2019年度热门电影分析报告

一、背景介绍

本文爬取2019年度在豆瓣有评分的1100余部电影的豆瓣和猫眼数据作为数据基础，选取2019年度票房过亿的前十部电影作为研究对象。同时爬取了《哪吒之魔童降世》的豆瓣评论及豆瓣影评，采用Sentiment Classify算法，对影片评论的情感倾向加以计算，同时给出豆瓣评分。

二、指标设计

电影数据：

1. 电影名称：该变量指某部电影在中国的名称。
2. 电影票房：该变量指某部电影在中国的票房收入的累积量，单位：万（亿）元。
3. 上映日期：该变量指某部电影的上映年份、月份、日。
4. 时长：指某部电影的播放时长，单位：分钟。
5. 评分：指某部电影的豆瓣评分。
6. 情感极性：消极、中性、积极。

评论数据：

1. 评论内容：用户所评论的内容。
2. 所给分数：用户在豆瓣上给影片的评价，0-10的一位小数。
3. 情感极性：消极、积极、中性。
4. 属于积极类别的概率，取值范围[0,1]。
5. 属于消极类别的概率，取值范围[0,1]。

三、电影总体分析

数据分析过程

1. 爬取了https://movie.douban.com/tag/#/?sort=U&range=0,10&tags=2019,%E7%94%B5%E5%BD%B1
2. 由于此网站没有评分上映时间等信息，利用该网站每个电影的url，爬取了所有电影的数据。包括：
    - 电影名称
    - 评分
    - 上映时间
    - 时长
    - 每个星级占比

```
class movieInfo:
    def __init__(self):
        self.stars = [0.0, 0.0, 0.0, 0.0, 0.0, 0.0]

    title = ""
    rate = 0.0
    url = ""
    releaseDate = ""
    runTime = 0
    stars = [0.0, 0.0, 0.0, 0.0, 0.0, 0.0]
```

3. 将所有数据写入Excel。

```
def writeExcel():
    workbook = xlrd.open_workbook("2019movies.xls")
    rowNum = workbook.sheets()[0].nrows
    newbook = copy(workbook)
    newsheet = newbook.get_sheet(0)
    for i in range(0, len(movieList)):
        row = []
        row.append(movieList[i].title)
        row.append(movieList[i].runTime)
        row.append(movieList[i].rate)
        row.append(movieList[i].releaseDate)
        row.append(movieList[i].stars[1])
        row.append(movieList[i].stars[2])
        row.append(movieList[i].stars[3])
        row.append(movieList[i].stars[4])
        row.append(movieList[i].stars[5])
        for j in range(0, len(row)):
            newsheet.write(rowNum + i, j, row[j])
    newbook.save("2019movies.xls")
```

4. 由于豆瓣没有票房属性，在时光网爬取票房属性后，匹配Excel中的电影并填入票房。
5. 利用pyecharts库做出”电影数量—电影票房“与”电影评分分布“柱状图，并用snapshot_phantomjs导出图像。

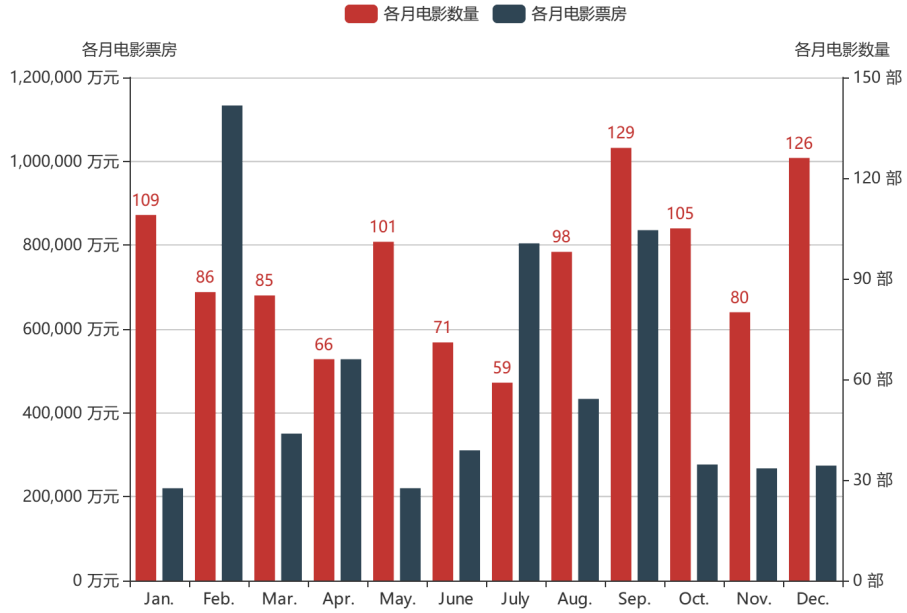
如下是导出的Excel文件

1	哪吒之魔童降世	110分钟	8.5	2019-07-26(中国大陆)	0.7%	2.0%	14.7%	39.2%	43.4%
2	流浪地球	125分钟	7.9	2019-02-05(中国大陆)	2.3%	4.8%	22.5%	38.3%	32.1%
3	我和我的祖国	155分钟	7.7	2019-09-30(中国大陆)	0.9%	3.1%	27.1%	46.8%	22.1%
4	少年的你	135分钟	8.3	2019-10-25(中国大陆)	1.1%	1.8%	15.2%	46.5%	35.5%
5	误杀	112分钟	7.7	2019-12-13(中国大陆)	0.7%	3.1%	26.4%	52.2%	17.5%
6	寄生虫	132分钟	8.7	2019-05-21(戛纳电影节)	0.3%	0.6%	8.5%	43.1%	47.4%
7	复仇者联盟4：终局之战	181分钟	8.5	2019-04-24(中国大陆)	0.8%	1.9%	15.5%	35.5%	46.4%
8	1917	119分钟	8.5	2020-08-07(中国大陆)	0.1%	0.8%	11.4%	49.3%	38.5%
9	天气之子	112分钟	7.1	2019-11-01(中国大陆)	1.2%	7.4%	40.1%	36.4%	14.9%
10	中国机长	111分钟	6.7	2019-09-30(中国大陆)	2.2%	11.5%	44.6%	32.0%	9.7%
11	诛仙 I	101分钟	4.5	2019-09-13(中国大陆)	39.2%	22.6%	20.6%	8.5%	9.1%
12	飞驰人生	98分钟	6.9	2019-02-05(中国大陆)	1.9%	9.5%	41.6%	36.0%	11.0%
13	白蛇：缘起	99分钟	7.9	2019-01-11(中国大陆)	0.5%	3.0%	24.9%	46.3%	25.4%
14	小丑	122分钟	8.7	2019-08-31(威尼斯电影节)	0.7%	1.2%	10.6%	38.4%	49.1%
15	疯狂的外星人	116分钟	6.4	2019-02-05(中国大陆)	4.1%	15.7%	45.0%	27.0%	8.1%
16	惊奇队长	124分钟	6.9	2019-03-08(美国/中国大陆)	1.2%	7.3%	47.8%	34.0%	9.7%
17	利刃出鞘	130分钟	8.2	2019-11-29(中国大陆)	0.2%	0.9%	16.5%	54.5%	28.0%
18	冰雪奇缘2	103分钟	7.2	2019-11-22(美国/中国大陆)	0.8%	6.4%	41.6%	36.3%	14.9%
19	罗小黑战记	101分钟	8.2	2019-09-07(中国大陆)	0.6%	2.6%	19.0%	43.1%	34.7%
20	老师·好	111分钟	6.7	2019-03-22(中国大陆)	2.0%	10.9%	46.7%	31.2%	9.2%
21	婚姻故事	136分钟	8.6	2020-05-20(中国大陆网络)	0.1%	0.6%	10.3%	47.5%	41.5%
22	蜘蛛侠：英雄远征	127分钟	7.7	2019-06-28(中国大陆)	0.6%	3.0%	27.5%	46.8%	22.1%
23	半个喜剧	111分钟	7.4	2019-12-20(中国大陆)	1.0%	4.6%	32.8%	46.4%	15.3%
24	狮子王	118分钟	7.3	2019-07-12(中国大陆)	0.7%	4.8%	37.3%	40.7%	16.5%
25	银河补习班	147分钟	6.3	2019-07-18(中国大陆)	5.6%	17.9%	41.8%	24.6%	10.1%
26	玩具总动员4	100分钟	8.6	2019-06-21(美国/中国大陆)	0.1%	0.8%	12.8%	42.5%	43.7%
27	82年生的金智英	118分钟	8.6	2019-11-07(中国香港)	0.2%	0.8%	11.3%	45.7%	42.0%
28	小妇人	135分钟	8.1	2020-08-25(中国大陆)	0.3%	2.0%	19.6%	49.2%	29.0%
29	受益人	112分钟	6.6	2019-11-08(中国大陆)	1.3%	9.5%	52.3%	30.9%	6.0%
30	叶问4：完结篇	107分钟	6.9	2019-12-20(中国大陆)	1.6%	8.8%	43.4%	35.1%	11.0%
31	速度与激情：特别行动	137分钟	6.3	2019-08-23(中国大陆)	2.2%	14.0%	54.7%	23.9%	5.3%
32	沉睡魔咒2	119分钟	6.1	2019-10-18(美国/中国大陆)	4.3%	20.4%	49.1%	19.7%	6.5%
33	烈火·英雄	120分钟	6.5	2019-08-01(中国大陆)	4.2%	13.9%	45.0%	26.5%	10.4%
34	使徒行者2：谍影行动	98分钟	6.3	2019-08-07(中国大陆)	2.1%	15.0%	53.5%	23.9%	5.5%
35	驯龙高手3	104分钟	7.5	2019-03-01(中国大陆)	0.4%	3.6%	36.0%	43.1%	16.9%
36	新喜剧之王	91分钟	5.7	2019-02-05(中国大陆)	12.7%	25.4%	35.6%	17.6%	8.7%
37	一条狗的使命2	108分钟	7.0	2019-05-17(美国/中国大陆)	1.0%	8.1%	45.4%	32.4%	13.2%
38	“大”人物	107分钟	6.5	2019-01-10(中国大陆)	2.0%	11.6%	52.2%	28.3%	5.9%
39	X战警：黑凤凰	114分钟	5.8	2019-06-06(中国大陆)	4.4%	24.2%	51.4%	16.1%	3.9%
40	阿丽塔：战斗天使	122分钟	7.5	2019-02-05(中国台湾)	0.7%	4.6%	32.5%	43.3%	18.9%
41	乔乔的异想世界	108分钟	8.4	2020-07-31(中国大陆)	0.2%	1.2%	13.9%	47.5%	37.2%
42	大侦探皮卡丘	104分钟	6.5	2019-05-10(美国/中国大陆)	1.8%	12.6%	51.5%	26.4%	7.6%
43	攀登者	125分钟	6.1	2019-09-30(中国大陆)	5.1%	18.9%	47.8%	22.0%	6.2%
44	饥饿站台	94分钟	7.8	2019-09-06(多伦多电影节)	0.4%	2.1%	24.7%	51.7%	21.2%
45	宠爱	108分钟	6.0	2019-12-31(中国大陆)	5.7%	20.6%	47.1%	20.6%	6.1%
46	扫毒2天地对决	99分钟	6.0	2019-07-05(中国大陆)	3.1%	20.7%	54.4%	17.5%	4.3%
1076	爆裂老兵	92分钟	4.7	2019-09-21(美国)	16.9%	41.5%	33.5%	4.6%	3.5%
1077	醉酒夫妻	97分钟	4.6	2019-04-19(美国)	18.5%	42.1%	31.6%	5.7%	2.0%
1078	血量子	96分钟	4.6	2019-09-05(多伦多电影节)	15.7%	44.8%	34.9%	2.3%	2.3%
1079	异世浮生	89分钟	4.6	2019-08-23(美国)	23.2%	35.5%	32.3%	7.1%	1.9%
1080	天灵盖	102分钟	4.5	2019-04-19(越南)	22.7%	39.6%	29.6%	6.2%	1.9%
1081	致命狙杀	89分钟	4.5	2019-10-24(中国大陆)	19.4%	41.8%	35.8%	1.5%	1.5%
1082	猎鬼姐妹	105分钟	4.5	2019-04-04(泰国)	22.9%	39.1%	30.7%	4.5%	2.8%
1083	局外人	0	4.5	2019-06-14(美国)	24.1%	38.4%	31.3%	2.7%	3.6%
1084	非常市长	115分钟	4.4	2019-08-30(威尼斯电影节)	22.3%	47.9%	17.0%	10.6%	2.1%
1085	绑定	90分钟	4.5	2019-07-05(美国)	23.0%	39.2%	28.4%	9.5%	0.0%
1086	化为灰烬	0	4.4	2019-07-19(美国)	24.9%	43.3%	24.6%	3.5%	3.8%
1087	监狱13	88分钟	4.4	2019-08-30(日本)	25.6%	40.9%	24.7%	4.2%	4.7%
1088	丰岛园	81分钟	4.4	2019-05-10(日本)	28.0%	37.1%	25.2%	7.0%	2.8%
1089	八子	120分钟	4.3	2019-06-21(中国大陆)	42.3%	22.5%	20.7%	7.9%	6.6%
1090	兴风作浪	85分钟	4.3	2019-01-11(中国大陆)	26.4%	41.3%	25.3%	5.2%	1.9%
1091	响尾蛇	85分钟	4.4	2019-10-25(美国)	24.0%	42.7%	27.0%	4.2%	2.2%
1092	我家的执事如是说	90分钟	4.3	2019-05-17(日本)	30.2%	37.6%	24.3%	4.5%	3.4%
1093	陷阱之中	90分钟	4.3	2019-10-31(意大利)	18.5%	51.8%	27.2%	2.1%	0.5%
1094	虎胆凤威	77分钟	4.3	2019-07-26(新加坡)	26.3%	45.3%	19.0%	5.8%	3.6%
1095	追随者	98分钟(圣	4.2	2019-01-27(圣丹斯电影节)	25.4%	49.6%	18.7%	4.4%	2.0%
1096	难以置信的怪物	84分钟	4.2	2019-03-09(Cinequest电影节)	28.2%	43.1%	20.1%	6.3%	2.3%
1097	大三元	111分钟	4.2	2019-02-01(中国台湾)	35.9%	33.1%	20.0%	4.8%	6.2%
1098	死亡水域	90分钟	4.1	2019-07-26(加拿大)	25.0%	48.5%	22.1%	2.9%	1.5%
1099	灰狗攻击	80分钟	4.2	2019-04-09(美国)	57.4%	13.1%	6.6%	9.8%	13.1%
1100	绝命47小时	91分钟	4.1	2019-10-04(加拿大)	29.1%	42.4%	23.5%	3.3%	1.6%
1101	第16集	93分钟	4.1	2019-06-28(美国)	33.2%	35.8%	24.8%	4.0%	2.2%
1102	怨灵古堡	83分钟	4.1	2019-12-27(美国)	29.7%	46.2%	16.9%	4.1%	3.1%
1103	匿名杀手	95分钟	4.0	2019-06-28(美国)	33.3%	40.5%	22.0%	2.7%	1.5%
1104	的士惊魂	90分钟	4.0	2019-08-26(英国)	33.6%	42.1%	18.7%	1.9%	3.7%
1105	新封神之哪吒闹海	95分钟	3.9	2019-05-14(中国大陆)	42.2%	36.7%	11.0%	6.4%	3.7%
1106	午夜之后	0	3.9	2019-04-26(Tribeca Film Festiva	35.2%	39.8%	20.3%	3.1%	1.6%
1107	滚蛋吧，大魔王！	79分钟	3.7	2019-02-20(中国大陆)	54.7%	24.0%	9.3%	4.0%	8.0%
1108	黑色圣诞节	92分钟	3.7	2019-12-13(美国)	42.0%	36.9%	16.0%	2.3%	2.8%
1109	齐天大圣之大闹龙宫	0	3.5	2019-02-01(中国大陆)	54.7%	24.4%	15.1%	4.7%	1.2%
1110	流氓战争	103分钟	3.5	2019-10-04(美国部分上映)	45.2%	39.3%	10.7%	2.4%	2.4%
1111	尘行者	0	3.3	2019-10-12(澳大利亚)	56.7%	26.9%	11.9%	1.5%	3.0%
1112	宋慈洗冤录	81分钟	2.9	2019-06-11(中国大陆)	62.7%	28.0%	9.3%	0.0%	0.0%

基于上映时间与票房属性分析

9月和12月上映电影数量非常多，1月、5月、8月和10月上映数量其次。可以看出影片更倾向于选择暑假、十月黄金月，12月跨年月，以及春节档上映。

但其中2月的票房得益于去年春节爆火的《流浪地球》与春节黄金档期众多优秀电影，票房遥遥领先。7月和9月分别得益于《复仇者联盟4：终局之战》以及《哪吒之魔童降世》，票房也遥遥领先其他月份。

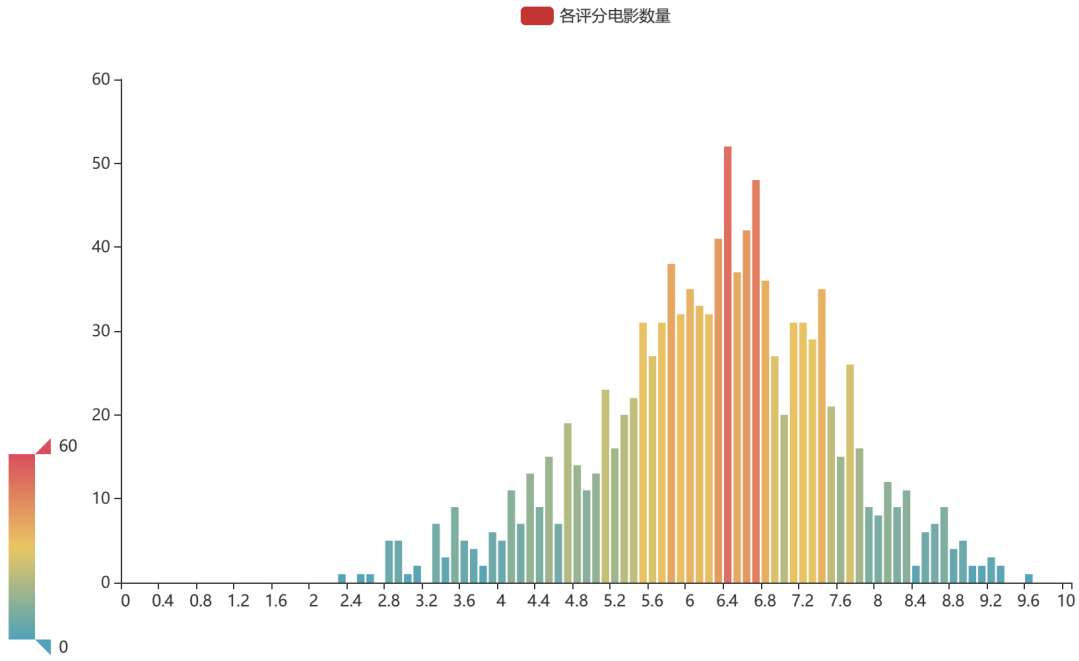


基于豆瓣评分分析

如此多的电影，对于用户来说，其价值如何呢?

我根据豆瓣评分对电影进行了分析。豆瓣评分从0-10，由于评分为0的记录有可能是没有用户评分，因此这里将评分限制在大于0的区域。

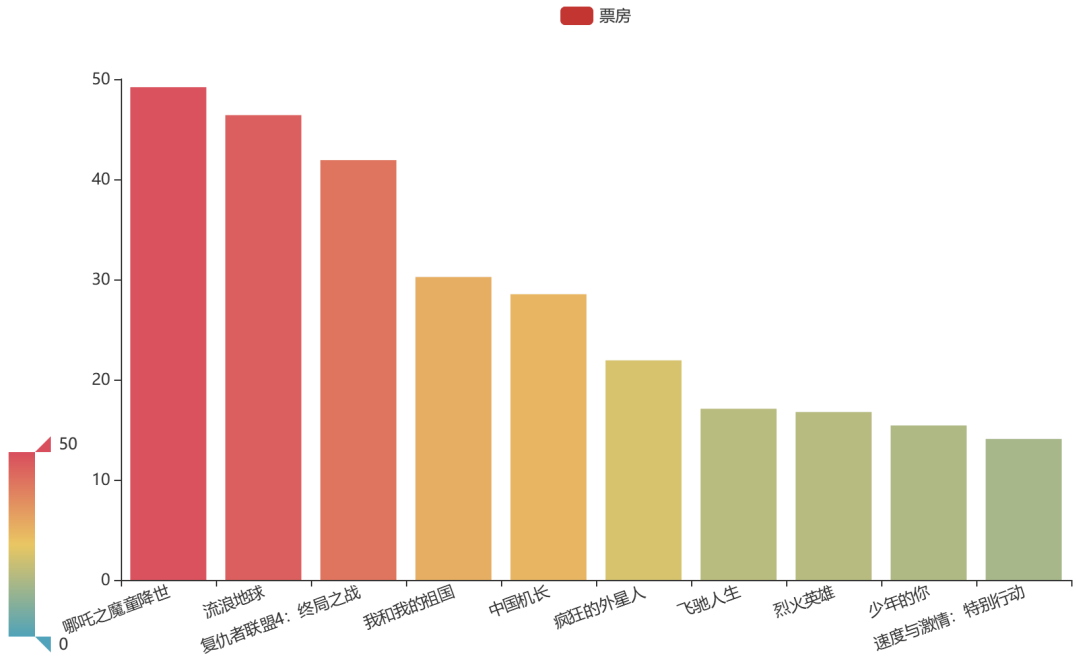
总的来讲，平均评分为6.6分，根据各个评分的数据，绘制了如下分布图。电影评分整体上还是呈现出正态分布的形式，中心在7.0左右。



四、票房前十电影分析

电影名称	票房（亿元）	上映时间
哪吒之魔童降世	49.19	07.26
流浪地球	46.40	02.05
复仇者联盟4：终局之战	41.91	04.24
我和我的祖国	30.24	09.30
中国机长	28.52	09.30
疯狂的外星人	21.92	02.05
飞驰人生	17.09	02.05
烈火英雄	16.76	08.01
少年的你	15.42	10.25
速度与激情：特别行动	14.07	08.23

可以看出，热门电影集中在春节档、暑期档和国庆档上映，偶尔在其他时间爆出热款。



五、《哪吒之魔童降世》分析

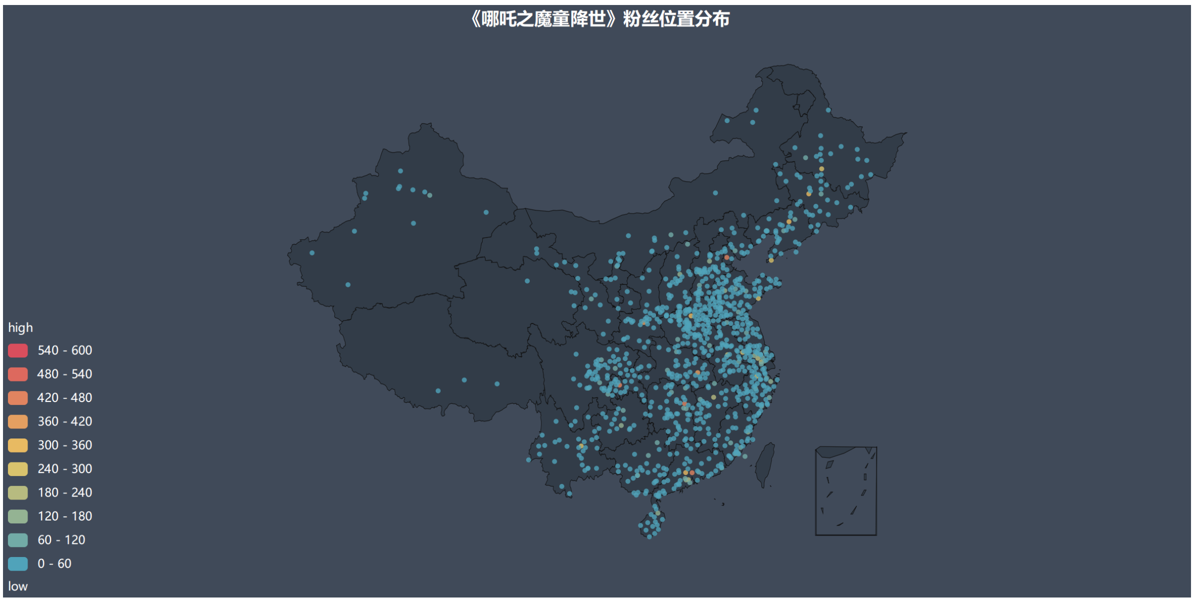
> 上映第 1 天：89分钟，中国动画最快破 1亿纪录。
>
> 上映第 2 天：中国影史首部单日票房破 2亿的动画电影。
>
> 上映第 4 天：中国影史第66部破 10亿影片！
>
> 上映第 8 天：正式登顶！破 16亿，超过《疯狂动物城》，创中国影史动画电影票房新纪录！
>
> 上映第 9 天：成为中国影史第 17 部破 20亿 影片！
>
> 上映第 10 天：破 23亿！连续10天单日票房过亿，连续10天获得单日票房冠军！

由于评论前排优先显示好评，为了避免评论分析造成的误差太大，本文爬取3万余条《哪吒之魔童降世》的评论，着重对该创造历史的影片进行全面的分析。

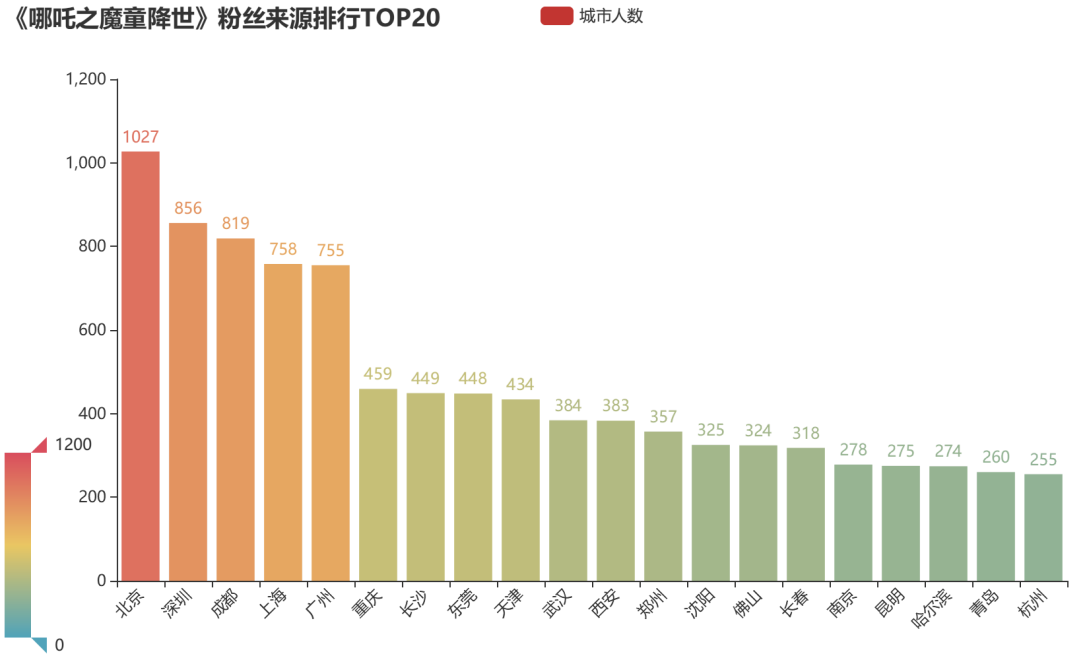
	_id.$oid	cityName	nickName	user_id	movieId	gender	content
2	5d492c89b701b97668b50c31	沛县	gJD876480704	1076095436	1211270	暂无	真的无敌推荐国漫巅峰必须二刷三刷(^〇^)
3	5d492c89b701b97668b50c32	上海	纸海拾珍	1076103821	1211270	1	我把哪吒当作自己了 特别感动
4	5d492c89b701b97668b50c33	眉山	那一寸光	1076107294	1211270	2	有笑点有泪点，看完引人沉思，不错的动漫
5	5d492c89b701b97668b50c34	漯河	漯河小美女吖	1076103818	1211270	暂无	好看好看好看请给我一个敖丙的男朋友
6	5d492c89b701b97668b50c35	合肥	luo11611	1076102889	1211270	1	好看，看看，好看……
7	5d492c89b701b97668b50c36	上海	宗宗	1076095434	1211270	2	挺好看的动画片，和老姐最后一部
8	5d492c89b701b97668b50c37	西安	Lanvy···	1076105461	1211270	暂无	超级好看。。。
9	5d492c89b701b97668b50c38	广州	Strange.	1076074373	1211270	1	**，太好看了。（吃了没文化的亏）
10	5d492c89b701b97668b50c39	南宁	凌晨	1076103808	1211270	2	总体还可以吧，个人不是很喜欢这类型的电影，孩子4岁多，带着一起去看，定不下来，孩子不喜欢看，没看完就要回家😂吵得我没能认真看，带眼镜不舒服，不带看不清，还有哪吒的配音不喜欢，场面壮观
11	5d492c89b701b97668b50c3a	青岛	玉浩装饰老宋	1076105455	1211270	1	棒极了！推荐观看！
12	5d492c89b701b97668b50c3b	蒙自市	郭的旅程	1076095427	1211270	1	故事完整，特效不错！客观的来说，出彩点就是哪吒的塑造和父母的情感。敖丙桥段相对过于仓促，结局后劲不足。是国产一个新的动画票房标杆，对动画行业来说有不小的促进作用
13	5d492c89b701b97668b50c3c	深圳	fdu189656425	1076082245	1211270	暂无	很好很棒很不错
14	5d492c89b701b97668b50c3d	靖西市	7	1076107280	1211270	1	李靖完美诠释父爱
15	5d492c89b701b97668b50c3e	深圳	LOV叶叶子	1076102869	1211270	暂无	好看！！！！
16	5d492c89b701b97668b50c3f	涟水	qca603449189	1076105450	1211270	暂无	好看，好看，好看
17	5d492c89b701b97668b50c42	南宁	幽幽	1076095417	1211270	暂无	3D效果不是很好
18	5d492c89b701b97668b50c43	厦门	夜慢慢其修远兮	1076098268	1211270	2	敖丙哪吒给我磕死！！！
19	5d492c89b701b97668b50c44	高州	🐟	1076105446	1211270	1	好看，有节奏感
20	5d492c89b701b97668b50c45	常宁市	夕颜ღ	1076105445	1211270	暂无	超级好看，国漫的崛起啊！！！
21	5d492c89b701b97668b50c46	衢州	lmQ224945255	1076105444	1211270	暂无	好看😉可以
22	5d492c89b701b97668b50c47	金华	ASL	1076097277	1211270	1	太好看了！！！超震撼！！！期待第二部！支持啊啊啊！！
23	5d492c89b701b97668b50c48	郑州	Just me	1076103790	1211270	2	整体观影 氛围比较 放松 有 笑点又不 失深刻 我命由我 不由天
24	5d492c89b701b97668b50c49	眉山	_qqnd41434162736	1076082240	1211270	暂无	我命由我不由天
25	5d492c89b701b97668b50c4a	东莞	牧歌尽天下!	1076104304	1211270	暂无	挺好看的！影票花得值！
26	5d492c89b701b97668b50c4b	六安	柳柳	1076098267	1211270	2	很好看，儿子也喜欢
27	5d492c89b701b97668b50c4c	南宁	My demons	1076082239	1211270	2	哪吒和敖丙特别好看😍
28	5d492c89b701b97668b50c4d	太原	qjz613224607	1076097286	1211270	暂无	二刷准备中，很燃，很爆，好看到无法形容，没看的小伙伴们，强烈推荐
29	5d492c89b701b97668b50c4e	包头	张小Q	1076107257	1211270	2	燃，国漫崛起
30	5d492c89b701b97668b50c50	云浮	刺猬与玫瑰Love	1076104298	1211270	暂无	很好看，喜剧搞笑。
31	5d492c89b701b97668b50c51	济南	汉溢文谦	1076098264	1211270	暂无	爱死(□°з°)-♡
32	5d492c89b701b97668b50c52	南昌	去年网恋被骗八千	1076105430	1211270	暂无	贼**好看哦
33	5d492c89b701b97668b50c53	文昌	北茗颐春-代远道	1076107251	1211270	1	神话与现代精神内核的结合，体现出了我们传统文化的巨大潜力与我国艺术创新力，不可多得!
34	5d492c89b701b97668b50c54	太原	小王小王 混世魔王	1076105425	1211270	2	电影好看 哪吒与父亲间的亲情感人之深 剧情超赞 情理之中 意料之外
35	5d492c89b701b97668b50c55	乌海	绅士的绝对领域	1076105424	1211270	1	真的很不错，我为国漫打call
36	5d492c89b701b97668b50c56	上海	遇见＆Yee888	1076105422	1211270	暂无	我觉得还行，没那么特别好看。
29531	5d4945bde3ed5945084cc3b1	铜川	Nzs341444635	1074077750	1211270	暂无	整体感觉还可以
29532	5d4945bde3ed5945084cc3b2	西安	℃	1074105064	1211270	2	一定的吐槽一下华东影院—太糟糕了，影片明明写着有彩蛋的字幕，还是将影院所有灯打开，要不是有观影观众提出，还准备清场，这样真的太糟糕，糟糕透了……
29533	5d4945bde3ed5945084cc3b3	崇州	清醒	1074103519	1211270	1	支持国漫特别是这种看着让人心涌澎湃得
29534	5d4945bde3ed5945084cc3b4	乌鲁木齐	倾窈	1074105060	1211270	暂无	国漫的崛起，很精良的制作，我终于在看了无数部与哪吒有关的影片，小说后，看到了李靖对哪吒那深沉的父爱，好片子，我命由我不由天♡
29535	5d4945bde3ed5945084cc3b5	北京	Vtk219474323	1074105059	1211270	暂无	比大圣、大鱼海棠、白蛇缘起都高了一个档次，各方面都是!
29536	5d4945bde3ed5945084cc3b6	哈尔滨	pmE18346156295	1074105052	1211270	1	笑点燃怎泪点
29537	5d4945bde3ed5945084cc3b7	哈尔滨	osL116616434	1074023434	1211270	暂无	非常好，非常好
29538	5d4945bde3ed5945084cc3b9	永州	玄机娘娘	1074103511	1211270	暂无	二刷！我终于说服我的室友来看哪吒啦！好笑又好哭！！！
29539	5d4945bde3ed5945084cc3ba	青冈县	一把雨伞打天晴	1074103510	1211270	2	一念成魔，，，，
29540	5d4945bde3ed5945084cc3bb	佛山	佩	1074102403	1211270	2	故事，情节，图像，特效都特别棒！国产电影真的可以和好莱坞相媲美了，真的超级好看！！给打100分！！
29541	5d4945bde3ed5945084cc3bc	韶关	喜喜快跑	1074103505	1211270	2	电影节奏 把握很好 人物形象 鲜明 打斗看的 很爽 故事很棒 看完只想 表白整个 团队!
29542	5d4945bde3ed5945084cc3bd	信阳	@深呼吸@	1074076862	1211270	2	很不错的电影!
29543	5d4945bde3ed5945084cc3be	南充	用什么暖你一千岁	1074077747	1211270	1	国漫，很好看，很搞笑。加油强烈安利大家观看
29544	5d4945bde3ed5945084cc3bf	昆山	素手抚鸾筝	1074103497	1211270	2	有笑有泪，剧情还可以再丰富一些。个别细节处理的特别棒，看的出来是花了心思的。
29545	5d4945bde3ed5945084cc3c0	武汉	啷个哩个啷	1074103499	1211270	1	H好看美滋滋有道理
29546	5d4945bde3ed5945084cc3c1	济宁	柚寂	1074100879	1211270	2	哪吒简直 太帅了 傲娇小天 使 浓浓亲情 藕饼我磕 了 国漫崛起 之路 nice!
29547	5d4945bde3ed5945084cc3c2	靖江	空城	1074076860	1211270	1	超级好看，特别燃，也有一些搞笑，很喜欢，期待2的上映!
29548	5d4945bde3ed5945084cc3c3	吉林	小鱼儿	1074105033	1211270	2	整个影片特别棒！看完的感觉就是从酸楚到落泪到感动，到温暖！我命由我不由天！命运不是要顺其自然的，而是争取来的，这就是哪吒!
29549	5d4945bde3ed5945084cc3c4	南宁	sXo191839524	1074105032	1211270	暂无	超级超级好看，最后准备到结局眼泪汪汪掉下来
29550	5d4945bde3ed5945084cc3c5	广元	颜成3777	1074104189	1211270	暂无	真好看！！！！！！！！！
29551	5d4945bde3ed5945084cc3c6	兰州	smE110926216	1074104188	1211270	暂无	特别好看，国产动画就是棒!

基于粉丝分布分析

由下图可见，粉丝人数主要集中在沿海一带。



从上图可以看出，《哪吒之魔童降世》的观影人群主要集中在沿海一带，这些地方经济相对发达，城市人口基数庞大，极多的荧幕数量和座位、极高密度的排片场次，让观众便捷观影，活跃的观众评论也多，自然也就成为票房的主要贡献者。



粉丝来源排名前20的城市依次为：**北京、深圳、成都、上海、广州、重庆、长沙、东莞、天津、武汉、西安、郑州、沈阳、佛山、长春、南京、昆明、哈尔滨、青岛、杭州**

电影消费是城市消费的一部分，从某种角度来看，可以作为考察一个城市购买力的指标。这些城市在近年的GDP排行中大都居上游，消费水平较高。

基于评论情感分析

> 以下情感分析参考博客https://www.csuldw.com/2019/09/28/2019-09-28-comment-sentiment-analysis/

数据集处理

原始的电影评论数据共3万多条，数据里面还夹着一些没有评分的数据，将这些评分为NaN的过滤掉之后，剩下来的数据还有3万多条。

除了数据评分取值的问题，CONTENT里的文本有的还是繁体，所以在进行情感分析之前，我们还需要对文本的格式统一起来。

情感类别定义

将1-2颗星定义为：negative

将3颗星定义为：normal

将4-5颗星定义为：positive

数据预处理

对于中文文本分词，这里采用的jieba，同时文本数据进行了去停留词处理，代码如下：

```
import jieba
from sklearn.feature_extraction.text import CountVectorizer
import re


def get_stopwords():
    stopwords = [line.strip() for line in
open('stopword_normal.txt',encoding='UTF-8').readlines()]
    return stopwords


stopwords = get_stopwords()
def text_process(text):
    text = re.sub("[\s+\.\!\/_,$%^*(+\"\']+|[+——！，。？、~@#￥%……&*（）]",
"",text)
    ltext = jieba.lcut(text)
    res_text = []
    for word in ltext:
        if word not in stopwords:
            res_text.append(word)
    return res_text


X = dataset.CONTENT
y = dataset.label
bow_transformer = CountVectorizer(analyzer=text_process).fit(X)
X = bow_transformer.transform(X)
```

模型训练与评估

采用sklearn将数据集划分为训练集和测试集，比例为9:1。

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size=0.1,
random_state=99)
```

由于是多分类，所以就直接采用了MultinomialNB作为baseline，核心代码如下：

```
from sklearn.naive_bayes import MultinomialNB
nb = MultinomialNB()
nb.fit(X_train, y_train)
preds = nb.predict(X_test)
```

模型估计

得到了预测的preds值之后，直接调用sklearn的metrics里面的方法，就可以轻松地将相关的模型评估指标值计算，代码如下：

```
from sklearn.metrics import confusion_matrix, classification_report
print(classification_report(y_test, preds))
```

输出结果

特别喜欢，也特别有意思	positive	0.94
带着两个孩子看的，有笑点，有泪点，很有感触，孩子们也要求再看一次。很好。	positive	0.85
好看感动🐯	positive	0.96
国产精品，意犹未尽	positive	0.82
丑萌丑萌的，看的时候一定要带上纸巾！又哭又笑！措手不及！	negative	0.25
好看 各方面都非常棒 特效音效等等 国产动漫的骄傲！	positive	0.83
动作顺滑，情节流畅，笑点多多又不失感动与坚定，主题一直很明确，还有许多小细节。这	positive	0.83
完美，可以吗？	positive	0.67
买的是碧江广场保利电影院得票，下午18.40的，当时是自动取票机三台都没打印纸了，很久	negative	0.37
一部充满爱的国产动漫电影：慈母的爱[强]——殷夫人严父的爱[强]——李靖师傅的爱[强]——	negative	0.45
感觉前面部分很逗比，正义永远不会迟到。	negative	0.28
值得一看！非常精彩！	positive	0.97
好看！！！！	positive	0.99
这个电影真的很好，不仅幽默，诙谐，看起来有笑点。而且还特别有教育意义，不仅是对小	positive	0.86
真的很不错，目前为止，国产动漫的巅峰了，个人感觉，有剧情，有画面，把一个大家都了	positive	0.81
你值得拥有的	positive	0.98
支持国漫，一句他是我儿，父爱如山	positive	0.98
很好看，我看到后面差点哭了	positive	0.52
超级好看的！	positive	1
非常好看，👍	positive	1
非常值得一看 推荐	positive	0.98
童年的回忆，很完美，代入感很强，值得推荐	positive	0.98
好看，超级，喜欢。	positive	0.99
影片既感人又搞笑，只是母亲的配音跟影片不太协调，其他人物可以负责搞笑，母亲只负责	positive	0.71
哈哈(ටωට)hiahiahia	positive	0.6
电影是很好看的！就是银泰的华谊兄弟电影院观影环境太差，影片放映到后期，影院的打扫	positive	0.58
超好看的，强力推荐	positive	0.99
很热血，也很励志，哪吒很可爱啊	positive	0.87
非常好看，感动	positive	0.97
两个字好看！剧情还是比较丰满，泪点低的我，看到哪吒知道爸爸要跟他换命的时候，泪如	positive	0.71
看的第二遍 还是很喜欢	positive	0.98
不错非常好看	positive	1
还可以，孩子们说好看，喜欢~	positive	0.83
不错，现在的动画大人都喜欢看了，多少年前，只有国外的动画有这个效应，我们国家现在	positive	0.89
影片里表达的东西很多：友情，亲情，命运，人性，家庭教育……这些一点点地融入剧情，细	positive	0.64
完美励志片，国漫崛起的第一神作，期待有更多的优秀国漫出现	positive	0.84
挺好的。中国动漫加油	positive	0.99
精彩绝伦。。。。	positive	0.71
真的值得一看	positive	0.96
棒呆👌国漫加油！！！！！！	positive	0.98
不错！不认命，就是哪吒（每个人）的命！	positive	1
龙是中国帝王的象征！影片似乎有些亵渎	negative	0.3

可视化

x轴为评论给出的评分，认为1-2为negative，3为normal，4-5为positive；

y轴为情感分析给出的置信度，大于0为positive，等于0为normal，小于0为negative。

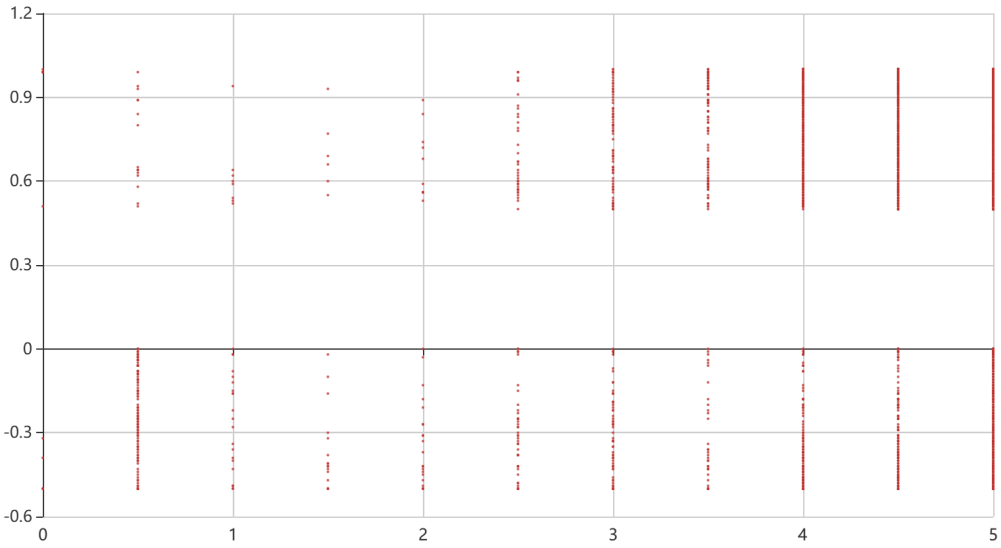
预计当x>3时，数据在x轴上方，x越接近5，y越接近1；

当x<3时，数据在x轴下方，x越接近0，y越接近-1。

但是如下图所示，x轴上方的分布大致符合预期，但是x轴下方似乎在左右两端都有很多negative的分布。

观察情感分析后的数据，发现了可能原因：《哪吒之魔童降世》探讨关于“善恶”的主题，于是把“恶”当作了negative，所以才有了x>3，x轴下方不符合预期的表现。

评分-情感置信度



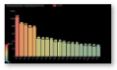
六、总结

以下是整个项目，代码已上传至GitHub：https://github.com/Ianpundar/crawlerDouban

名称	日期	类型	大小	标记
.git	2020/11/3 22:10	文件夹		
.idea	2020/11/1 15:50	文件夹		
__pycache__	2020/11/2 22:58	文件夹		
results	2020/11/5 16:39	文件夹		
2019年度热门电...	2020/11/1 16:44	Markdown File	11 KB	
README.md	2020/11/3 22:10	Markdown File	1 KB	
2019年度热门电...	2020/11/5 23:10	Microsoft Edge ...	3,300 KB	
10box.py	2020/11/4 22:01	Python 源文件	1 KB	
cityGeo.py	2020/11/5 15:51	Python 源文件	4 KB	
cmtsVis.py	2020/11/5 21:40	Python 源文件	2 KB	
date_box.py	2020/11/4 18:25	Python 源文件	5 KB	
date_box-Bron...	2020/11/2 21:43	Python 源文件	3 KB	
Douban.py	2020/11/1 17:40	Python 源文件	10 KB	
Nezha.py	2020/11/4 22:58	Python 源文件	3 KB	
nezhaVis.py	2020/11/5 12:55	Python 源文件	3 KB	
Rate.py	2020/11/4 20:25	Python 源文件	1 KB	
runTime.py	2020/11/4 20:23	Python 源文件	0 KB	
sentiment.py	2020/11/5 17:55	Python 源文件	2 KB	

名称	日期	类型	大小	标记
html	2020/11/5 16:39	文件夹		
png	2020/11/5 16:40	文件夹		
2019movies.xls	2020/11/3 23:25	Microsoft Excel ...	235 KB	
cmtsAnalysis.xl...	2020/11/5 22:19	Microsoft Excel ...	21,024 KB	
nezha.csv	2020/11/5 12:49	Microsoft Excel ...	4,032 KB	

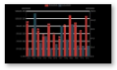

10box.png

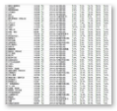

city_bar.png

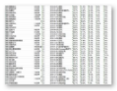

cmtsAnalysis.png


cmtsPre.png

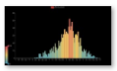

cmtsPre1.png


date_box.png


excel1.png


excel2.png


geo.png


Rate.png


rate_emo.png


res1.png


res2.png


sex_pie.png


结构.png