# 基于情感词典的豆瓣电影评论数据分析

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。

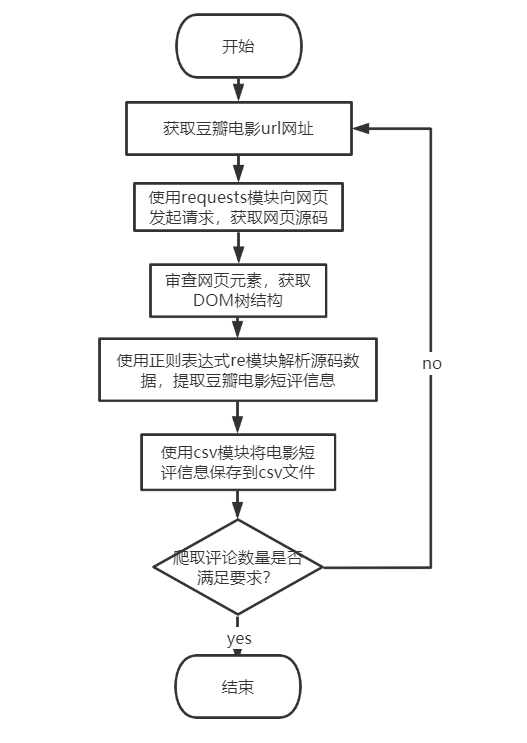
接下来我们进行了数据预处理，删除了一些未爬取到的空数据，以及重复数据。

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。

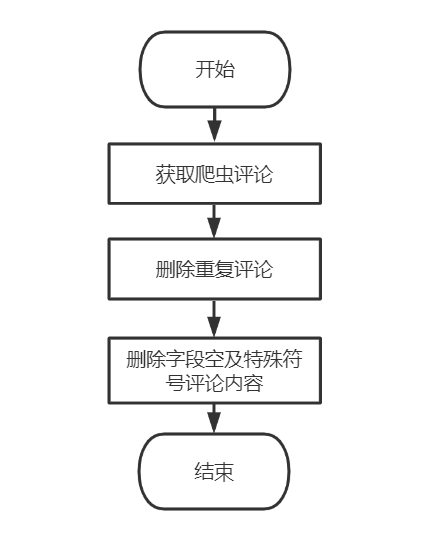
#### 1.数据获取阶段

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。



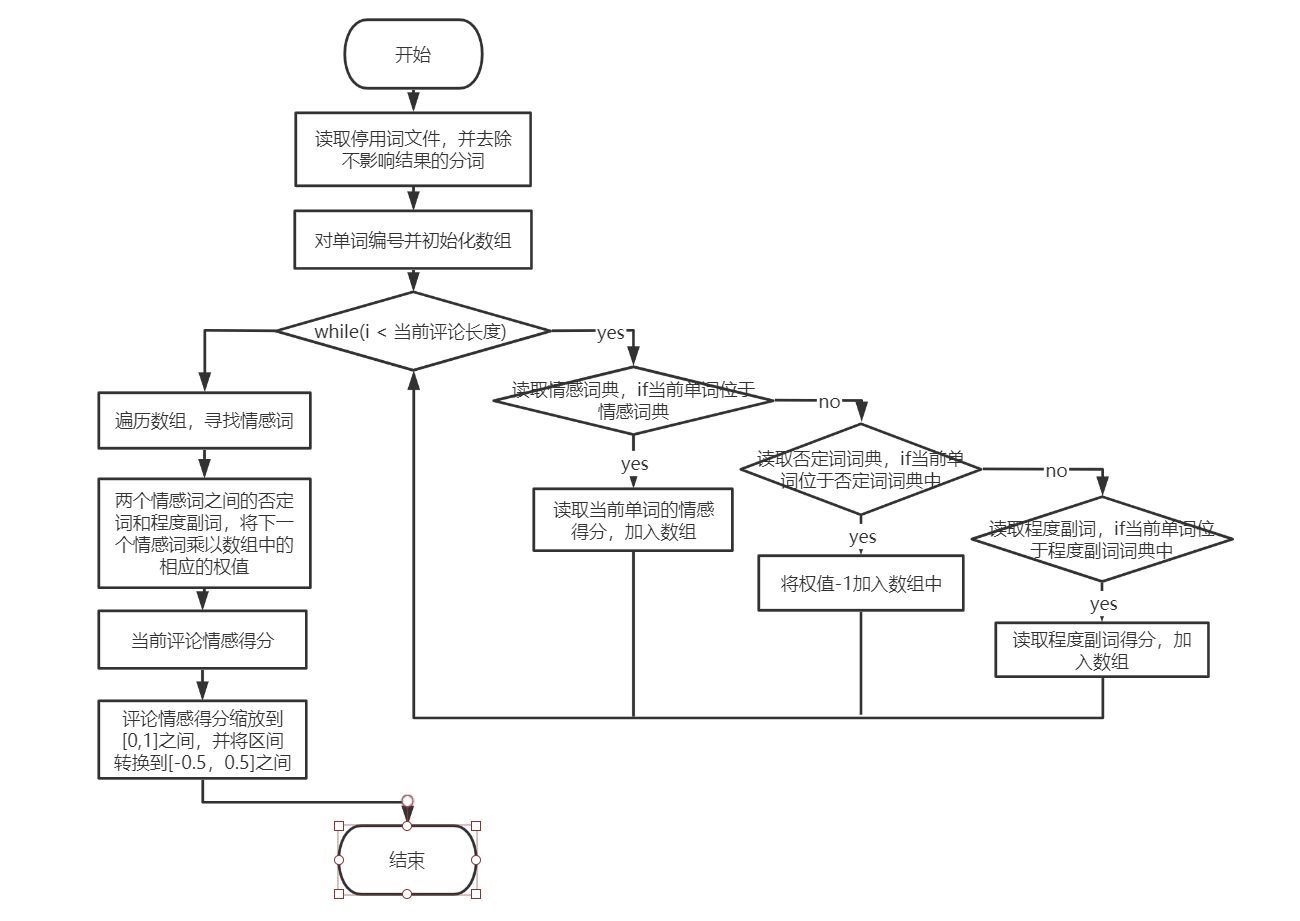
#### 2.数据预处理阶段

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。



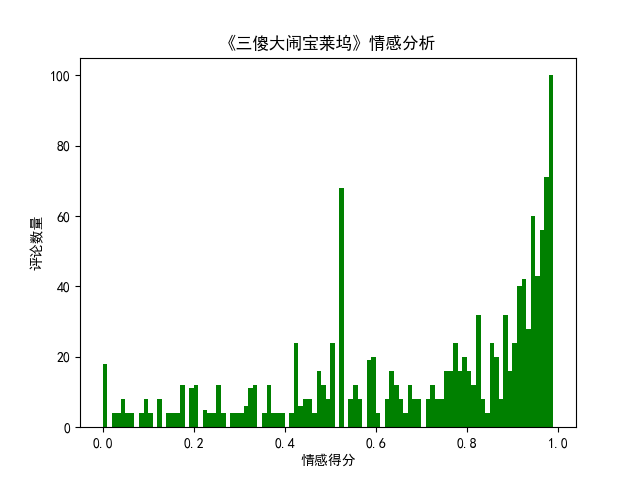
#### 3.情感分析阶段

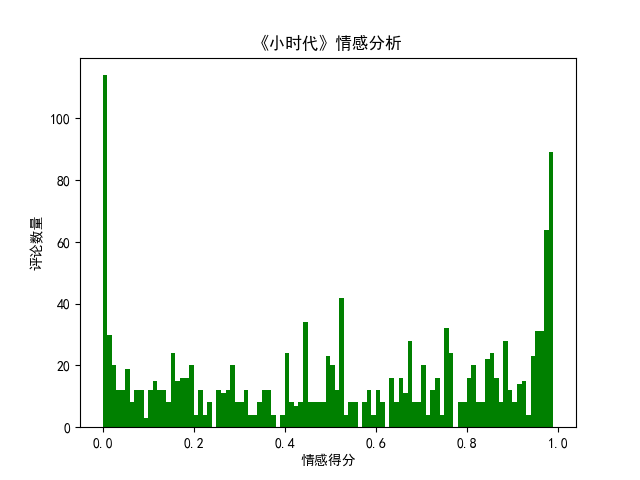
首先将每条评论进行jieba分词，将每段评论分成单独的词，然后调用我们手动添加的停用词词典，将一些停用词删除，删除之后根据bosonNLP词典定位情感词出现的位置和它的分数，对于每个情感词，其前置的否定词和程度副词可以影响它的得分。查询我们手动添加的否定词和程度副词，否定词出现，就将它乘以-1，程度副词出现，就将它乘以相应的权值，最后汇总情感词的得分，即得出单条评论的得分，区间为0~1。对每部电影每条评论进行打分，该阶段完成。



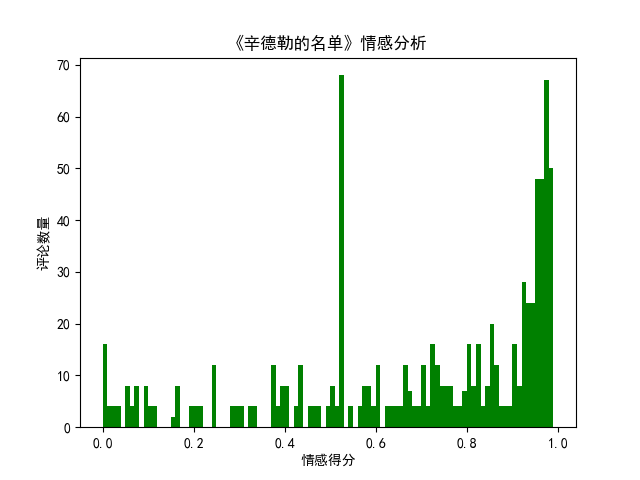
#### 4.可视化阶段

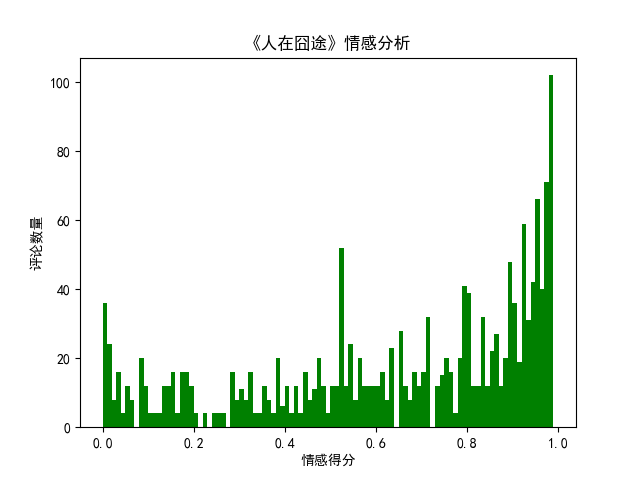
我们将每部电影的评分进行可视化，以影评情感得分为x轴，影评数量为y轴。生成相应的柱状图。



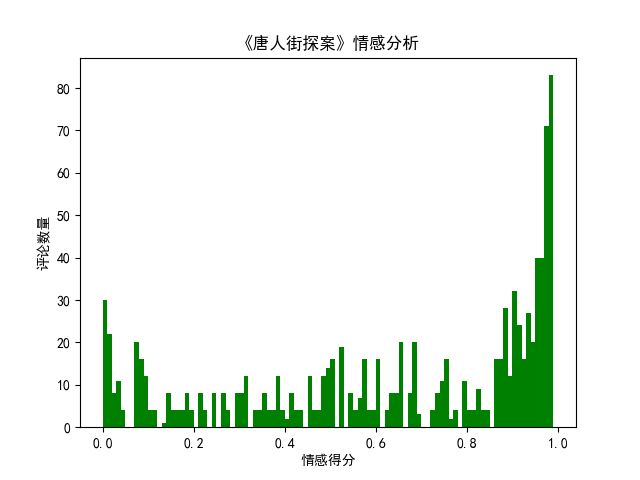


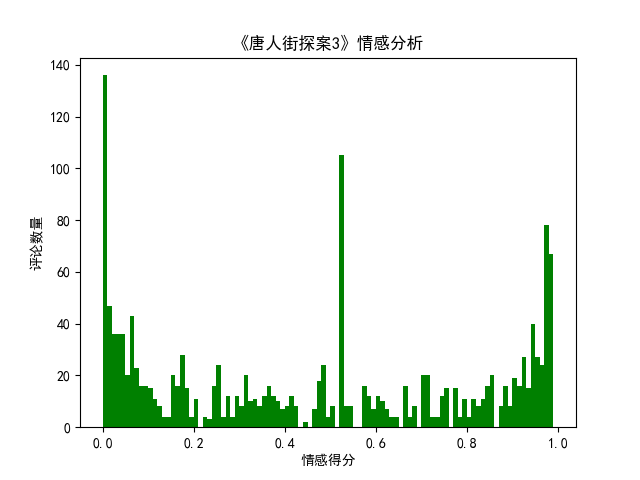
如上所示，根据柱状图的内容我们可以很清晰的看到小时代的低分评价明显多于三傻大闹宝莱坞，这也和我们平常对两部电影的印象大致相仿，小时代作为郭敬明导演的作品，一出来就被贴上了拜金，内容空洞的标签，而三傻作为印度宝莱坞的大作，在幽默的同时穿插着积极向上的能量，是公认好评如潮的大作。情感得分聚集于1.0周围的影评数量相当之多，足以见得电影的优秀，与之相对的是小时代的窘境，评论得分大多聚焦于0分，也显示了电影的拙劣，以及诸多网友评论的客观。





作为口碑都较为不错的电影，辛德勒的名单高分评论肉眼可见的高于人在囧途。这背后的原因也可能是类型的差别，辛德勒的名单作为高分的战争剧情类电影，可能是人们心中对于这种沉痛且真实的历史抱以敬意，加上电影对真实事件的艺术渲染，让很多人感同身受以及反思过往，大多数人会给出0.8以上的情感倾向。而人在囧途作为一部公路喜剧片，片中可能会有一些低俗玩笑，一些人喜欢这些，而另一些人厌恶，他们会给出情感倾向适中的评论。因此该片高分很多，但是得分处于0.4~0.8中位数的也比较多。战争题材的电影相对沉重，大多数人由于尊重历史会给出较高的分，喜剧片相对轻松，很多人给分比较随意，因此处于中位数的影评得分也会较多。





唐人街探案作为近期大火的悬疑+喜剧类型，刚开播得到了观众的一致好评，烧脑的同时不缺幽默，合家欢的同时又不缺硬核。但是由图可见，唐探1的情感得分明显高于唐探3的情感得分，单就同一IP而言，两部电影硬质量其实是相差不大的，导致这个结果的可能原因就是时间因素。当唐探1横空出世的时候，中国电影圈很少有这种喜剧和烧脑结合比较完美的作品，所以得到了一致好评，而唐探3作为今年新作品，不管是烧脑还是喜剧上都很难创新，当观众满怀期待的等待唐探3给他们带来惊喜，却迎来了这种说不上好也说不上坏的感觉，有些粉丝可能觉得IP没有创新，愤愤的打了低分。因此作为两部质量差不多的电影，有些电影充当了第一个吃螃蟹的人，取得了成功，而后来者没有做到突破，只能得到一个平平无奇的评价。

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。



以上是韩国近期火热电影寄生虫的影评词云图，根据该词云图就能推断出这部电影大概讲了什么。寄生虫深刻地揭示了阶级矛盾，电影中充满了隐喻，穷人生活在地下室，而富人生活在独栋大别墅，穷人靠寄生在富人身上生活，却怎么也改变不了阶级差距的现实。根据该词云图也许能为电影厂商做出提醒，这种富含阶级和隐喻类型的电影也许能成为未来电影行业的趋势呢。



以上是电影哪吒之魔童降世的词云图，可见大部分人对这部电影还是认可的。而其中的几个关键字值得注意一下。国产，动画，这两个关键字可能是取得成功的关键，同样的剧情，同样的表现力，配上国产动画的标签可就完全不一样了。众所周知，中国动画市场已经萎靡很长时间了，全方位被日本，美国打压，甚至整个动画都被打上低龄幼儿的标签。而这样一部剧情人设双在线的动画电影毫无疑问的为中国动画电影行业版砖添瓦，观众乐意为这样的电影买单，可能也宣示着中国动画电影的强势回归。

#### 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。