热门电影影评数据爬取与分析项目分析报告

**1 问题提出**

本项目主要通过python获取《复仇者联盟4》评论用户的基本信息，影评内容，评分情况等内容，并对获取的信息进行分析处理，帮助出品方了解用户喜好。

电影出品方需要通过短评等信息，来了解观众的喜好，他们所期待的电影类型，并且与观众的相关地区联系，观众所在地区可能会对他们对于电影的评价有所影响，以及观众的年龄段都有所影响，电影出品方通过对不同方面的分析，来尽可能的达到观众的需求。

**2 短评分析**

**2.1 词云分析**

通过对比好评、中评、差评、短评（爬取的所有影评）的词云发现漫威一词在四个图中都有出现，而且大多高频词汇都相同，说明人们对这部电影的讨论不管是好是差，围绕的关键都是一样的。



图 1 全部短评词云

**2.2 用户发表短评数量随时间的变化**

通过短评数量随着日期的变化可以发现，电影上映当天评论数量最多，随后直线下降，一般人评论都会在看完电影后当即进行评论，那个时候也是观众印象最深刻的时候，以免时间过长，导致看电影时的心态变化。而在电影上映后还有一次评论数量较多的时候，通过查找发现2019-4-27这一天是周六，周六是休息日，看电影的人不在少数，因此评论数量出现了一个小高峰。

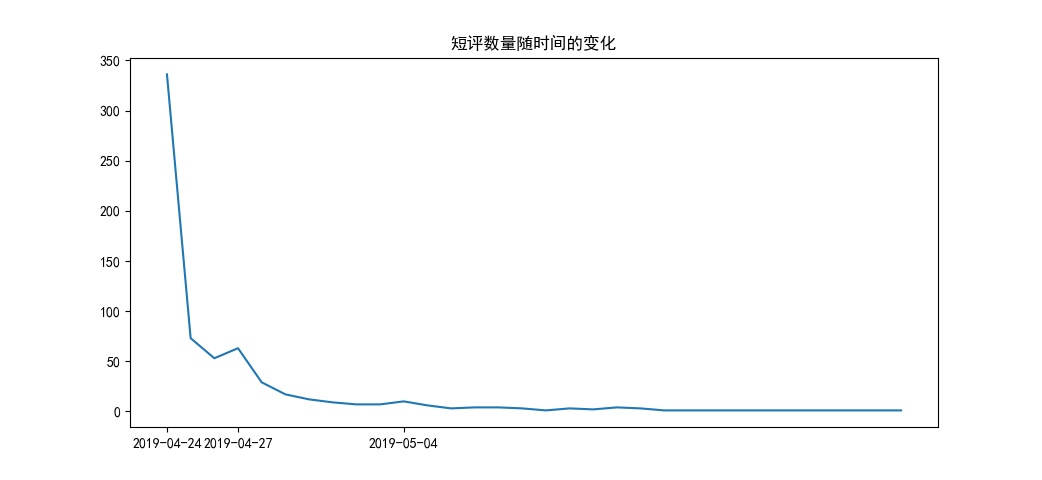


图 2 短评数量随日期变化

2.3 短评数量与评分的关系

由图可知，评分为3星的短评数量最多，其次是2星和5星，说明大多观众对复联4这部电影评价一般，这点需要电影出品商引起重视。

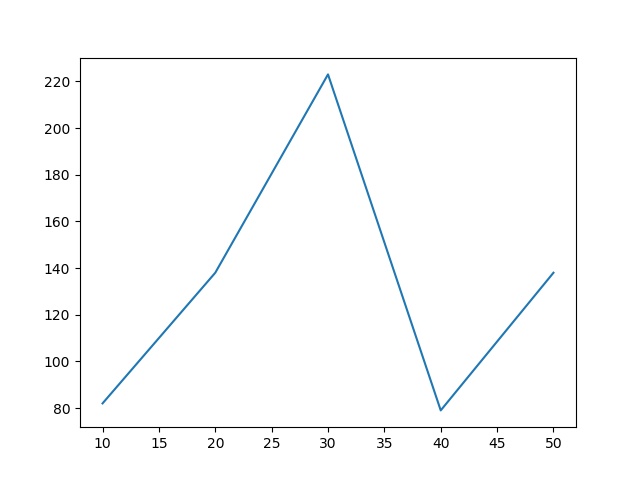


图 3 短评数量与评分的关系（10为1星，以此类推）

**3 城市分析**

**3.1 用户常居城市分布**

通过数据统计发现，短评用户居住城市最多的是北京，其次是上海。北京与上海，一个是首都，一个是经济及其发达的地方，居住人口都十分多。其次，经济繁荣的地方相对看电影的人较多。

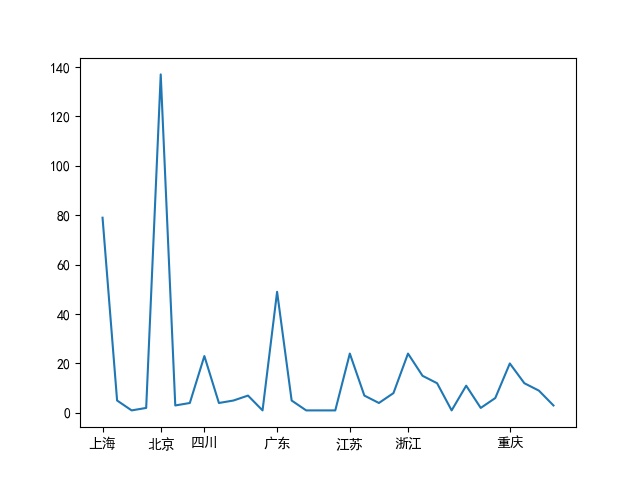


图 4 用户常居城市分布

**3.2 不同城市评分情况**

计算出每个城市的短评评分平均数。其中山西地区有两条短评，其中一条为介休，两条短评一条给了5星，一条给了1星。除此之外，其他地区的评分大多是在2星到3.5星左右波动，可以看得出各地短评都有好有差，整体还是偏好的居多。

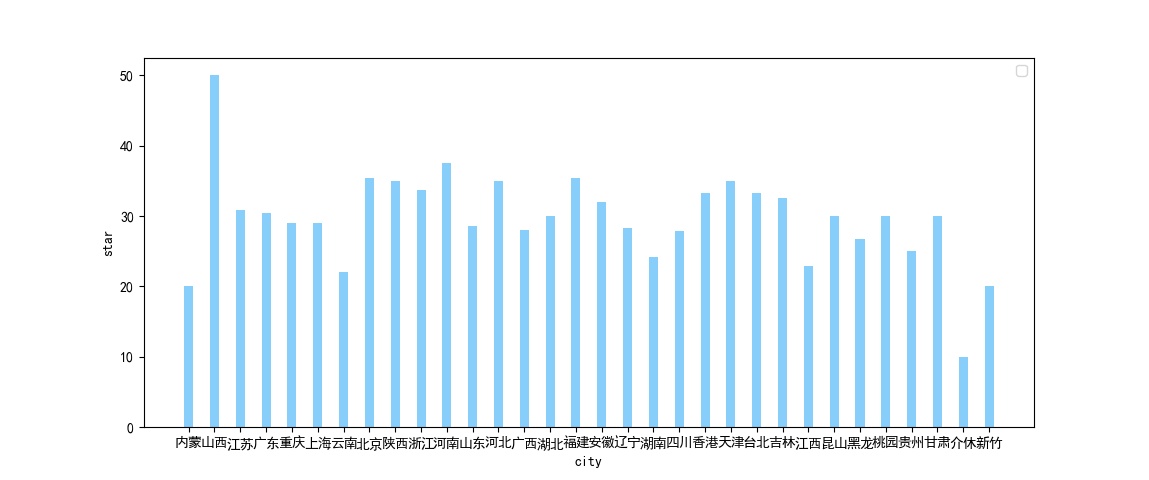


图 5 不同城市评分情况（10代表1星，以此类推）

**4 用户会龄分析**

**4.1 用户会龄分布**

提取出用户加入豆瓣的年份进行统计分析，发现发表复联4短评的用户在2010年加入豆瓣的人最多，而2005年、2014年、2019年的人是最少的。产生这种波动的原因，一个可能是跟复联电影有关，看复联相关电影的人的年龄分布与其加入豆瓣的时候有所关联；一个可能跟豆瓣当年的用户加入数等有关。

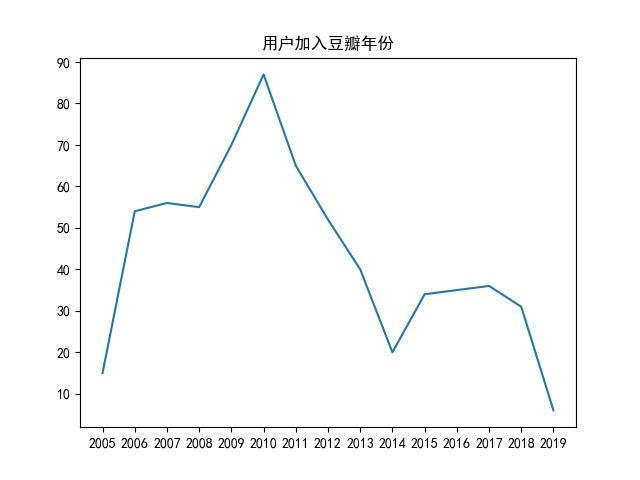


图 6 用户会龄分布

**4.2 会龄对评分的影响**

由图可知，2019年的用户给的平均分最低，2015年入会的用户给的平均分最高，其余年份的用户给分平均分普遍在27.5到32.5中间，总体给分为高分，说明复联4这部电影在大众中口碑还是极佳的。

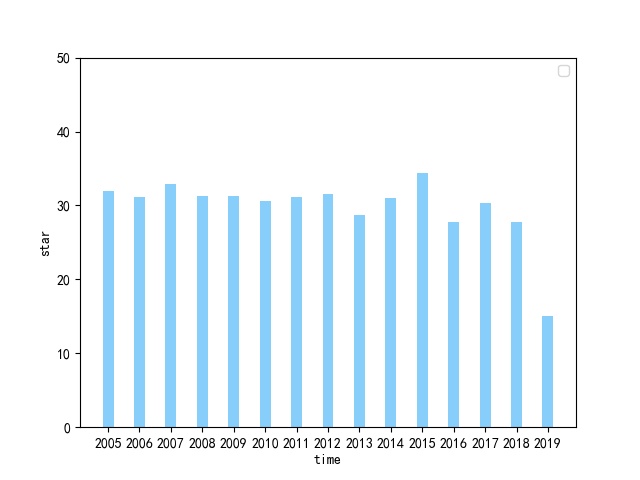


图 7 会龄对评分的影响（10为1星，以此类推）