# 基于情感词典的豆瓣电影评论数据分析

## 一、摘要

本文以豆瓣作为媒介，深刻调研了豆瓣评论的情感得分，以影评的角度去给各种电影打分，给读者以及电影厂商一些另外的角度去思考，总结。

本文第一部分是数据爬取部分，我们根据每部电影的url地址写了一段爬虫代码，爬取了6大类电影，每个类型10部电影，每部电影2000条数据，一共12万条数据。

接下来我们进行了数据预处理，删除了一些未爬取到的空数据，以及重复数据。

然后就是情感分析，我们采用了BosonNLP情感词典，并手动额外添加了一些停用词，程度副词和否定词，扩充了一下情感词典。利用JIEB分词将影评切分。把程度副词当作权值，用以修饰其后跟随的形容词，对每部电影的每条评论进行打分。最后汇总成整部电影的得分。

最后就是可视化部分，我们将各类型的各电影分别打分并可视化，共计60张柱状图。并加以分析，总结，给各位读者以另一种角度审视电影这个行业。

目录

[基于情感词典的豆瓣电影评论数据分析 1](#_Toc11549)

[一、摘要 1](#_Toc16387)

[二、绪论 3](#_Toc18771)

[三、需求分析 3](#_Toc20490)

[四、设计 3](#_Toc23004)

[五、总结 8](#_Toc4276)

[六、参考文献 9](#_Toc9392)

[七、致谢 9](#_Toc24288)

## 二、绪论

随着网络技术的不断发展，微博、论坛等众多网站为用户提供了更加广阔的平台来交流信息、表达意见。这些网站的用户评论信息往往蕴含着巨大的商业价值，不仅反应了用户的情感态度，更反应了社会集体的情感状态。用户评论信息的挖掘可以让其他用户更加了解产品，同时也可以让商家更好地识别用户的产品需求和喜好，提高市场竞争力。

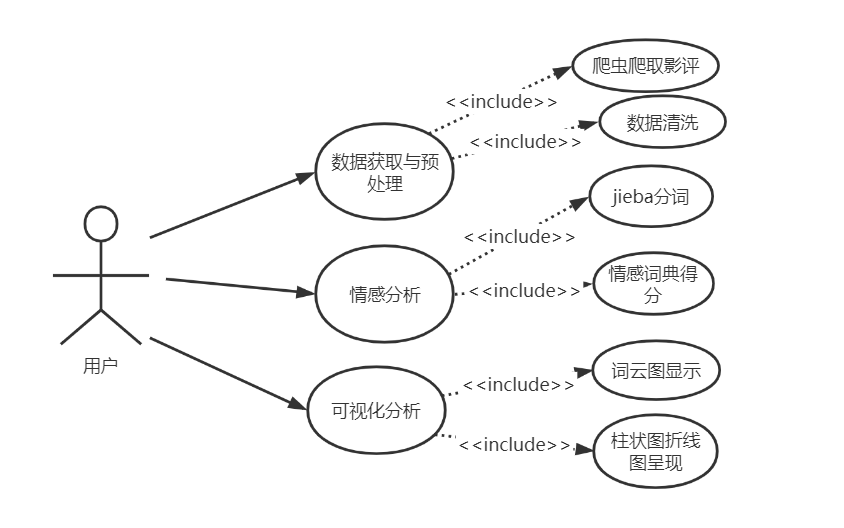
其中，豆瓣网作为中国最大最权威的电影评论网站之一，它对电影的评价在人们选择和认知电影的过程中扮演着非常重要的作用。豆瓣评分直接反应了豆瓣网对电影的评价，但豆瓣评分往往只关注了用户对电影的评分信息，而忽视了用户的评论信息，使得人们看到的最终评分未必能反应这部电影的真实情况。

为了帮助浏览者有效的解读影评文本，了解影评中的情感因素，我们对影评文本进行情感分析并计算得到电影评论的情感得分，通过对电影评论情感得分与豆瓣原生用户评分进行综合对比与分析，使用户更好地了解电影的整体评价。

## 三、需求分析

鉴于网络用户评价已成为人们选择产品或服务时的重要参考指标，旨在了解打分评价和评论评价之间的关系，同时为仅有评论评价的网站提供符合潜在分值的排序和推荐功能。

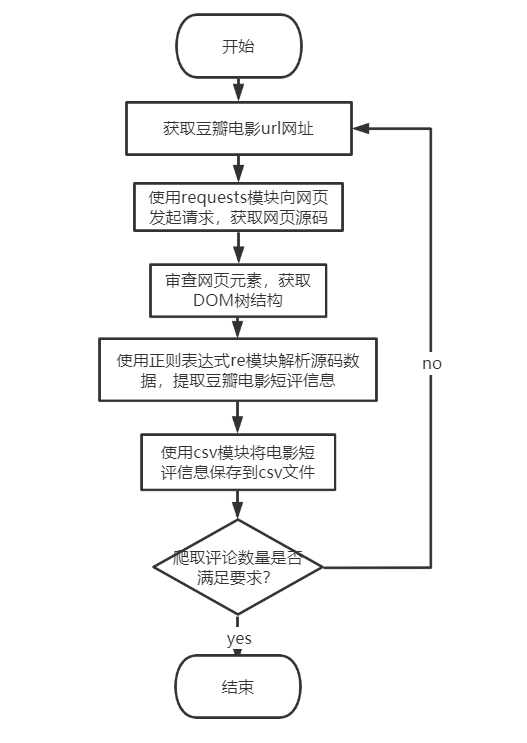
将情感得分结果进行可视化分析，以图表形式展现打分评价和评论评价之间的区别和联系，并从多个维度去考虑电影市场的兴衰与变化，也同时为电影厂商出厂作品提供分析与建议。



## 四、设计

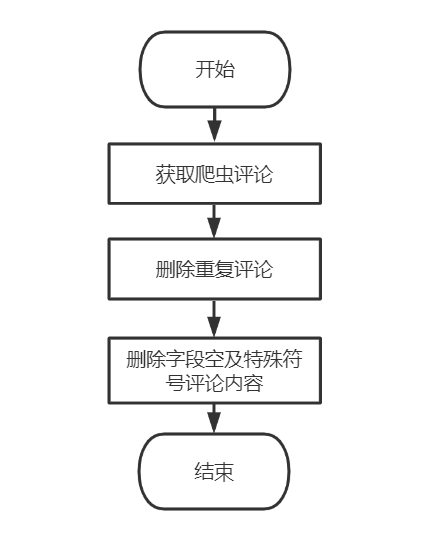
#### 1.数据获取阶段

首先获取豆瓣网站电影相关的url地址，使用requests模块向网页发起请求，获取网页源码。审查网页元素，获取dom树结构，再利用正则表达式匹配我们想要的comment域，提取豆瓣电影的短评信息，由于豆瓣每页的评论数是20，所以我们每部电影爬取100页，根据url地址的limit后缀让程序自动换页并将这些评论存入csv文件。当评论总数达到要求时，将他们改成txt格式。



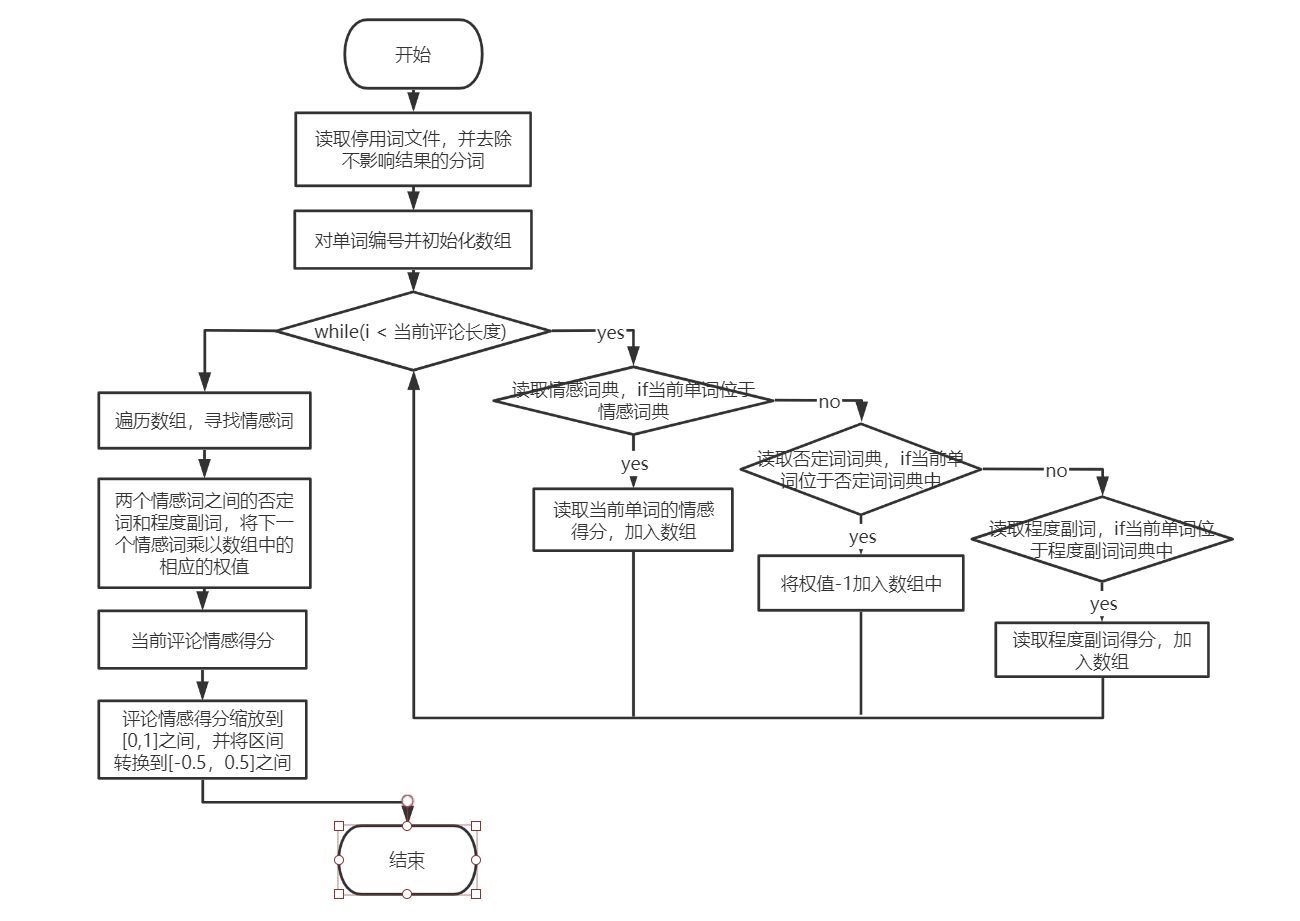
#### 2.数据预处理阶段

将txt格式的数据审查，检查到哪一行和上面一行一致或者字段为空，就将该数据视为含缺失值，并将其删除。



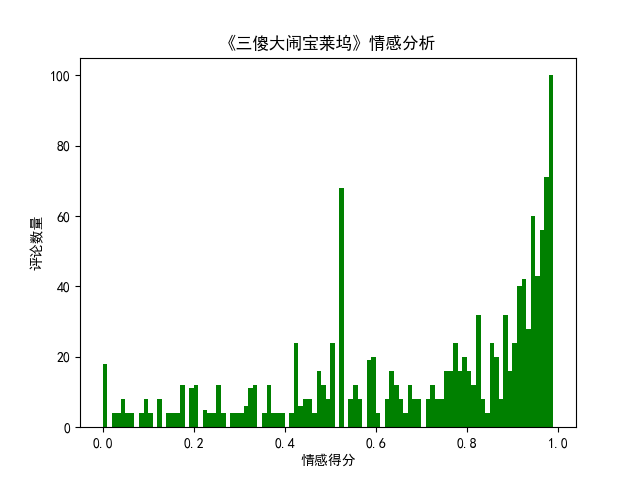
#### 3.情感分析阶段

首先将每条评论进行jieba分词，将每段评论分成单独的词，然后调用我们手动添加的停用词词典，将一些停用词删除，删除之后根据bosonNLP词典定位情感词出现的位置和它的分数，对于每个情感词，其前置的否定词和程度副词可以影响它的得分。查询我们手动添加的否定词和程度副词，否定词出现，就将它乘以-1，程度副词出现，就将它乘以相应的权值，最后汇总情感词的得分，即得出单条评论的得分，区间为0~1。对每部电影每条评论进行打分，该阶段完成。



#### 4.可视化阶段

我们将每部电影的评分进行可视化，以影评情感得分为x轴，影评数量为y轴。生成相应的柱状图。



还可以对每部电影根据分词和去停用词之后生成相应的词云图。



## 五、总结

这次工程实践也让我们获益匪浅，我们在系统的学习情感分析技术的同时，也对电影行业有了更深刻的认识。

对于技术，我们系统的学习了爬虫代码的撰写，情感得分的依据。在原有词典上添加自己编写的词典，并对副词，否定词进行相应的处理，让电影情感评分的得出显得更加客观。我们熟练掌握了情感分析得分，可视化出图的要领，也深刻地学习到了NLP的详细过程。

对于电影，评分较高的电影和评分较低的电影之间也有千丝万缕的联系，从中找到这些联系，找到他们的因果，也是我们这次工程实践最主要的收获。有些因素可能是电影高分的关键，通过大量数据进行可视化分析，从图中找到它们的联系，推断出形成差异的原因，也为电影厂商作出了参考，为电影行业未来的发展做出贡献。

## 六、参考文献

1.情感分析：挖掘观点、情感和情绪 : 挖掘观点、情感和情绪

[美] 刘兵（Bing Liu） / 刘康、赵军 / 机械工业出版社 / 2017-9-20 / 99.00

2.Sentiment Analysis and Opinion Mining Liu, Bing / 2012-5

3.NLP-文本情感分析

<https://blog.csdn.net/weixin_42398658/article/details/85222547>

4.豆瓣评论情感分析

<https://blog.csdn.net/eastmount/article/details/85118818>

## 七、致谢

该工程实践是在我们的导师袁老师的亲切关怀和悉心指导下完成的。他严肃的科学态度，严谨的治学精神，精益求精的工作作风，深深地感染和激励着我。从课题的选择到项目的最终完成，袁老师都始终给予我细心的技术指导和不懈的支持。在此谨向袁老师致以诚挚的谢意和崇高的敬意。

　　在此，我们还要感谢在一起完成该实践的队友和其他组的同学们，正是由于你们的帮助和支持，我们才能克服一个一个的困难和疑惑，直至本实践的顺利完成。

　　在论文即将完成之际，我的心情无法平静，从开始进入课题到论文的顺利完成，有多少可敬的师长、同学、朋友给了我无言的帮助，在这里请接受我诚挚的谢意！谢谢你们！