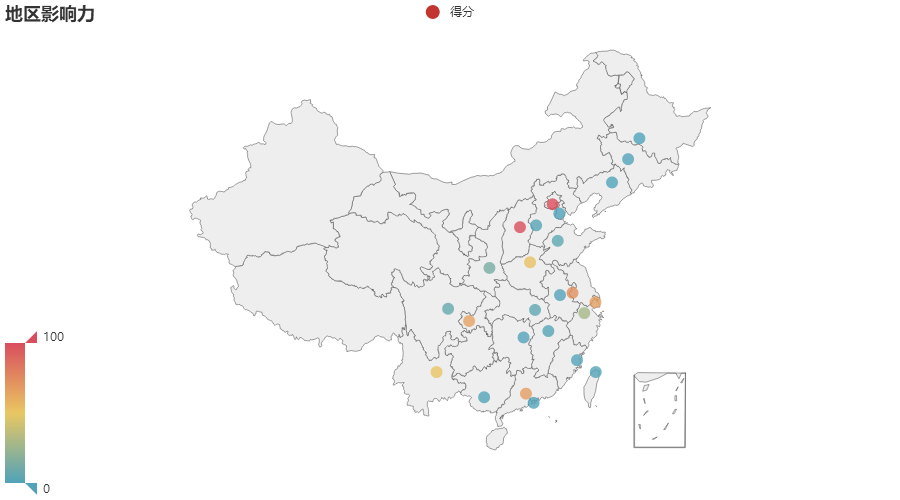
五、城市分析

统计用户的常居城市，利用pyecharts库绘制用户城市分布地图，如下所示：



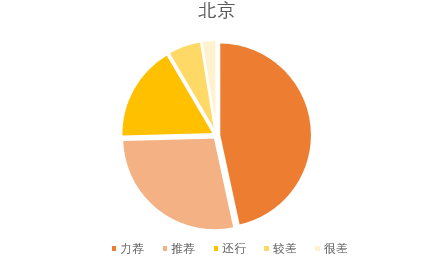
从图中可以看出，用户最多的地区为北京、上海和广东。且用户集中分布在中部和东南沿海，东北地区和南部较少，而西北地区几乎没有。这从侧面也可以反映出经济发达地区观影人数较多。

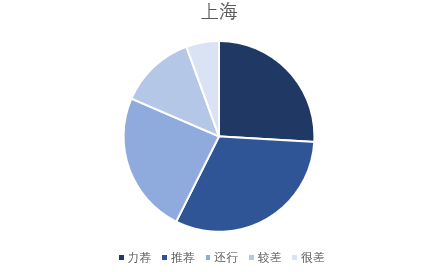
在豆瓣评论中，每一条评论都有认为该条评论有用的人数。认为有用的人越多，可以认为该条评论的影响力越大。爬取该人数，结合用户的地理位置，可以得出不同地区的评分影响力，和对这部电影的关注程度。地区的评论影响力除了评论的认可度外，还有评论的人数。为了能综合体现这两个方面，同时便于展示，对所有短评的认为有用人数进行标准化，并按省份求和，绘制地图如下：

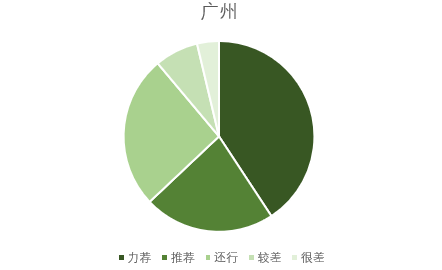


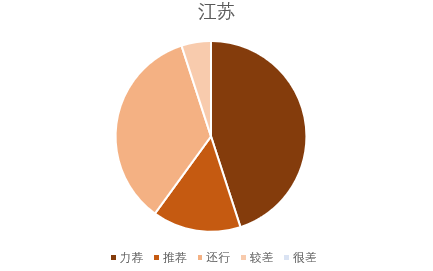
图中的点颜色越接近红，说明影响力越大。从图中可以看出，影响力最大的是北京和山西，其次是上海、江苏、广东、重庆。说明经济较好的大城市，对电影的影响越大。

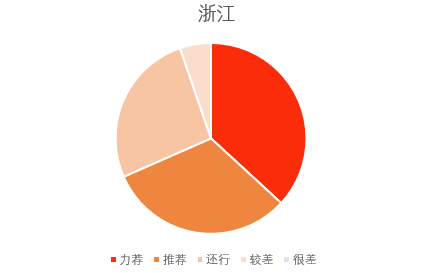
对发表短评人数前5的省份的评价进行统计分析，如下所示：











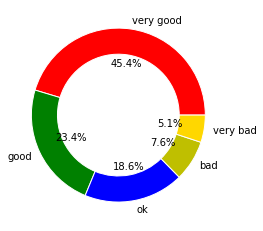
以上可以看出观影人数较多的5个城市中，超过半数的人认为值得观看，说明这是一部被大众认可的电影。

六、用户会龄分析

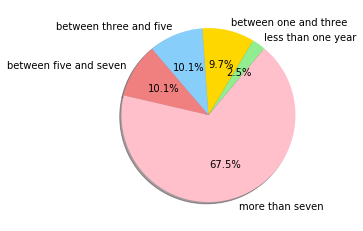
删除部分缺失数据后，创建不同会龄段给出不同评价的人数的透视表，统计数据如下：

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 会龄 | 力荐 | 推荐 | 还行 | 较差 | 很差 | 总和 |
| 小于一年 | 6 | 1 | 0 | 2 | 3 | 12 |
| 一至三年 | 28 | 5 | 6 | 5 | 2 | 46 |
| 三到五年 | 30 | 5 | 6 | 4 | 3 | 48 |
| 五到七年 | 19 | 10 | 12 | 6 | 1 | 48 |
| 大于七年 | 132 | 90 | 64 | 19 | 15 | 320 |
| 总和 | 215 | 111 | 88 | 36 | 24 | 474 |

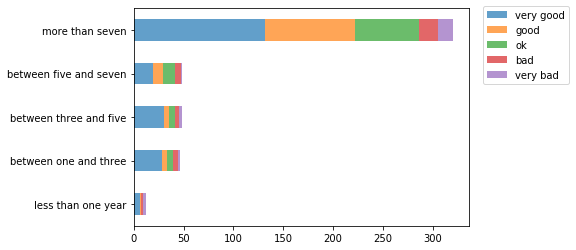
根据评分为人数绘制环形图如下，可以得出认为“力荐”的人数比例为45.4%，认为“推荐”的人数比例为23.4%，加起来超过半数。说明大部分人认为这部电影值得一看。



绘制饼图可以看出发表评论的用户中，会龄大于七年的用户高达67.5，说明这部电影吸引的观众是有一定观影经历的人群，多为成年用户。所以这些影评很大程度上能反映观众的感受，是比较客观的。



绘制条形图可以看出，在每个会龄段，给电影评分为“力荐”的人数占总体的三分之一到一半。说明无论是观影经历多还是少，人们都对这部电影较为认可。



七、总结

通过爬取豆瓣上的500条《复仇者联盟4》的短评进行分析，可以得到以下结论：

1、大部分人认为《复仇者联盟4》是一部好的电影，值得一看。观众认为这部电影体现了英雄侠义和情怀。认为这部电影不好看的观众主要对剧情不是很满意。

2、作为漫威系列的结局之作，拥有一定观影经历的漫威迷对这部电影有很大的热情，大部分的评论者会龄在七年以上。但无论观影经历多，还是所在城市不同，大部分观众能认可这部电影。

3、电影在东南地区反响比西北地区热烈，大城市对电影票房有大的影响。

4、电影话题的高峰在首映当日，之后热度大幅下降，在第三天出现小高峰的回暖，之后趋于平静。