
利用文本挖掘探索
摔跤吧！爸爸
DANGAL
基于豆瓣短评的文本聚类
林禹 14353189

利用文本挖掘探索电影《摔跤吧！爸爸》的成功因素

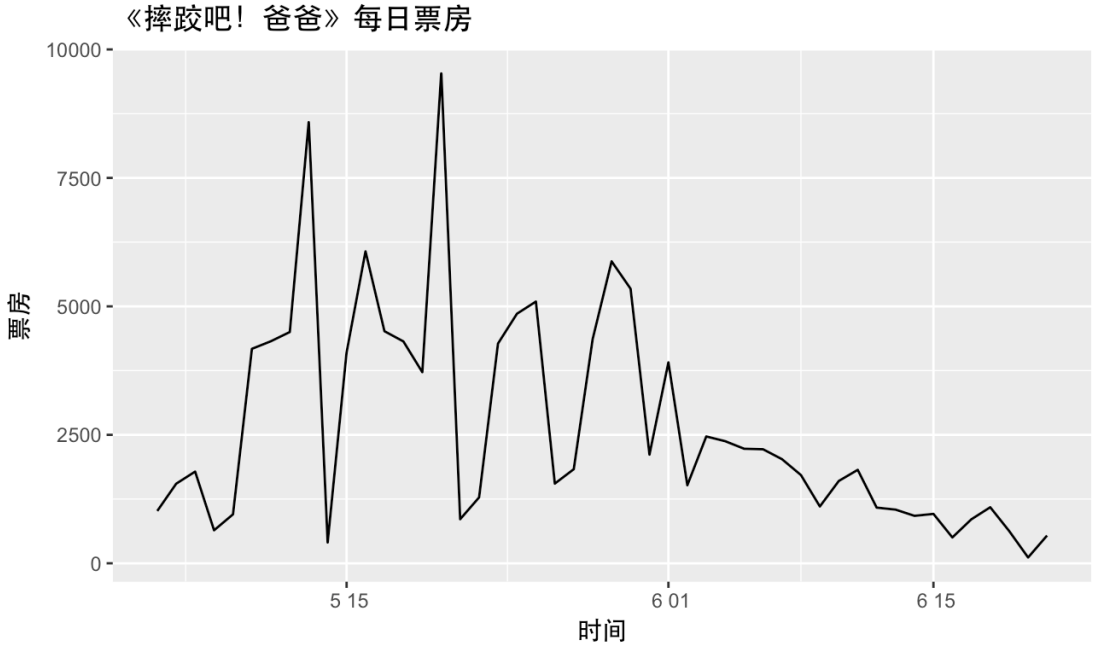
——基于豆瓣短评的文本聚类分析

摘要：5 月 5 日上映的印度电影《摔跤吧！爸爸》在国内上映后引起了巨大反响，获得了广泛的好评。为了探究这部电影的成功之处，本文采用文本挖掘中的文本聚类方法，探究观众对这部电影的关注点，并基于所得到的结果，也对国产电影的发展提出建议。

关键词：电影，文本挖掘，文本聚类，豆瓣电影

一、 选题简介

5 月 5 日，一部名为《摔跤吧！爸爸》的印度电影在国内上映。上映初期，电影票房爆冷，但随着时间推移，媒体、社交平台对这部电影的口碑以及评价的不断提升，越来越多的人进入电影院观看该影片。从票房走势图可以看到，该片的票房峰值出现在第二次高峰，而非第一次高峰。在豆瓣上，这部电影获得了 9.2 分的高分，赢得了广泛的好评。



相比之下，现在的中国电影市场商业化过重，近年来鲜有优秀的国产影片呈现给观众。当前电影市场被严重资本化，为保证投资回报，投资方、制片方常常启用一些当红的流量“小鲜肉”、“花旦”，同时大手笔购买时下热门“IP”，制作

出的电影往往是缺乏灵魂的“快销品”，演员演技乏善可陈，电影内容空洞虚无。

因此，不妨通利用数据挖掘的技术分析《摔跤吧！爸爸》这部优秀的印度电影，探究这部电影的闪光点，同时也为当前乏善可陈的中国电影市场发展提供借鉴。而豆瓣电影作为国内电影爱好者聚集地，用户群体中既有专业的电影工作者，也有普通的网民，他们在这个平台上为电影打分，给出自己对电影的评价，反映着国内广大观影群众的心声，具有很高的代表性。尤其是相比充斥着大量明星粉丝、网络营销、网络水军的微博，豆瓣电影的影评更为客观，是观影用户的真实心声。所以，本文从《摔跤吧！爸爸》在豆瓣电影中 13 万多条的评论中爬取了排名靠前的 4 万条评论数据，利用文本挖掘的技术进行分析，探寻这部电影的优秀之处，也借此反思中国电影的发展现状以及未来。

二、 数据情况

为了获得豆瓣电影上的用户短评，我用 python 编写了豆瓣短评爬虫获得用户名、发表时间、评级、评论支持数量以及评论内容，并将文件存储为 xls 格式。下面是代码片段：

```
Author = "LinYu"

reload(sys)
sys.setdefaultencoding("utf-8")

headers = {
    'User-Agent':'Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/57.0.
    'Referer':'https://accounts.douban.com/login?alias=&redir=https%3A%2F%2Fwww.douban.com%2F&source=index

}
session = requests.Session()
session.headers.update(headers)

def getComment(url):
    #解析网址
    url = url_init(url)
    login(redir = url)
    message = session.get(url,headers = headers)
    names = re.findall('<title>(.*?)</title>',message.content)
    print "指定的爬取任务是%s"%names[0]

    pattern = '<div class="avatar">[\s\S]*?</p>'
    blocks = re.findall(pattern,message.content)

    #获得短评总数
    pattern = "<li class=\"is-active\">[\s\S]*?<span>.....(.*?).</span>"
```

Line 19, Column 2 | Tab Size: 4 | Python

获得的数据总数为 40020 条，其中“评级”这一变量存在缺失值，在数据爬取的过程中已经被标注为 N，其他变量不存在缺失值。

用户名	时间	评级	支持数	评论内容
蝙蝠翔	2016-12-23	"力荐"	15456	去他娘的摔跤吧爸爸 鄙视国内这种通过降低片名档次从而降低印度电影关注度的做法
王静宇	2017-04-16	"力荐"	9790	关于本片价值观一不看社会背景就评判三观就是耍流氓。在印度的社会状况下，女性是没有自由选择职业的氛围的，先破才能立。就像片中父亲对女儿
有点迷人	2017-01-04	"力荐"	9688	阿米尔一人饰演男主角19、29、55岁三个人生段。他首先完成了19岁的戏，然后疯狂增肥，一个月时间变成了体重194斤的55胖大爷，拍完老年戏后，
Allindis	2017-04-11	"力荐"	8646	好看死了。拜托大家认真看完电影再做评论。爸爸有约定一年为期，不行就不勉强。吉塔在赢得了第一场摔跤比赛之后，也主动请爸爸带自己参赛。
DespicableMe	2016-12-24	"力荐"	9562	这么好的一部电影。。弄得这路马什么破译名。。听起来像儿童片似的。。
Jacqueline	2016-12-31	"力荐"	7043	就凭电影，谁也没资格叫人家阿x
同志亦凡人中	2016-12-23	"推荐"	6646	根据真人真事改编，讲述印度一块女子摔跤金牌的诞生。剧情热血逗乐，竟技惊心动魄，挑不出什么毛病的宝莱坞电影，父女情尽在不言中。阿米尔
栗色雪	2017-01-18	"推荐"	4119	豆瓣文青们可能已经成功被白左的西方政治正确洗脑了：不放在具体语境下，以西方女权衡量一切本来就是一种偏见和武断。在整体状况还不如中国
菜问之	2017-05-04	"推荐"	4458	女权癌们归根结底还是她们看不起摔跤这项运动，要是把爸爸让女儿练习摔跤改成鼓励女儿去选美，当模特，变成大明星，估计就不会有那么多愤慨
豆丁	2016-12-31	"力荐"	6713	印度排名第一的电影，全球口碑已炸裂
延延	2017-04-16	"力荐"	3597	#北京电影节#我不觉得这个片子囿于父权的藩篱，父亲虽然有在女儿身上圆梦的渴望，但也建立在发现女儿的天分的前提上。摔跤让两个女孩子挣脱
大奇特(Grinch	2017-04-16	"力荐"	2961	（看的完整版）取材真实，剧作扎实，用体育竞技片来拍父女情节剧，并改写了上到封建礼教、下到官僚制度的不变法则。有人提出片中受父权压制
豆花酱不爱馋	2017-05-05	"力荐"	2664	一群骂父亲剥夺女儿未来的白左真令人作呕。我们现在有资格成为女运动员女演员女记者女程序员，是因为有牺牲童年牺牲安逸的开路人。女儿对未来
Mr.çŒ«	2016-12-24	"力荐"	3175	阿米尔汗就是印度电影的金字招牌，而且他每部电影都反映了印度方方面的问题，从《三傻》反映的教育，《我神》反映的信仰，到这部《摔跤》前一
阿草的斯托卡	2017-02-05	"力荐"	2243	父权你mb啦，一个能发现儿女自身优点并开发出来的这个也是父母责任之一啊。反推现在的世界冠军们哪个小时候过得是自由自在吃吃喝喝的日子啊。
é°Ÿ	2017-01-04	"还行"	1440	在印度看得169分钟，一部处处衬着男权和父权的女性电影，女儿没有任何选择权，被父亲残暴地教育成为世界冠军，这个冠军的“正确”结果就意味着
蟑螂	2017-01-07	"较差"	1102	父亲价值观简直令人作呕，以梦想、金钱、冠军强制女儿人生，所谓父女情不如说是满足私欲后的奖励。电影讲述打破性别偏见，其实自己已经偏见
翻滚吧！蛋堡	2017-01-10	"还行"	834	父亲堪称年度最讨厌角色，女性意识的觉醒不是建立在自我的基础上，而是在父权的逼迫下强制产生，这种价值感实在不敢苟同。此外，本片将体育
眠去	2017-05-07	"力荐"	1513	中国真是太多中华田园伪女权了，丢人

三、 分析过程

爬虫所获得的数据维度囊括了豆瓣短评页面的所有元素，基于这些获得的文本内容可以进行多方面的文本挖掘，如情感分析、聚类分析等。本次分析的最初构想是，通过对评价内容进行无监督的情感分析，得出正向情感词语以及负向情感词语，以此探索出这部电影令人满意的要素以及令人失望的原因。但是，这部电影的积极评价过高，负面评论只占极少数，进行情感分析只能得到显著的正向结果。然而，正向的结果同样可以通过解析文本，对文本进行聚类，发现观众的关注点以及类别实现，所以最终决定使用文本聚类方法进行文本挖掘。

文本聚类的算法通常划分为分割方法、层次方法、基于密度的方法以及基于网格的方法。这里简单介绍其中常用的分割方法中的 K-MEANS 算法：

K-MEANS 算法指的是从 n 个数据对象任意选择 k 个对象作为初始聚类中心，并根据其他对象与这些聚类中心的相似度（距离），分别将它们分配给与其最相似的（聚类中心所代表的）聚类，然后再计算每个所获新聚类的聚类中心（该聚类中所有对象的均值），不断重复这一过程直到标准测度函数开始收敛为止。一般都采用均方差作为标准测度函数. k 个聚类具有以下特点：各聚类本身尽可能的紧凑，而各聚类之间尽可能的分开。

具体过程:

(1)选取 k 个对象作为初始的聚类种子;

(2)根据聚类种子的值,将每个对象重新赋给最相似的簇;

(3)重新计算每个簇中对象的平均值,用此平均值作为新的聚类种子;

(4)重复执行(2)、(3)步,直到各个簇不再发生变化。

K-means 算法的研究对象通常为向量，不能直接处理文本，所以，在文本聚类的过程中,其实是将分词后的结果转化为文本空间向量,再进行文本聚类操作。所以，在文本聚类中，过程可以表述为：

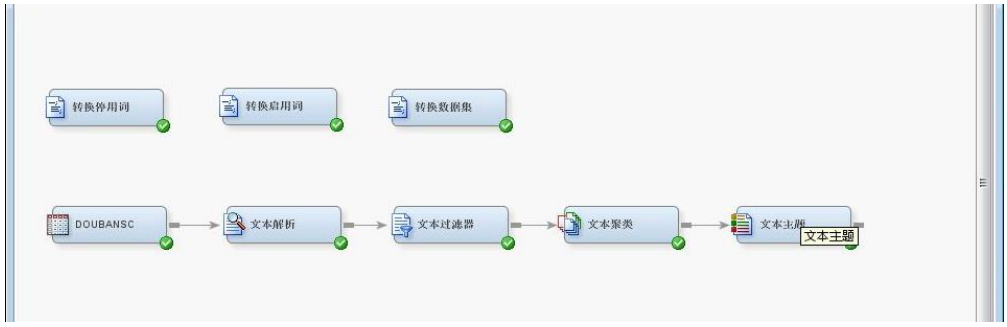
（1）从 N 个文档随机选取 K 个文档作为质心

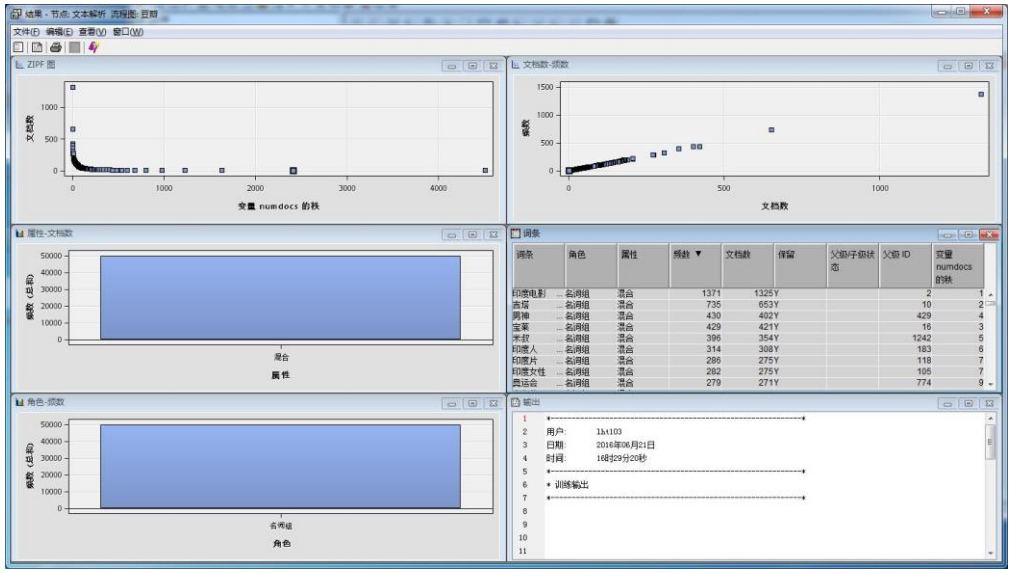
（2）对剩余的每个文档测量其到每个质心的距离，并归到最近的质心的类

（3）重新计算已经得到的各个类的质心

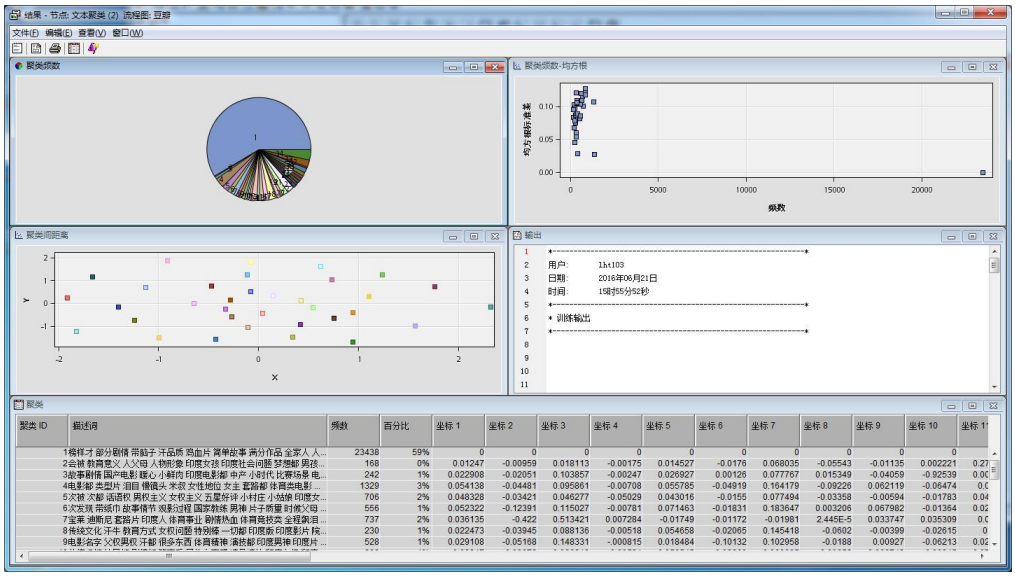
（4）迭代 2～3 步直至新的质心与原质心相等或小于指定阀值，算法结束

由于文本聚类不需要验证和测试，所以不需要对原始数据进行数据分区，按照思路，最终在 SAS 中建立的流程图如下：





在文本解析的节点，设置了保留的词性为动词、形容词、名词、专有名词，可是最终文本解析的结果却只剩下了名词，导致后面的文本聚类以及文本主题的结果都是对名词的聚类。



接下来的文本聚类，节点设置为最大聚类数 45，最大词数 15。

同时，在以 SAS 分析为主体的基础上，用 R 语言辅助进行分析工作。

与 SAS 不同，在 R 中需要自行对文本进行数据清洗，包括去停用词，去无实义单词等，这一点既是优点也是缺点。好处是灵活度高，缺点是不够便捷。

米	5860
摔跤	5157
汗	5138
太	4876
爸爸	4155
女儿	4024
真	3985
志	3917
励	3885

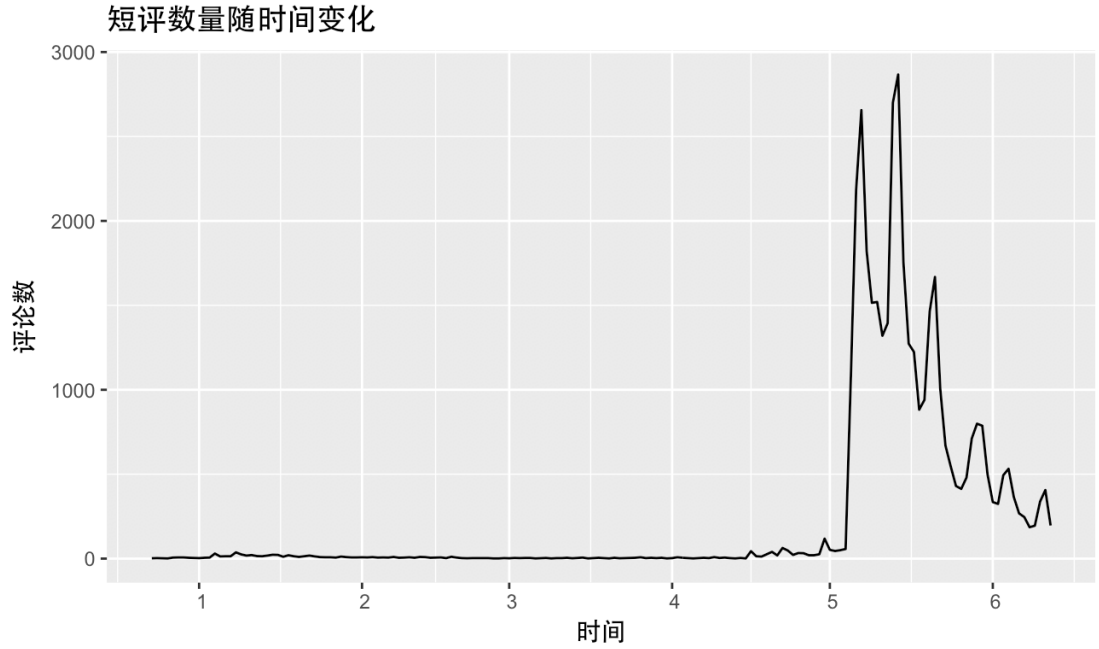
对评论内容进行数据清洗之后（去停用词），进行词频排序，很明显可以看到，主演“阿米尔汗”的名字被分词拆开了，“印度电影”也被拆分成了两个词语，还有“励志”也被拆开了，所以在分词前先插入自定义字典，重新进行分词的步骤，再次进行词频统计。这样的情况同样出现在 SAS 的文本解析结果中，可是 SAS 节点对中文的支持存在一定局限，进行文本解析的时候无法导入自定义词典，所以这样的情况难以改善。这样的限制带来的结果是在得到文本聚类以及文本主题的结果中会出现一些残缺的成分，甚至是无意义的成分，比如说“每次都”，观众在影评中经常出现这一词组，而 SAS 却将该短语解析成“每”“次都”，结果是每字被当作数词忽略掉，而“次都”被保留了下来。不过好在根据这些残缺的词语，能够推测出该词语本身，若是无意义的词语，可以在结果分析中忽略，当然，这些存在于文本聚类中的残缺词语，也影响了文本聚类的准确性。

电影	12936
印度	7250
摔跤	5157
阿米尔汗	4241
爸爸	4155
女儿	4024
励志	3877
父亲	3527
哭	3180
好看	3063

SAS 中的文本解析结果较难更改，但是在 R 语言中，还是比较好干涉文本的数据清洗工作的。在加入了自定义字典（主要包括了演员的名字，一些网络用语）之后，进行分词，去停用词之后，得到词频统计排行有所改观，但是仍然存在一些问题，所以手动进行错词删除，将右图中类似的，单字无意义但词频较高的字从列表中删减（主要删除了词频排名前 240 条的数据，因为 R 语言绘制词云主要选取排名靠前 240 的词语，排名 240 之后的词语对最终所绘制的词云影响不大）。

```
> wordrank = subset(wordrank,Var1!="太")
> wordrank = subset(wordrank,Var1!="真")
> wordrank = subset(wordrank,Var1!="点")
> wordrank = subset(wordrank,Var1!="说")
> wordrank = subset(wordrank,Var1!="片")
> wordrank = subset(wordrank,Var1!="想")
> wordrank = subset(wordrank,Var1!="拍")
> wordrank = subset(wordrank,Var1!="一部")
> wordrank = subset(wordrank,Var1!="中")
> wordrank = subset(wordrank,Var1!="里")
> wordrank = subset(wordrank,Var1!="片子")
> wordrank = subset(wordrank,Var1!="完")
> wordrank = subset(wordrank,Var1!="星")
> wordrank = subset(wordrank,Var1!="真的")
> wordrank = subset(wordrank,Var1!="高")
```

四、 分析结果



简单统计短评的发表时间，可以看到，《摔跤吧！爸爸》于 2016 年 12 月 23 日印度上映，当时已经有网友看过该电影并且在豆瓣上给出了电影的短评，而到了 5.5 号电影在中国大陆上映之后，评论数迅速增加，并且是在上映过后的几天，评论数才出现了最高峰，这也验证了文章开头所提到的，这部电影在中国上映一段时间后的关注度高于刚上映时的关注度。

文本聚类的结果中，最令人感兴趣的是第一条聚类结果，频数高达 23438，

描述词		频数 ▼
榜样才 部分剧情 带脑子 汗品质 鸡血片 简单故事 满分作品 全家人 人热泪 傻差 印度少女 印度主旋律电影 创造奇迹 主旋律影片 巴胺	...	23438
残粉 大陆版 五分好评 印度电影 中国电影 歌舞剧 音乐棒 部电影都 好莱坞套路 经典电影 运动员都 韩国电影 体育竞技电影 题材都 好多泪	...	1360
电影都 类型片 泪目 慢镜头 米叔 女性地位 女主 套路都 体育类电影 印度国宝 印度良心 印度電影 广电 迟暮 大荧幕	...	1329
表演都 超级棒 传记片 大神 电影人 方面都 故事本身 故事都 汗男神 汗叔 可以拍 平权 商业片 社会意义 神片	...	862
奥运会 奥运金牌 爸爸妈妈 比赛都 大段 很多电影 刻板印象 米汗 社会背景 体育类 印度背景 印度歌舞 印度国歌 印度人都 中国人	...	847

在上文展示的聚类结果饼图中，占比 59%，可见多数人都属于这一类群，大家对这部电影的评价中提到了“鸡血片”[1]、“满分作品”、“创造奇迹”、“印度主旋律电影”、“多巴胺”（描述词中是，巴胺，但是根据推测应该是多巴胺），可见观众中多数人喜欢这一类宣扬正能量的热血电影。

从频数排名第二的聚类中可以看到，“歌舞剧”、“音乐棒”，表明这一类用户比较强调印度电影中巧妙融合歌舞音乐的特点，这也是印度电影的一大特色，同时，观众也强调了“体育竞技电影”、“运动员”，表明观众对运动这一题材也颇具好感。

[1] “鸡血片”是网络用语，指故事很励志，很鼓舞人心

此外从聚类结果中还能提炼出来的另一类用户，是比较关注电影的社会现实

男权主义 女权主义 五星好评 小村庄 小姑娘 印度女性地位 点俗套 很多影评 印度女权问题 非常热血 印度女子 男权电影 ... 706

意义的，喜欢探讨电影中所折射的社会环境，如大家提到的“男权主义”、“女权主义”、“印度女性地位”、“印度女性问题”，这种具有现实意义的电影也是为观众所津津乐道的。

下图展示的是文本主题的结果，其中最大的问题是“烂片”一词多次出现在

主题	类别	文档截止值	词条截止值
印度人,印度电影,宝莱,拍电影,商业片	多个词条	1.781	0.464
宝莱,印度电影,烂片,中国电影,印度片	多个词条	1.610	0.442
男癌,直男,癌电影,很多人,女权主义	多个词条	1.304	0.403
国产电影,宝莱,印度电影,印度电影都,神片	多个词条	0.887	0.331
人都,很多人,男癌,印度电影,吉塔	多个词条	0.842	0.325
全片,片都,尬舞,宝莱,印度电影	多个词条	0.803	0.318
世界冠军,印度女性,吉塔,生儿子,宝莱	多个词条	0.794	0.316
电影本身,社会意义,宝莱,印度人,国产片	多个词条	0.775	0.313
印度女性,吉塔,故事情节,社会地位,体育精神	多个词条	0.637	0.283
正能量,传记电影,烂片,全片,商业片	多个词条	0.601	0.275
小女孩,印度人,家庭主妇,印度电影,社会环境	多个词条	0.586	0.272
观影,商业片,吉塔,米叔,主旋律电影	多个词条	0.587	0.273
电影都,男神,烂片,宝莱,米叔	多个词条	0.586	0.271
部电影,整部,印度人,整个故事,宝莱	多个词条	0.576	0.272
鸡血,宝莱,印度人,体育精神,印度电影	多个词条	0.572	0.269
男权社会,男权主义,社会环境,女权电影,印度女人	多个词条	0.481	0.250
印度女孩,教育方式,男权主义,体育精神,中国人	多个词条	0.476	0.247
中国人,印度人,故事情节,教育方式,烂片	多个词条	0.486	0.247
爸妈,米叔,吉塔,电影都,人都	多个词条	0.475	0.244
国家队教练,宝莱,体育精神,平权,戏剧冲突	多个词条	0.464	0.244
印度电影,国产片,男神,故事情节,烂片	多个词条	0.521	0.252
女权主义,男权主义,主旋律电影,体育精神,男神	多个词条	0.471	0.247
男主,社会环境,教育方式,男权主义,整个故事	多个词条	0.445	0.239

不同主题中，对这样一部好评如潮的电影来说，并非令人满意的结果。探究原始评论，发现多数短评在褒奖这部电影的同时，也会联系中国电影进行对比，多数表达了中国电影多是“烂片”的观点，当然也有对一些认为这部电影是烂片的观

宝莱,印度电影,烂片,中国电影,印度片　　印度电影,国产片,男神,故事情节,烂片

点的否定，所以烂片一词才会多次出现在文本主题中。

从文本主题中可以看到，评论的主要主题可以归结为两个，一个是专注于印度电影与中国电影的比较，另外一个是探讨电影展现的现实意义，包括男权、女权、直男癌等。反倒是描述自己观影体验的内容较少，当然这也和 SAS 在文本解析中将形容词也自动忽略的原因密切关联。因为从词频生成的文字云中可以看到，不少用户都提到了“感人”、“泪目”、“热血”一类的词语。

从词云中可以看到，观众最经常提到的就是这部电影的国别属性。同时，主演阿米尔汗的名字也被多次提到，实际上“阿米尔汗”包含许多其他的同义词，如观众会称其为“阿米尔·汗”、“米叔”等，这说明人们对主演关注度非常高，也表明主演确实成为了这部电影一个出彩的地方。不仅如此，人们对演员的努力更是给予了莫大的肯定，在词云中体现的是“摔跤”“比赛”一词，稍稍浏览原始评论数据就可以发现，人们常提到电影中拍摄的摔跤比赛片段有多么精彩，主

演为了学习摔跤付出了多大的努力，这都是“摔跤”一词高频出现的原因。

此外，观众也常常描述自己的观影体验，包括“热血”、“燃”、“感动”、“哭”等等，这从一个侧面反映出电影不仅成功地向人们传达了积极正向的价值观，也凭借自身的故事情节打动了观众。观众热烈地讨论着电影本身的故事剧情，包括“爸爸”和“女儿”之间的情感、冲突等，人们也多次提到了故事描绘出的“励志”地成为优秀摔跤运动员的体育“梦想”。

五、 结果讨论与建议

从上文的分析中，可以归纳出以下几点结论与建议：

一是观众喜爱这种宣扬正能量的电影，这一点不论是在文本聚类、文本主题还是文字云中都有所体现。《摔跤吧！爸爸》这部电影根据真实事件改编，讲述了女儿最终替爸爸实现摔跤梦以及整个国家的奥运会摔跤比赛冠军梦的故事，也正是这样一个热血、励志的故事深深感染了观众。反观近年来的中国电影市场，这种同时涉及到梦想、励志、热血的电影屈指可数，甚至可以说是一片空白，从去年至今，电影院确实有上映过口碑较好的国产热血影片，如《湄公河行动》，也有优秀励志的国产影片，如《滚蛋吧！肿瘤君》《夏洛特烦恼》，但却找不出一部将梦想、热血、励志融合在一起的优秀国产电影。而实际上，中国电影市场并不缺乏这一类型的影片，如传达出“美国梦”的好莱坞超级英雄系列，不过，此类电影更吸引人的地方是高科技手段制作的特效以及动漫英雄故事本身的情怀，

鲜有因其本身故事感染力强而广受赞誉。在这种情况下，中国观众遇到这部热血励志的印度电影的时候，自然如久旱逢甘露一般，深深地被这部电影的内容所吸引，观众内心长久以来的空缺得到了填补。

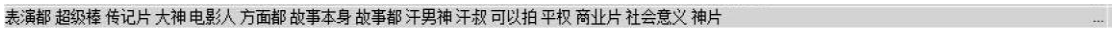
从这一点来说，未来的中国电影可以挑选融合梦想、热血、励志的故事剧本，改变之前温和的“心灵鸡汤”，为观众献上这样的“热血鸡汤”。但是，选择类似的主题绝对不是将故事内容生搬硬套，这种翻拍国外电影的行为往往是观众所厌恶的。推出相同主题的电影同时也存在一个问题，那就是观众在已经看过《摔跤吧！爸爸》这样优秀的电影之后，是否还会对相似主题的电影感兴趣。不过，电影类型空缺确实在一定程度上助力《摔跤吧！爸爸》强烈的市场反响，这一点启发当前中国电影发行方不妨用大数据的手段，去发掘市场近年来空缺的电影类型，以及广受好评的电影类型。在决定拍摄之前，先基于科学手段分析市场，更能创造出人民群众喜闻乐见的优秀电影作品。

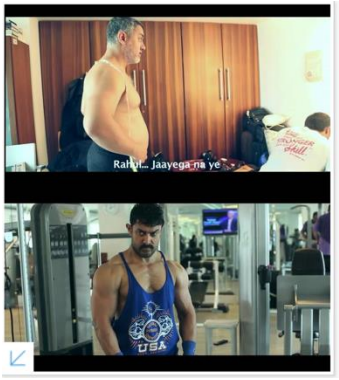
二是观众喜欢有深度的电影，尤其是与社会大环境密切关联，反映社会现实的电影。《摔跤吧！爸爸》的价值观引起了很激烈的讨论，片中的父亲从某种意义上来说确实“强迫”自己的女儿实现了自己的梦想，但联系印度重男轻女的社会背景，正是父亲的教导使得两个女儿能够摆脱多数印度女性年轻就嫁人的命运，破除女子不如男的论调，从这一点来看，影片是对印度女性地位低下的批判，价值观本身并不存在问题。不论是文本聚类还是文本主题，都可以看到结果中“男权”“女权”“社会环境”被多次提及，可见当今观众中存在许多能够解读电影背后的现实意义的人群，仍然有不少人在追求电影的艺术性与理性。近年来中国各种题材的电影，剧情空洞无趣的居多，有深度的占少数，能反映社会现实的更是寥寥无几：早已泛滥的文艺剧情片，以及过年贺岁档才能偶尔见到的科幻大作。甚至回溯这几十年来的国产电影，能说出的有社会深度的，只有寥寥几部，如《霸王别姬》、《活着》。

造成这种现象的原因是多方面的，但归结起来主要有两个，一方面是资本控制电影市场，注重回报，时下热门 IP 自然就会成为资本的宠儿，于是各类时下热门小说，尤其其中的网络言情小说，由当下的流量明星担任主演，纷纷被改编成电影电视剧；另外一方面与国家电影体制有关，国家对电影内容的监管过于严格，一些涉及社会敏感现实的电影不可能获得发行，很大程度上阻碍了电影艺术

的自由表达。但像《人民的名义》这样现实的电视剧已经走上了电视屏幕，我们也有理由相信这是一个良好的开端。所以导演应该用心挑选剧本，不论是翻拍文学经典，或者是创造剧本，一旦拥有了一个有深度有内涵的剧本，尤其是具有强烈现实意义的剧本，可以说整部电影已经成功了一半。

三是电影的成功很大一部分取决于主演，不仅仅是主演演技的问题，更能感染观众的，是主演的敬业精神。文本聚类中有一类就是关注电影主演（如下图中的汗叔、汗男神实是观众对阿米尔汗的称呼），观众提到主演并不是因为他们是





该演员的粉丝，更多的是在评论中盛赞他们为了这部电影所做出的努力。为了能更真实地饰演青年的父亲与老年体态发福的父亲，主演阿米尔汗暴增 28 公斤完成年老发福的父亲的戏份，后在 5 个月时间狂甩 25 公斤，完成电影中父亲年轻时的戏份。不仅仅是男主角，电影中饰演大女儿和二女儿的每一位女演员（电影讲述了女儿从小到大的成长，所以两个女儿人生的每一个阶段每一位角色都是由不同的演员饰演的），为了此片更是付出良多，尤其是饰演成年时二女儿的演员桑亚，为了电影里仅仅出现不到一分钟的摔跤片段，训练了九个月。当观众看到电影中精彩而真实的摔跤场景，自然会联想到演员们背后所付出的汗水与努力。

反观目前国产电影，多数被起用的明星不过是所谓的“流量明星”，除了所谓的“颜值”、粉丝数之外，没有任何可圈可点之处。这种电影在选角公示之后，往往就已经失去了一半的观众，也失去了内容质量的保证。当然，中国也不乏优秀敬业的老戏骨，如成龙，但年轻一代确实缺乏具有职业精神的代表人选，也有很大的可能是这种兢兢业业的演员从商业角度来看，不如那些“流量明星”的投资回报前景明朗，毕竟“流量明星”演技再差，总会有疯狂的粉丝为之买单。但这样的电影，只会是资本市场合格的商品，无法称为电影发展史上宝贵的艺术品。所以，要想真正打造一部优秀的国产影片，制片方在选角色时不仅要选择演技佳的演员，更应该选择真正敬业的演员，也许他们的票房前景看似没有流量明星明

朗，但他们的却能保证整部影片的质量。

总而言之，《摔跤吧！爸爸》电影的成功有许多值得国产电影学习的地方，它的成功不仅限于文本挖掘所发现的因素，还有其他未被发掘因素，是多种因素共同作用的结果。在当今，电影的风靡程度已经成了一个国家文化软实力的象征标志，商业上的成功不过是随之而来的副产品。在经历了近年来国产优质影片的匮乏，国内观众内心势必渴望着优秀的国产影片的再次出现，而我国巨大的电影市场也正好为优秀电影作品的出现提供了广阔的发展空间。

参考文献：

[1]胡小莉,李波,吴正鹏. 电影票房的影响因素分析[J]. 中国传媒大学学报(自然科学版),2013,(01):62-67+39.

[2]周昭涛. 文本聚类分析效果评价及文本表示研究[D].中国科学院研究生院（计算技术研究所）,2005.

[3] Tan A H. Text mining: The state of the art and the challenges[C]//Proceedings of the PAKDD 1999 Workshop on Knowledge Disocovery from Advanced Databases. sn, 1999, 8: 65-70.

[4] Basuroy S, Chatterjee S, Ravid S A. How critical are critical reviews? The box office effects of film critics, star power, and budgets[J]. Journal of marketing, 2003, 67(4): 103-117.

[5] Chintagunta P K, Gopinath S, Venkataraman S. The effects of online user reviews on movie box office performance: Accounting for sequential rollout and aggregation across local markets[J]. Marketing Science, 2010, 29(5): 944-957.