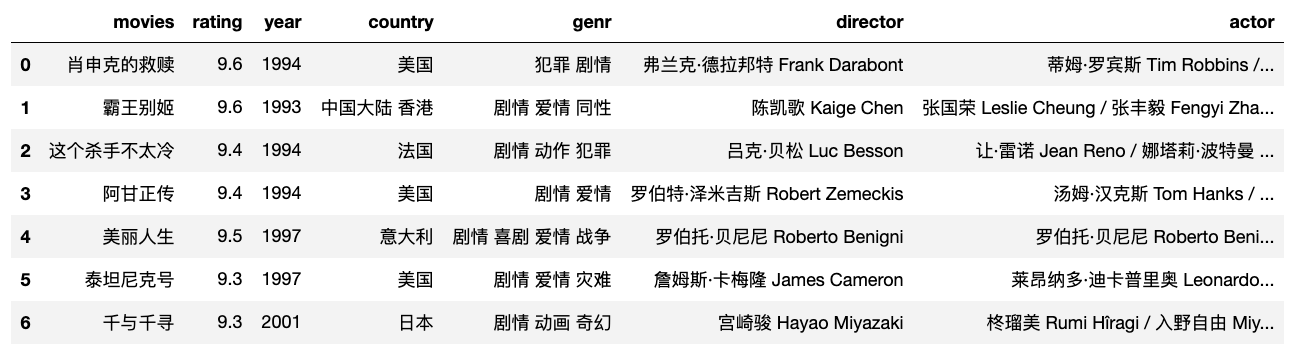
豆瓣电影Top250 python爬取及数据分析

最近学数据分析之余，在用豆瓣找电影调节心情， [Top250榜单](https://movie.douban.com/top250)的数据样本不多，也没有设置反爬虫机制，很适合爬虫入门练手和数据清洗练手。Kaggle之类的网站一般提供已经整理好的表单数据，模型建立完之后可以直接选取套用。但一般场景中数据事先并未提供，需要自己获取数据，本次试着用python爬取了豆瓣电影Top250的数据，体验了一下从网站抓取并探索的过程。

1、用爬虫获取数据

爬虫的机制是先获取页面html数据，然后进行解析，加工成适合的格式后进行储存。这里通过requests库获取页面，通过BeautifulSoup库解析页面。因为没有研究过正则表达式，所以只能先把所有数据先粗略抓取下来以后，通过pandas进行了数据的清洗处理。最终得到Top250里的电影名、豆瓣评分、发行时间、拍摄国家、影片类型、导演及演员信息。

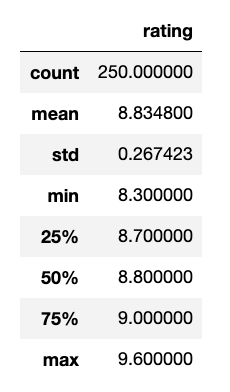


代码点击这里：

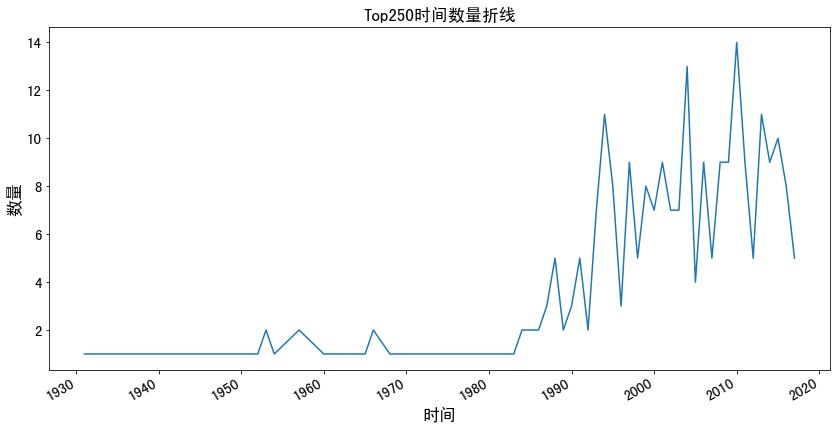
<https://github.com/gameboy4264/Kaggle-Comp/blob/master/crawler/douban3.py>

2、数据分析

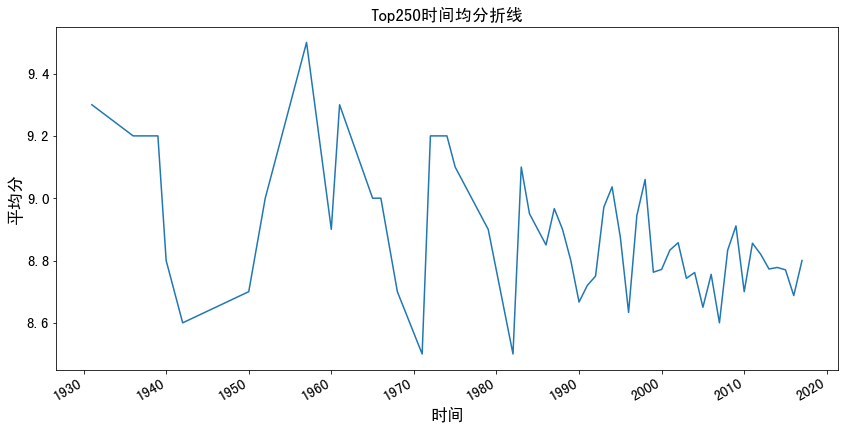
数据获取完之后，先要规整数据，大部分规整工作已经在爬虫程序里处理完了，这里发现两个问题，一个是发行时间，有重映时间的需要单独找出来替换为首次发行时间；二是多国家参与拍摄的片子按国家分析时需要拆开，影片种类也有同样的问题，所以源数据集复制了几分之后，演变成了针对各个单变量分组的子数据集。此榜单250部影片的平均评分为8.83分，标准差0.267， 最低8.3分，最高9.6分。



（1）发行时间

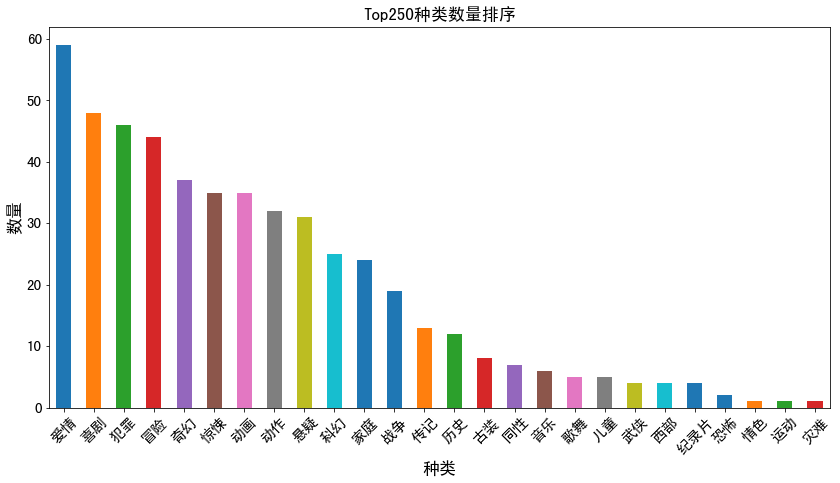


大部分影片为1990年以后发行，只有少量发行与1990年之前，如果不太能接受老旧风格的朋友，这个榜单是一份不错的寻片参考。

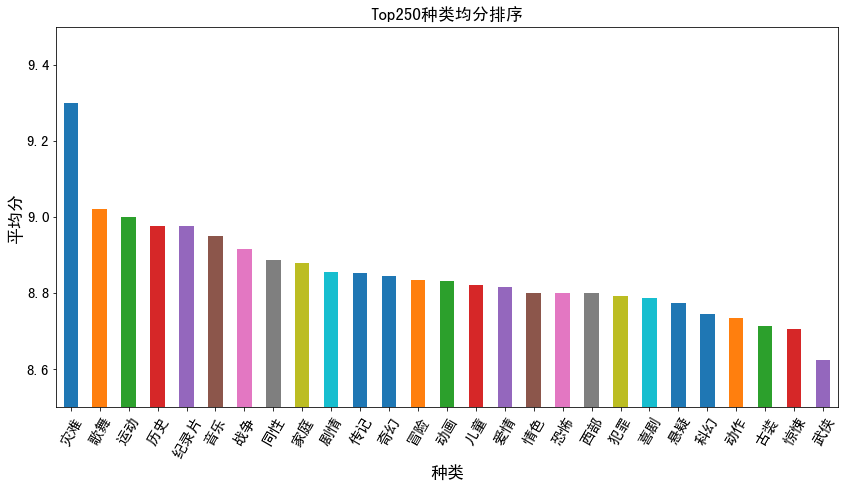


平均分沿着时间轴趋于稳定，在8.8分上下徘徊。

（2）影片种类

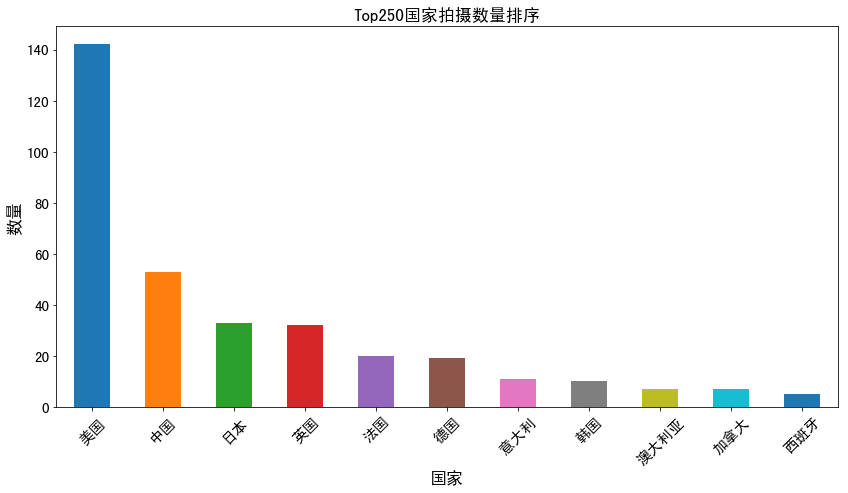


大部分电影种类标签不唯一，可以同时数据很多种类，比如“剧情、动作、犯罪”，按种类分割之后，发现其中188部都有剧情标签，想想也是，哪部电影能没有剧情呢╮(￣.￣)╭　。去除剧情标签后，发现数量最多的是爱情片，其次分别是喜剧、犯罪、冒险。

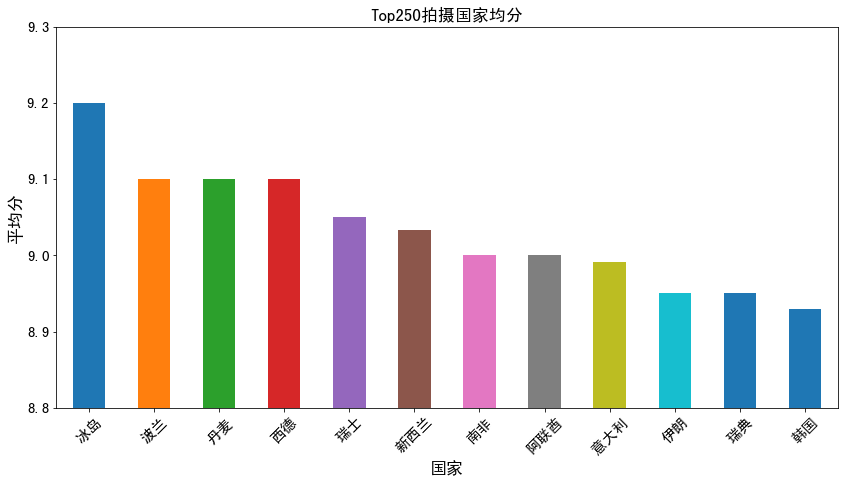


根据种类算出榜单影片的平均分，灾难片评分最高，其次为歌舞、运动、历史类，对照上图发现平均分比较高的种类都是数量较少的类型。经过查询发现250榜单中灾难片只有一部《泰坦尼克号》。

（3）拍摄国家

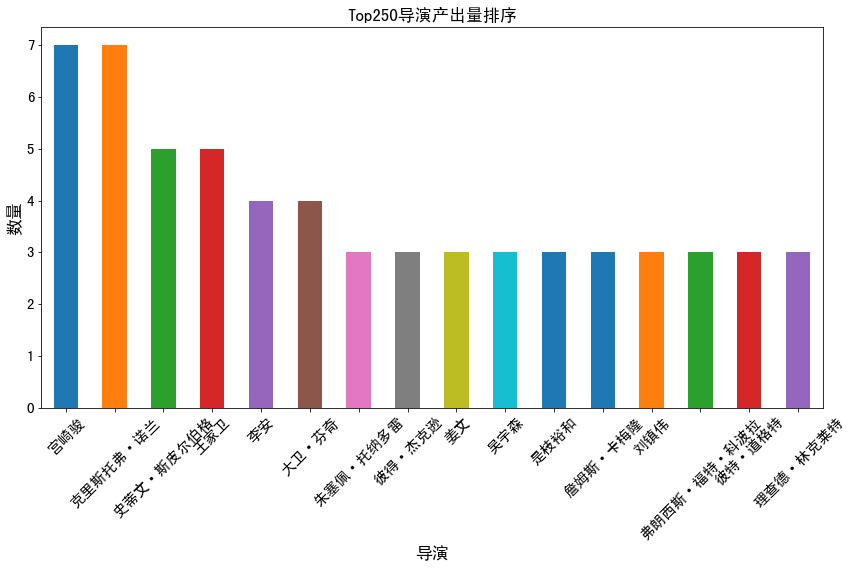


榜单中美国主导或者参与拍摄的影片占多数，为142部，随后为中国、日本、英国。



按平均分排列的话，小众影片的国家则排在前列，最高是冰岛，最后是波兰、丹麦、西德。

（4）导演产量



上图为有超过三部影片入榜的导演名单。最高产的是宫崎骏和诺兰，分别有7部影片入榜，斯皮尔伯格、王家卫、李安、大卫芬奇均有4部以上影片入榜。

3、遇到的问题

（1）数据解析处理问题

我的目的是要得到一个dataframe数据，所以先建了一个空的数据框，然后抓到各列的数据往里填。填完之后再通过pandas的apply用lambda函数进行提取。

（2）matplotlib默认不支持中文

Matplotlib默认不支持中文，开始生成的图片所有中文只显示一个个方框。需要下载一个中文字体，然后放到Matplotlib的字体库中。这样可以一劳永逸，永久解决中文画图事宜。

（3）一列数据有多种元素

很多电影有多国参与拍摄，同时又属于多个影片种类，分析的时候需要分割数据。这里针对每一列的分割生成了一个子数据及，这样单因素分析或者画图的时候其他列不受影响。

（4）时间序列需要单独处理

抓取完的数据为字符串类型，针对时间分析之前要通过to\_datetime函数转化为时间戳格式。

（5）数据类型多为字符串类型而非数值类型

此榜单数据除了得分列之外，其他多为字符串类型，所以只能根据数量或者平均分进行单一变量分析，无法进行散点图、热点图之类的多因素分析。