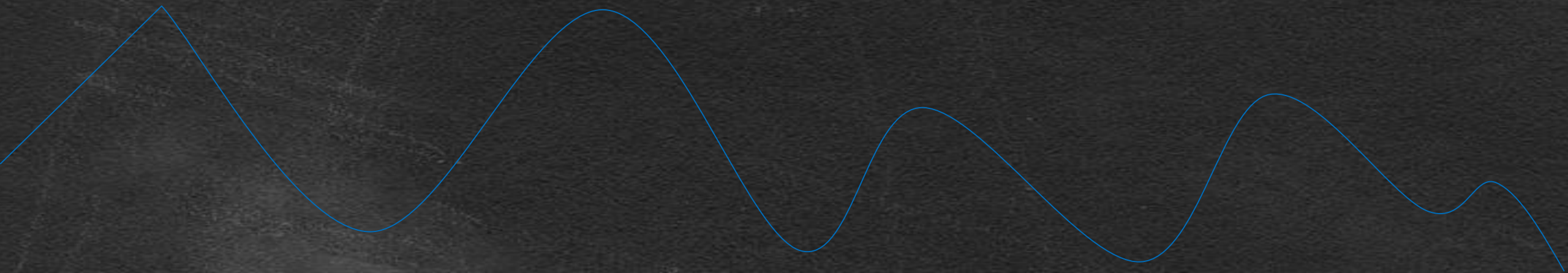
人邮教育
www.ryjiaoyu.com

大数据成就未来

Python爬虫实践《流浪地球》豆瓣影评分析

张敏



目录

背景与挖掘目标

- 豆瓣（douban）是一个社区网站。网站由杨勃（网名“阿北”） 创立于2005年3月6日。该网站以书影音起家，提供关于书籍、电影、音乐等作品的信息，无论描述还是评论都由用户提供（User-generated content，UGC），是Web 2.0网站中具有特色的一个网站。
- 网站还提供书影音推荐、线下同城活动、小组话题交流等多种服务功能，它更像一个集品味系统（读书、电影、音乐）、表达系统（我读、我看、我听）和交流系统（同城、小组、友邻）于一体的创新网络服务，一直致力于帮助都市人群发现生活中有用的事物。

豆瓣 douban

背景与挖掘目标

- 2019年2月5日电影《流浪地球》正式在中国内地上映。根据刘慈欣同名小说改编，影片故事设定在2075年，讲述了太阳即将毁灭，已经不适合人类生存，而面对绝境，人类将开启“流浪地球”计划，试图带着地球一起逃离太阳系，寻找人类新家园的故事。
- 《流浪地球》举行首映的时候，口碑好得出奇，所有去看片的业界大咖都发出了同样赞叹。文化学者戴锦华说：“中国科幻电影元年开启了。”导演徐峥则说，“里程碑式的电影，绝对是世界级别的。”

背景与挖掘目标

- 可是公映之后，《流浪地球》的豆瓣评分却从8.4一路跌到了7.9。影片页面排在第一位的，是一篇一星影评《流浪地球，不及格》。文末有2.8万人点了“有用”，3.6万人点了“没用”。
- 关于《流浪地球》的观影评价，已经变成了一场逐渐失控的舆论混战，如“枪稿”作者灰狼所说，“关于它的舆论，已经演化成‘政治正确、水军横行、自来水灭差评、道德绑架、战狼精神。’”

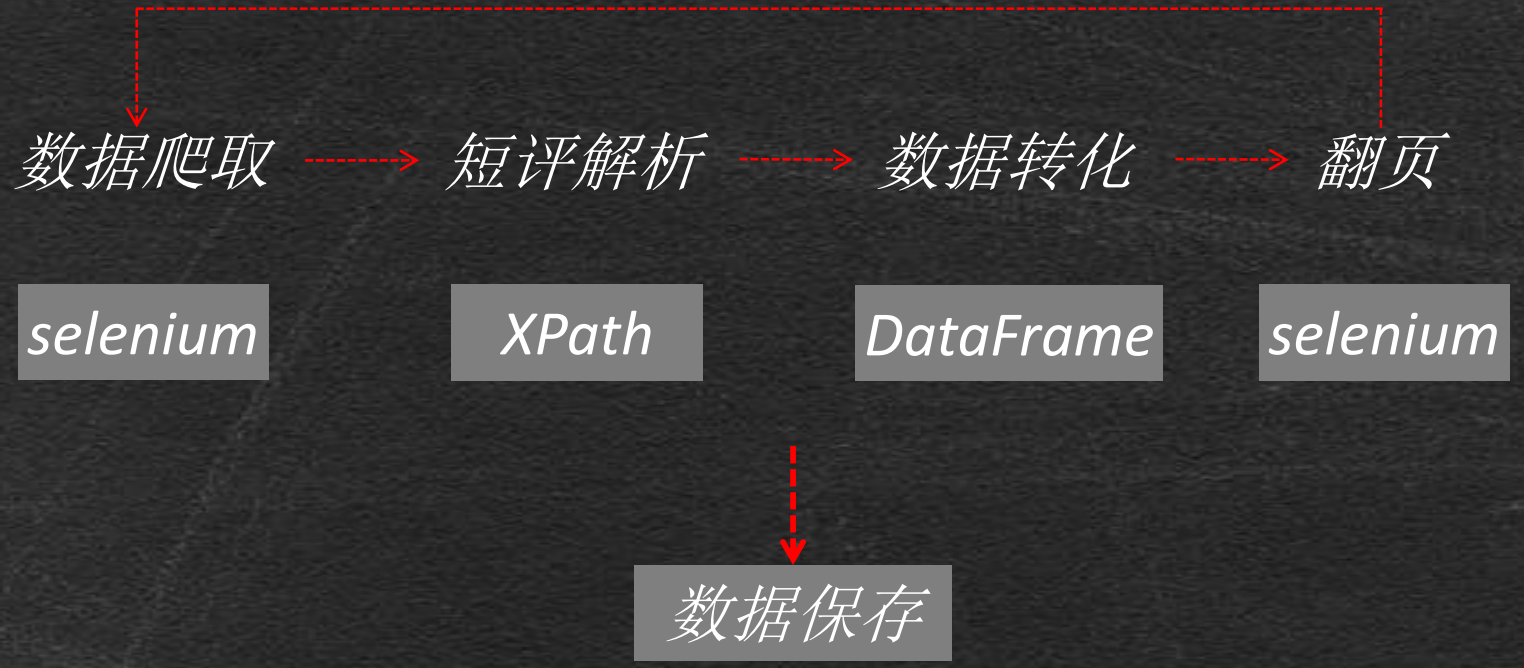
背景与挖掘目标

- 本案例的主要挖掘目标为根据豆瓣对《流浪地球》的短评数据进行文本挖掘及可视化的操作。
- 主要有以下内容：
  1. 获取豆瓣评论数据
  2. 分析好评与差评的关键信息
  3. 分析评论数量及评分与时间的关系
  4. 分析评论者的城市分布情况

目录

获取豆瓣短评数据

主要流程

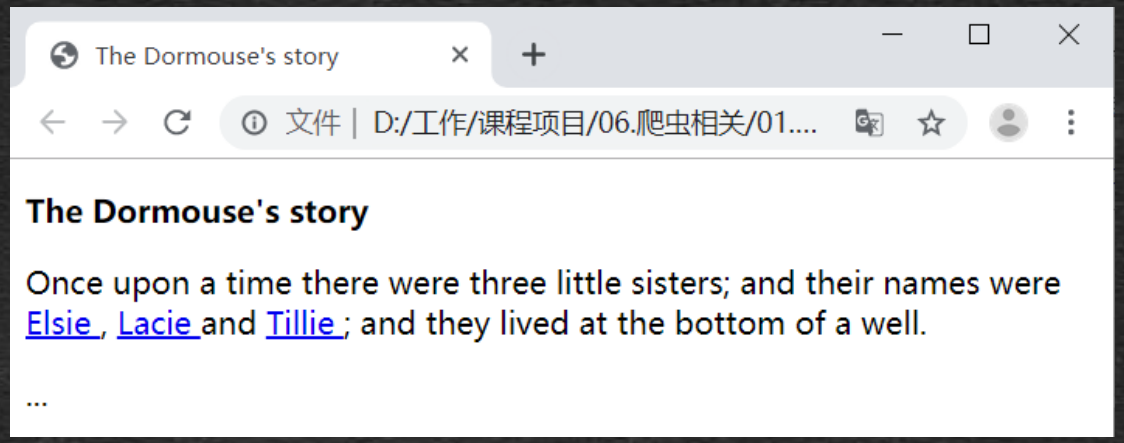


获取豆瓣短评数据

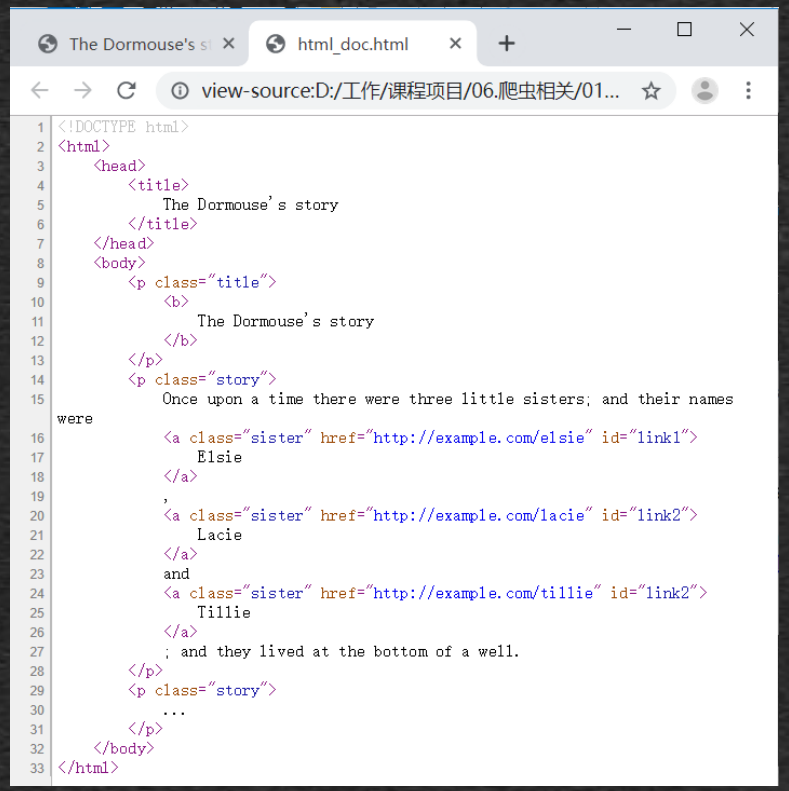
人邮教育 www.ryjiaoyu.com

数据爬取

网页



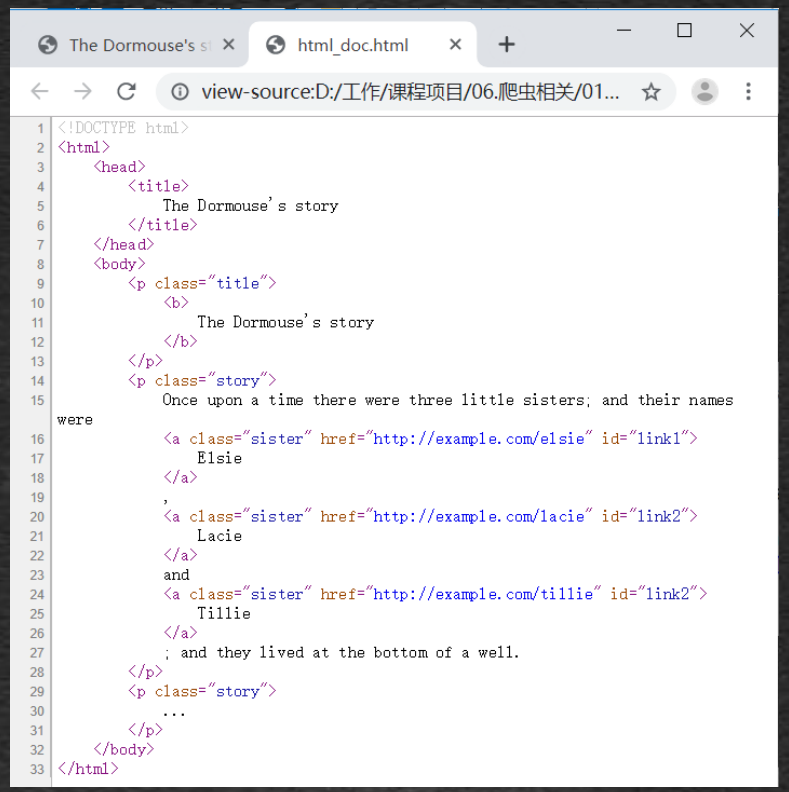
HTML源码



获取豆瓣短评数据

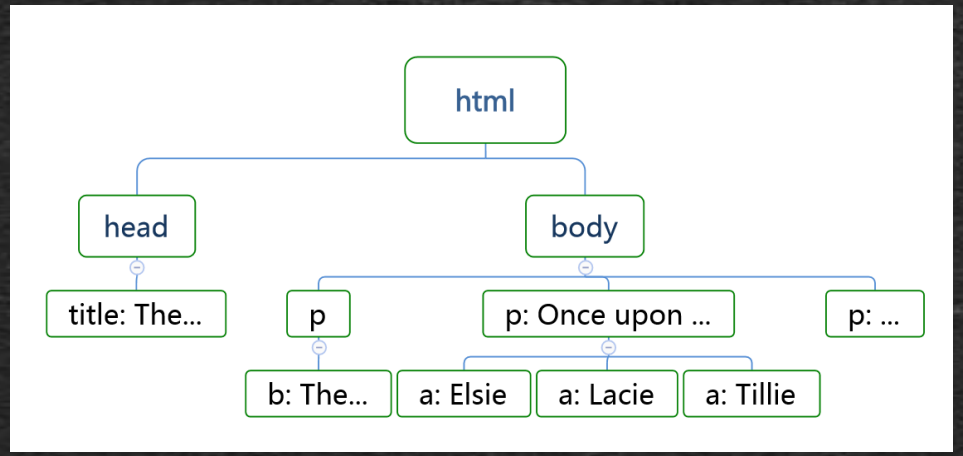
数据爬取

HTML源码




```
<!DOCTYPE html>
<html>
    <head>
        <title>
            The Dormouse's story
        </title>
    </head>
    <body>
        <p class="title">
            <b>
                The Dormouse's story
            </b>
        </p>
        <p class="story">
            Once upon a time there were three little sisters; and their names
were
            <a class="sister" href="http://example.com/elsie" id="link1">
                Elsie
            </a>
            ,
            <a class="sister" href="http://example.com/lacie" id="link2">
                Lacie
            </a>
            and
            <a class="sister" href="http://example.com/tillie" id="link2">
                Tillie
            </a>
            ; and they lived at the bottom of a well.
        </p>
        <p class="story">
            ...
        </p>
    </body>
</html>
```

DOM树



获取豆瓣短评数据

网页源码抓取

- from selenium import webdriver
- import pandas as pd
- url = 'https://movie.douban.com/subject/26266893/comments?status=P'
- drive = webdriver.Chrome()
- drive.get(url)

- Chromedriver下载链接： http://chromedriver.storage.googleapis.com/index.html
- 注意：需将Chromedriver放置在系统环境变量的路径中

1. 基本语法

Xpath使用类似正则的表达式来匹配HTML文件中的内容，常用匹配表达式如下。

表达式	说明
nodename	选取nodename节点的所有子节点
/	从当前节点选取直接子节点
//	从当前节点选取子孙节点
.	选取当前节点
..	选取当前节点的父节点
@	选取属性

2. 谓语

Xpath中的谓语用来查找某个特定的节点或包含某个指定的值的节点，谓语被嵌在路径后的方括号中。

表达式	说明
/html/body/div[1]	选取属于body子节点下的第一个div节点
/html/body/div[last()]	选取属于body子节点下的最后一个div节点
/html/body/div[last()-1]	选取属于body子节点下的倒数第二个div节点
/html/body/div[positon()<3]	选取属于body子节点下的下前两个div节点
/html/body/div[@id]	选取属于body子节点下的带有id属性的div节点
/html/body/div[@id=" content"]	选取属于body子节点下的id属性值为content的div节点
/html /body/div[xx>10.00]	选取属于body子节点下的xx元素值大于10的节点

3. 功能函数

Xpath中还提供功能函数进行模糊搜索，有时对象仅掌握了其部分特征，当需要模糊搜索该类对象时，可使用功能函数来实现，具体函数如下。

功能函数	示例	说明
starts-with	//div[starts-with(@id,”co”)]	选取id值以co开头的div节点
contains	//div[contains(@id,”co”)]	选取id值包含co的div节点
and	//div[contains(@id,”co”)andcontains(@id,”en”)]	选取id值包含co和en的div节点
text()	//li[contains(text(),”first”)]	选取节点文本包含first的div节点

4. 提取header节点下全部标题文本及对应链接

- 使用text方法可以提取某个单独子节点下的文本，若想提取出定位到的子节点及其子孙节点下的全部文本，则需要使用string方法实现。
- 使用HTML类将其初始化通过requests库获取的网页，之后使用谓语定位id值以me开头的ul节点，并使用text方法获取其所有子孙节点a内的文本内容，使用@选取href属性从而实现提取所有子孙节点a内的链接，最后使用string方法直接获取ul节点及其子孙节点中的所有文本内容。

页面等待

- Selenium Webdriver提供两种类型的等待——隐式和显式。显式的等待使网络驱动程序在继续执行之前等待某个条件的发生。隐式的等待使WebDriver在尝试定位一个元素时，在一定的时间内轮询DOM。
- 本节主要介绍显示等待。显式等待是指定某个条件，然后设置最长等待时间。如果在这个时间还没有找到元素，那么便会抛出异常。
- WebDriverWait函数是默认每500毫秒调用一次ExpectedCondition，直到成功返回。ExpectedCondition的成功返回类型是布尔值，对于所有其他ExpectedCondition类型，则返回True或非Null返回值。如果在10秒内不能发现元素返回，就会在抛出TimeoutException异常。
- WebDriverWait的语法使用格式如下。

WebDriverWait(driver, 等待时间)

预期的条件

方法	作用
frame_to_be_available_and_switch_to_it frame	加载并切换
invisibility_of_element_located	元素不可见
element_to_be_clickable	元素可点击
staleness_of	判断一个元素是否仍在DOM，可判断页面是否已经刷新
element_to_be_selected	元素可选择，传元素对象
element_located_to_be_selected	元素可选择，传入定位元组
element_selection_state_to_be	传入元素对象以及状态，相等返回True，否则返回False
element_located_selection_state_to_be	传入定位元组以及状态，相等返回True，否则返回False
alert_is_present	是否出现Alert

获取豆瓣短评数据

人邮教育 www.ryjiaoyu.com

数据展示

1	citys	content	evaluate	labs	nams	scores	times	user_info	votes
2	['北京']	一个悲伤的	推荐	看过	沙雕电影	['40']	#######	['1855738	35161
3	['北京']	电影比预其	推荐	看过	影志	['40']	#######	['tjz230 ', '	68629
4	['北京']	还能更土更	很差	看过	嘟嘟熊之父	['10']	#######	['duduxion	69686
5	['北京']	1.终于，轮	力荐	看过	陆支羽	['50']	#######	['luzhiyu ',	59980
6	[]	真为吴京的	很差	看过	侠侠	['10']	#######	[]	38488
7	['上海']	三星鼓励一	还行	看过	艾晨	['30']	#######	['satan163	18016
8	['重庆']	失望 一群	较差	看过	我是王大朋	['20']	#######	['cheer.o ',	13473
9	[]	野心远远大	很差	看过	妖孽	['10']	#######	[]	13873
10	[]	求求编剧，	很差	看过	迷眼看青山	['10']	#######	[]	12590
11	[]	台词做作居	较差	看过	gus	['20']	#######	[]	12540

获取豆瓣短评数据

注意事项

- 豆瓣封IP，白天一分钟可以访问40次，晚上一分钟可以访问60次，超过限制次数就会封IP。
- 在登录账号的情况下，豆瓣也只提供500条展示的数据。

目录

分析好评与差评的关键信息

预处理

	citys	content	evaluate	labs	nams	scores	times	user_info	votes
0	['北京']	一个悲伤的故事：太阳都要毁灭，地球都要流浪了，我国的校服还是这么丑.....	推荐	看过	沙雕电影	['40']	2019-02-05 00:24:35	['185573840 ', ' 2018-10-07加入']	35161
1	['北京']	电影比预期要更恢弘磅礴，晨昏线过后的永夜、火种计划、让地球流浪、木星推动地球…等等大小设定，...	推荐	看过	影志	['40']	2019-02-04 15:56:16	['tjz230 ', ' 2005-07-18加入']	68629
2	['北京']	还能更土更儿戏一点吗？毫无思考仅靠煽动，毫无敬畏仅余妄想。好的科幻片应该首先承认人类的无知，...	很差	看过	嘟嘟熊之父？	['10']	2019-01-28 22:06:27	['duduxiongzhifu ', ' 2008-01-28 加入']	69686
3	['北京']	1.终于，轮到我们仰望星空。2.后启示录死亡废墟，赛博朋克地下城，以及烟波浩渺的末日想象，缔...	力荐	看过	陆支羽	['50']	2019-01-29 20:10:48	['luzhiyu ', ' 2008-08-30加入']	59980
4	[]	真为吴京的演技尴尬，总是摆出一副大义凌然的样子，好奇为什么刘的作品中总有这种傻逼般的圣母存在...	很差	看过	侠侠	['10']	2019-02-05 01:55:20	[]	38488

分析好评与差评的关键信息

人邮教育 www.ryjiaoyu.com

预处理

	citys	content	evaluate	labs	nams	scores	times	user_info	votes	user_age
0	北京	一个悲伤的故事：太阳都要毁灭，地球都要流浪了，我国的校服还是这么丑.....	推荐	看过	沙雕电影	40.0	2019-02-05 00:24:35	['185573840 ', ' 2018-10-07加入']	35161	2018-10-07
1	北京	电影比预期要更恢弘磅礴，晨昏线过后的永夜、火种计划、让地球流浪、木星推动地球…等等大小设定，...	推荐	看过	影志	40.0	2019-02-04 15:56:16	['tjz230 ', ' 2005-07-18加入']	68629	2005-07-18
2	北京	还能更土更儿戏一点吗？毫无思考仅靠煽动，毫无敬畏仅余妄想。好的科幻片应该首先承认人类的无知，...	很差	看过	嘟嘟熊之父？	10.0	2019-01-28 22:06:27	['duduxiongzhifu ', ' 2008-01-28加入']	69686	2008-01-28
3	北京	1.终于，轮到我们仰望星空。2.后启示录死亡废墟，赛博朋克地下城，以及烟波浩渺的末日想象，缔...	力荐	看过	陆支羽	50.0	2019-01-29 20:10:48	['luzhiyu ', ' 2008-08-30加入']	59980	2008-08-30
4	None	真为吴京的演技尴尬，总是摆出一副大义凌然的样子，好奇为什么刘的作品中总有这种傻逼般的圣母存在...	很差	看过	侠侠	10.0	2019-02-05 01:55:20	[]	38488	None

分析好评与差评的关键信息

总体评论的词云展示

可以看到高频词“中国”、“地球”、“人类”表现出该片的主要人文思想，“特效”体现出特效镜头对科幻片的重要性，“科幻电影”体现出影迷对科幻类电影的浓厚兴趣。

分析好评与差评的关键信息

好评与差评的词云展示

好评词云

差评词云

目录

分析评论数量及评分与时间的关系

评分统计

可以看到用户对《流浪地球》的评价两极分化，但大多数倾向3星到4星。

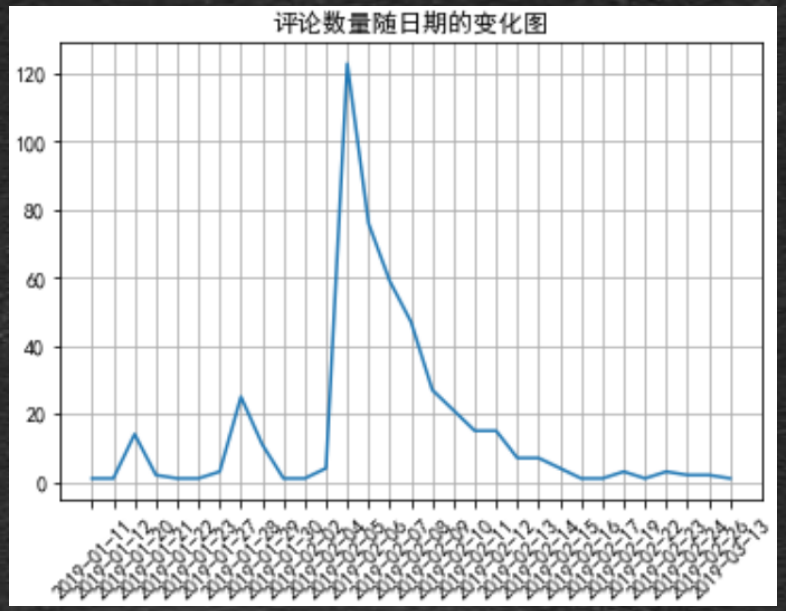
评分	数量
★☆☆☆☆	49
★★☆☆☆	77
★★★☆☆	124
★★★★☆	117
★★★★★	98

分析评论数量及评分与时间的关系

评论数量随日期的变化

- 点映时间：2019-1-20 、2019-1-28
- 正式上映时间：2019-2-5

- 总结：点映后会有小幅的评论数量增加，正式上映后，评论数据大幅上涨，达到了高峰。正式上映后，每日发布的评论数量逐渐减少。



分析评论数量及评分与时间的关系

评论数量随日期的变化

- 在影片上映3天内为评论高峰，这符合常识，但是也可能有偏差，因为爬虫获取的数据是经过豆瓣电影排序的，倘若数据量足够大得出的趋势可能更接近真实情况。
- 影片在上映前也有部分评论，分析可能是影院公映前的小规模试映，且这些提前批的用户的评分均值，差不多接近影评上映后的大规模评论的最终评分 ，从这些细节中，我们或许可以猜测，这些能提前观看影片的，可能是资深影迷或者影视从业人员，他们的评论有着十分不错的参考价值。

分析评论数量及评分与时间的关系

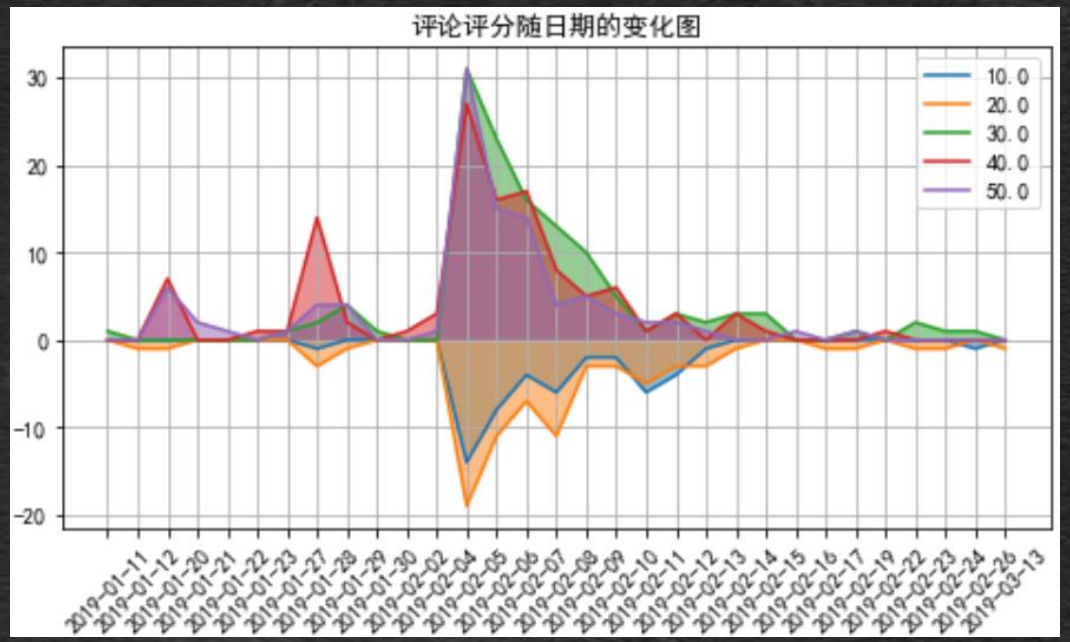
评论数量随时刻的变化

- 豆瓣用户发布短评的时间主要集中在晚上，17点至凌晨0点比例尤为明显。，随着时间向深夜推进，比例逐渐下降，凌晨4点达到最低值。这主要与用户的作息生活有关系。
- 同时短评一般在在观看完电影后发布的，所以用户可能偏向于观影结束回到家之后再进行对影片的评价行为。

分析评论数量及评分与时间的关系

豆瓣评分的时间趋势分析

1. 在点映期间，对电影的评价大部分是正面评价，但是电影上映后用户对《流浪地球》的评价开始两极分化。
2. 一星评价中，2019-2-11有个小高峰，而当天是星期一，好评的数量是小低谷，可能是刷负分的评价。

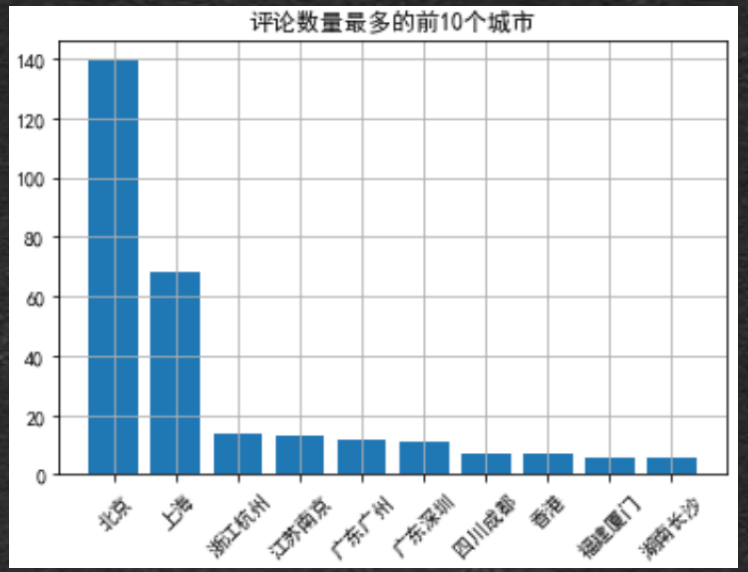


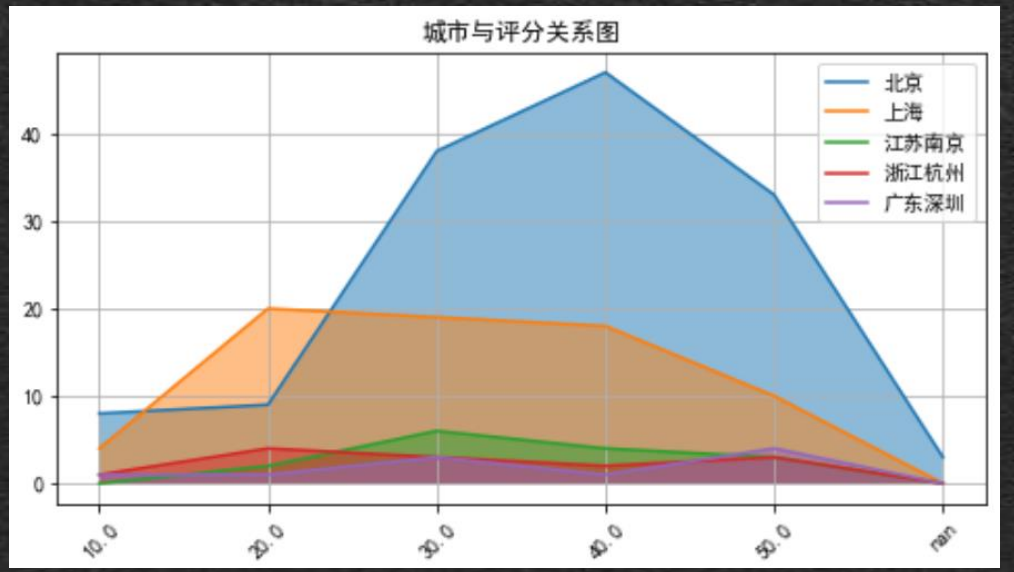
目录

分析评论者的城市分布情况

评论数量最多的前10个城市排名

总结：北京、上海的用户是最多的，同时对两个城市的数据统计时发现，北京的用户倾向于给该类型主旋律影片四星的评价，而上海地区打差评的评价更多一些。

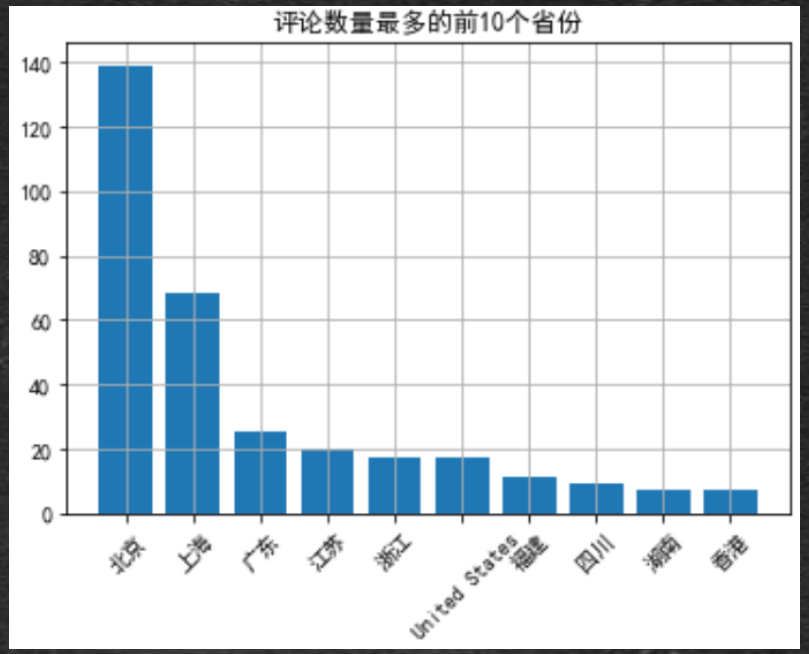




分析评论者的城市分布情况

评论数量最多的前10个省份排名

总结：北上广使用豆瓣进行评价的记录更多一下，可能是豆瓣的人文、企业文化受众多为一线城市的民众或受教育水平影响。



总结

- 本案例的主要挖掘目标为根据豆瓣对《流浪地球》的短评数据进行文本挖掘及可视化的操作。
- 从好评与差评的关键信息展示上可以看得出该影片是中国难得的科幻类型的影片，讲述了人类带着地球流浪的事情，好评主要因为特效和爱国，差评主要因为剧情生硬。
- 从日期上面去进行评论数量分布统计发现评论数量最多的在上映后一周内。点映时评论较好，但是上映后口碑两极分化。
- 北京上海的用户发表短评最多，常住北京的用户好评居多，上海的用户倾向给差评。

Thank you!