**网络爬虫与数据收集大作业**

基于豆瓣前200电影的数据收集与分析

学院：阿里云大数据应用学院

专业：数据科学与大数据技术

姓名：钟耀荣

学号：20200619

指导老师：周雷、李洁

1. 选题背景与意义

根据查询相关资料信息显示：豆瓣已经成为国内电影爱好者、影评人士的聚集地，而豆瓣电影数据分析是电影发展的衡量标准。豆瓣上积攒了大量电影数据为电影行业分析提供了重要资源，不仅可以帮助到电影爱好者挑选自己喜欢的电影，而且可以通过豆瓣电影来衡量国内外的电影的发展情况。因此豆瓣前200电影数据的收集与分析具有一定的背景和意义。豆瓣是中国最具影响力的电影社区之一，拥有着大量的电影信息和用户评价。豆瓣前200电影是豆瓣用户最喜爱的电影之一，它们代表了当前中国电影市场的一定趋势和用户口味。对这些数据进行收集和分析，可以帮助我们更好地了解中国电影市场和观众的需求，为电影产业提供一定的参考和指导。具体来说，数据收集和分析的背景与意义有以下几个方面：

1、了解电影市场现状

通过收集豆瓣前200电影的数据，我们可以了解当前中国电影市场的现状。比如，哪些类型的电影更受欢迎，哪些导演和演员备受认可，哪些电影的制作水平更高等等。这些信息可以帮助电影制片方更好地了解市场需求，也了解到电影制作的趋势，制作更符合观众口味的电影。

2、提高电影创作质量

通过分析豆瓣前200电影的数据，我们可以发现电影制作中的优点和不足。比如，哪些电影的叙事结构更为合理，哪些叙述结构更为符合大众的审美，哪些电影的角色刻画更为细腻，哪些电影的画面效果更为出色，那些电影可以引起大众的好评等等。这些信息可以帮助电影制作人员提高电影的创作质量和观赏体验。

3、探索电影市场趋势

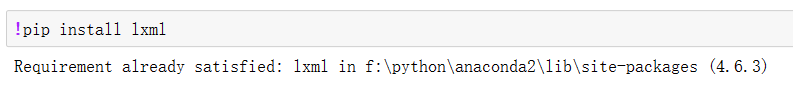
通过分析豆瓣前200电影的数据，我们可以发现一些电影市场的发展趋势。比如，随着观众对文艺片的需求增加，文艺片的市场占比是否在逐渐上升；随着科幻电影的热度不断攀升，科幻电影是否会成为未来电影市场的主流，爱国片是否可以引导青少年构建正确的社会价值观，爱国价值观并引起人们的反思等等。这些信息可以帮助电影从业者更好地了解电影市场的未来发展趋势，为电影产业提供更有针对性的发展方向。

综上所述，豆瓣前200电影数据的收集与分析对于了解电影市场现状、提高电影创作质量和探索电影市场趋势等方面具有重要的背景和意义。

1. 数据收集

这个从豆瓣爬取的数据是我自己写的爬虫代码到豆瓣电影Top榜单上进行爬取的，具有真实性，可用性，并且进行了调试测试了该网页可以进行爬取。

在数据爬取前，我先下载了一个必不可少的lxml库，lxml是使用Python语言编写的库，主要用于解析和提取HTML或者XML格式的数据，它不仅功能非常丰富，而且便于使用，可以利用XPath语法快速地定位特定的元素或节点。因此有了这个库才能更好的解析网址。如图：



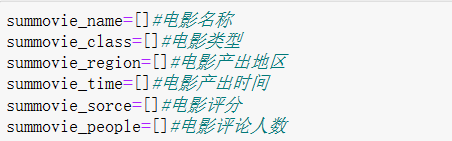
1. 确定我要爬取的网站：在我开始编写爬虫程序之前，我考虑了要获取数据的网站以及该网站的数据结构。并了解如何访问该网站并获取数据。
2. 我首先确定要获取的数据类型：在我开始编写程序之前，我明确了需要爬取的数据类型是csv文本类型。

3、编写程序：确定了要获取的数据类型和要爬取的网站后，我做了一个测试代码，测试了此网站的可连通性，判断出接下来的爬取是否可行，检查网站是否有反爬虫机制。如下图：



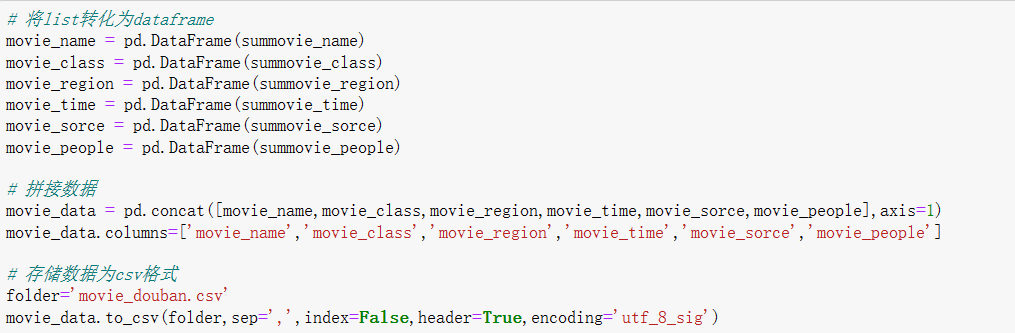
可以看到的是此网站是可以来连通的，我使用的python编程语言和库来获取网站数据。而我写的程序将模拟浏览器行为以访问网站并收集数据。

1. 爬取数据：当成功获取数据后，我先将其转为dataframe模式，存于movie\_data中。如图：





5、存储数据：数据爬取并抓化为dataframe格式后，将爬取的数据转换为csv格式，存储在文件中，为了方便后续数据分析时，可以直接调用到数据进行数据清洗。



1. “豆瓣前200电影”分析

对于豆瓣的前200电影进行分析，首先需要导入所必要的库，运用库已有的功能来作图。而对于这次作业，我将运用到matplotlib与pyecharts来进行作图分析。下面对matplotlib和pyecharts作图的意义进行阐述。

用matplotlib和pyecharts作图的意义在于让数据更加直观、易于理解和分析。Matplotlib是Python的一个绘图库，提供了丰富的绘图工具，可以制作多种类型的图表，如线图、散点图、条形图、饼图等。它的优点在于绘图功能强大，可以满足大部分数据可视化的需求，而且使用简单快捷。

Pyecharts是一个基于Echarts的Python可视化库，提供了一系列丰富的图表组件，可以帮助用户制作不同类型的图表，如折线图、柱状图、饼图、地图等。它的优点在于图表的交互性和美观性，能够制作出具有较高质量的可交互式图表。

综上所述，用matplotlib和pyecharts作图的意义在于让数据更加生动形象地呈现出来，更好地理解和分析数据，同时能够提高数据报告的可视化效果和吸引力。

我对“豆瓣前200电影”分析如下：

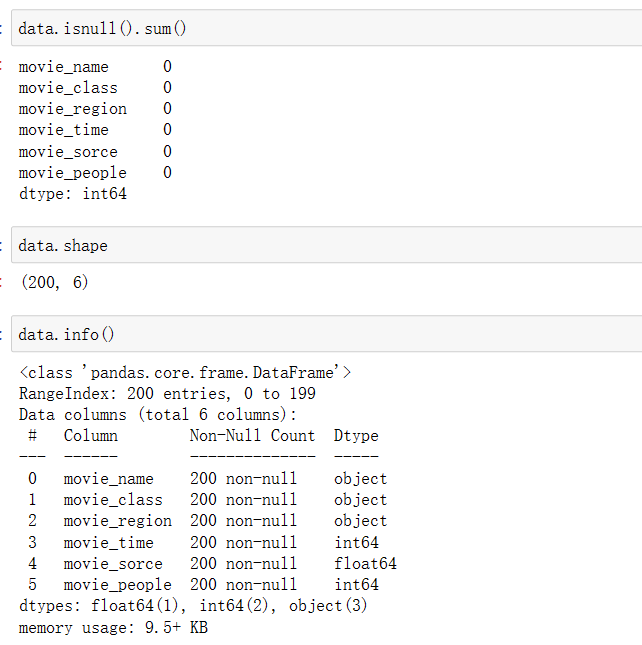
1. 首先对数据进行清洗，检查出数据是否有不正常的地方，是否存在空值、错误值、异常值等等情况，检查完成后，如果有异常情况可以对异常的数据用fillna()函数替换掉不正常的值，达到数据清洗的目的。清洗过后展现出数据的大小，对自己即将要分析的数据有一个初步的认识，并且查看各数据的类型。

数据清洗的好处主要包括以下几点：

(1)、提高数据的质量：数据清洗可以去除数据中的噪声、异常值和错误数据等，从而提高数据的准确性和可靠性。

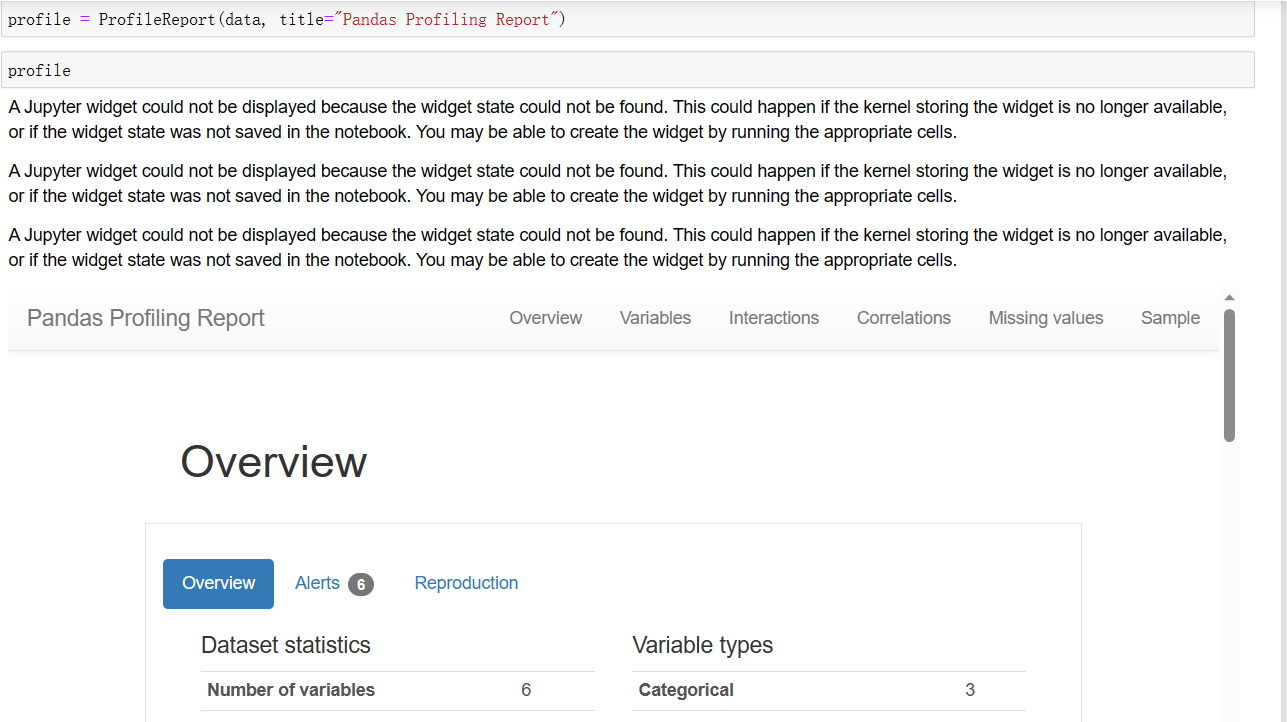
(2)、降低数据分析的难度：清洗后的数据更规范、易于理解和分析，可以减少分析人员的工作量和分析时间。

(3)、提高数据应用的效果：清洗后的数据可以更好地支持各种应用，如机器学习、数据挖掘、商业智能等，从而提高应用效果和价值。

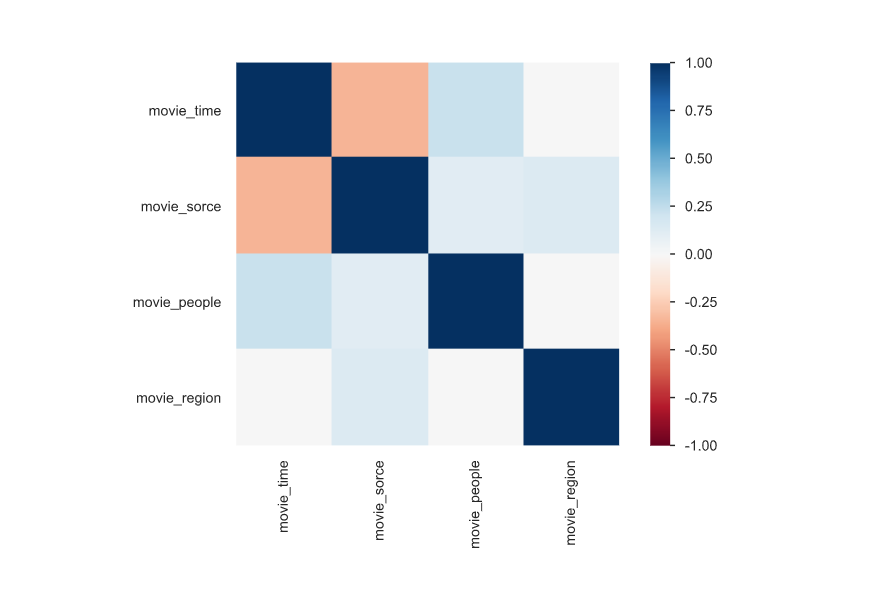


1. 清洗数据完过后开始对数据进行可视化分析，我导入了一个pandas\_profiling的库，它是一个Python库，可以自动生成一个数据报告，该报告包含有关数据集的各种统计信息和可视化。它可以帮助数据科学家更快地了解数据集，识别数据中的问题，并加快数据清理和探索的过程。

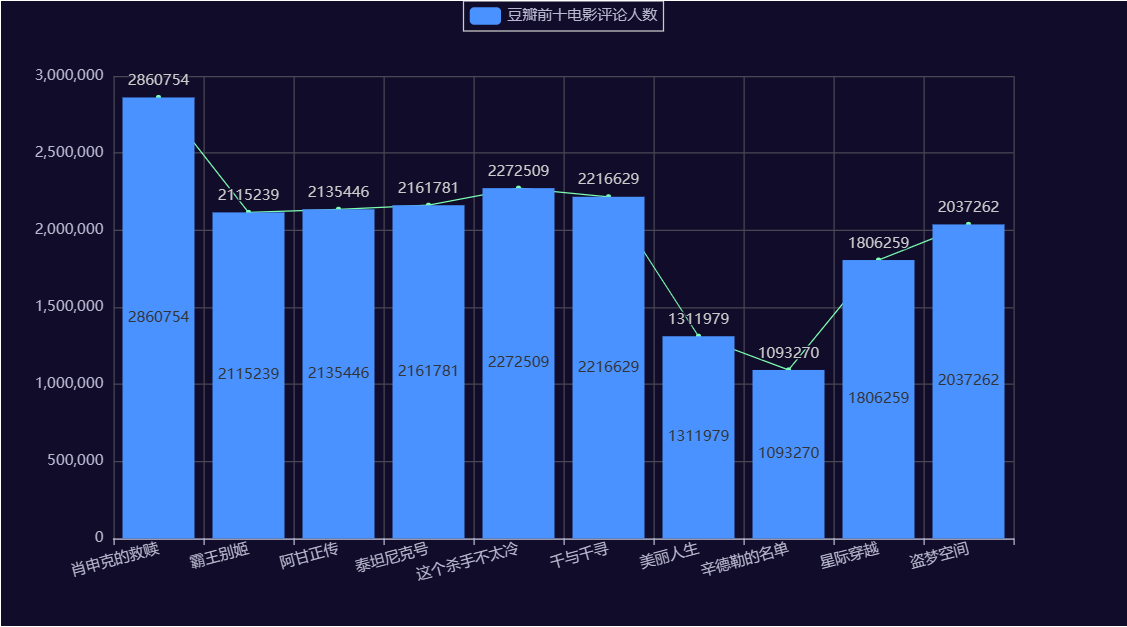
屏幕截图 2023-06-02 005929



1. 热力矩阵图分析各维度之间的关系：

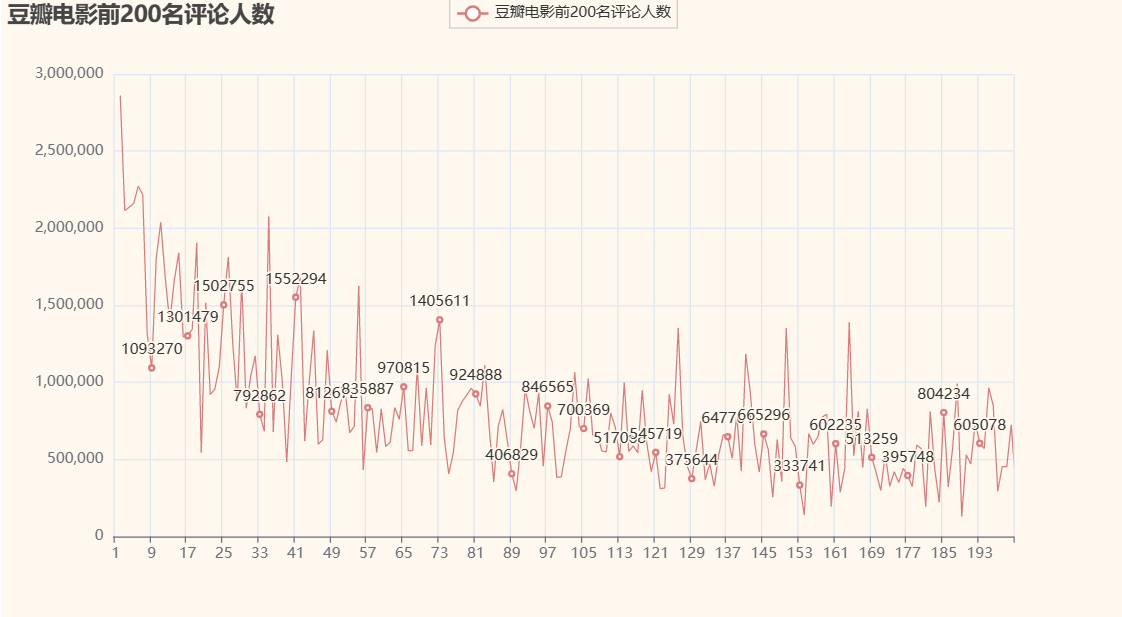


1. 用了pyecharts做了柱形图加折线图的豆瓣前十电影评论人数：



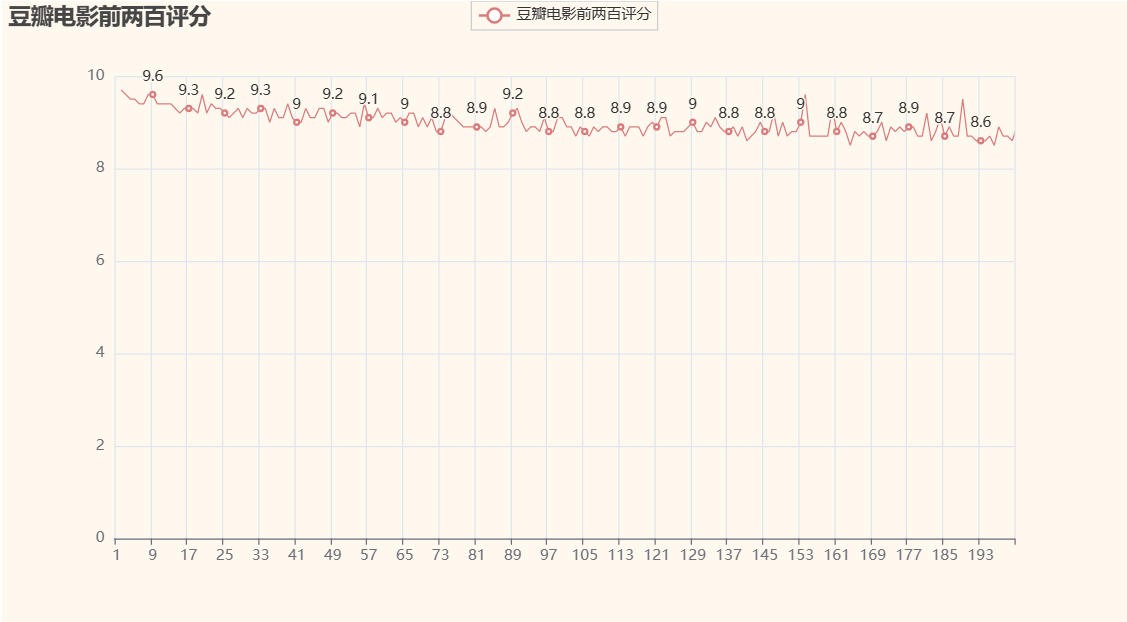
从评论人数可以看出第一名的讨论热度跟其他的电影存在着一定的差距，也说明了大众对于这部电影的话题讨论度挺高的，而其他九个除了第七、第八，都处于较为平均的数值。

1. 用pyecharts做了豆瓣前200电影评论人数的折线图：



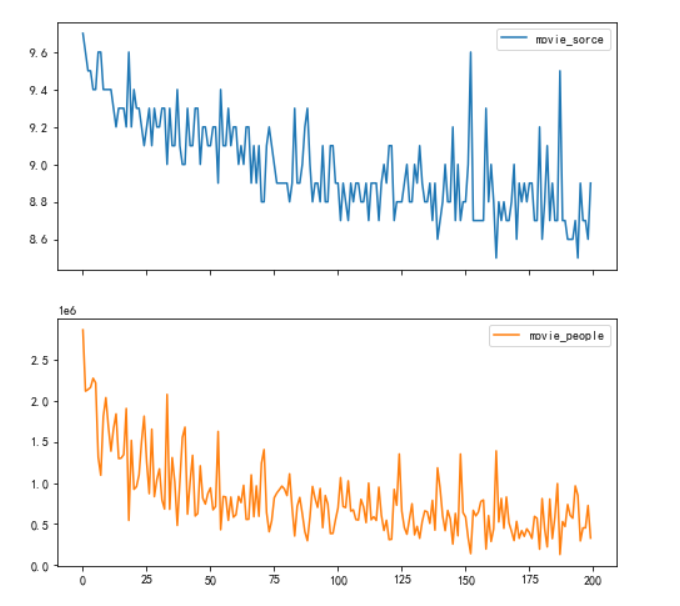
从图不难看出电影的排名与评论人数，话题讨论度成正相关关系，基本是从高到低排列。也说明了排名越靠前的电影，评论人数越多。

1. 用pyecharts做了豆瓣前200电影评论评分的折线图：



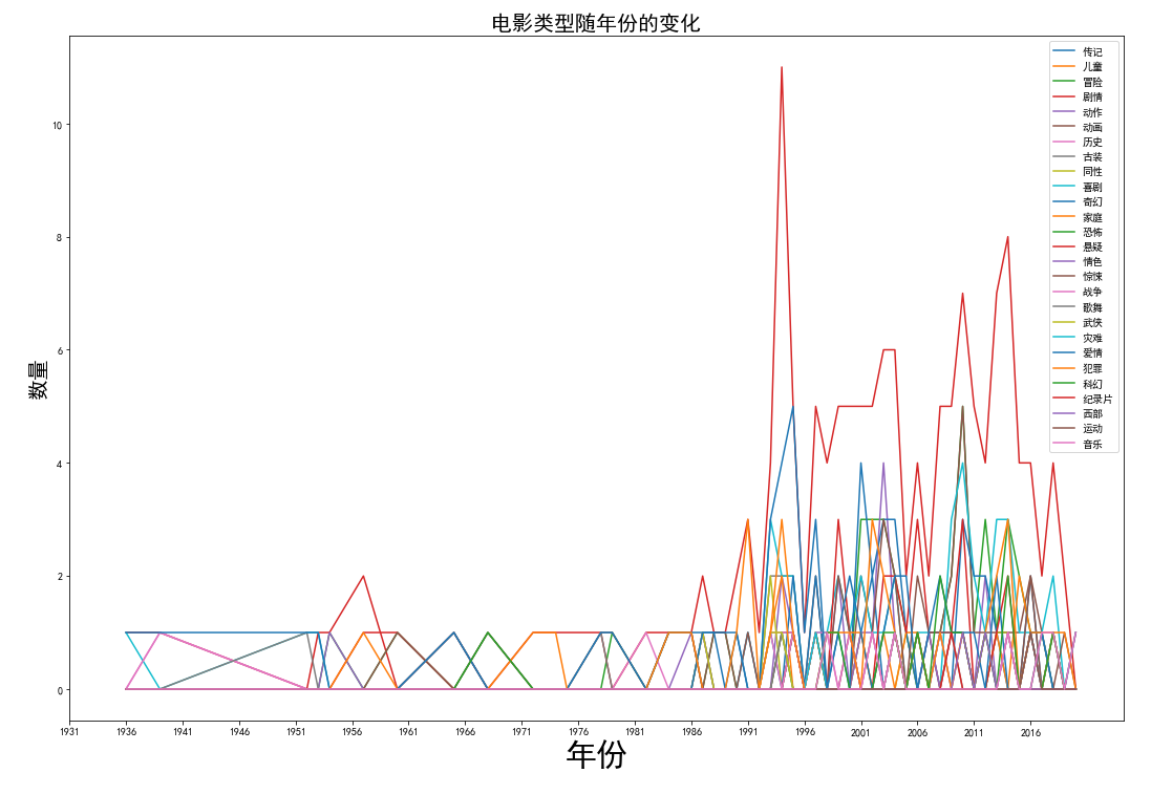
从图里可以看出，前两百名的评分排名整体下降趋势是非常缓慢的，并且没有任何人一部电影低于8分，可以说是大家公认的处于顶端的电影，也是大家共同评出来的。从中也可以看到整体评分是较为正常的与排名没有太大偏颇。

7、这里用了matplotlib可视化方法单独对豆瓣前200的电影做了评分和评论人数的直观折线图：



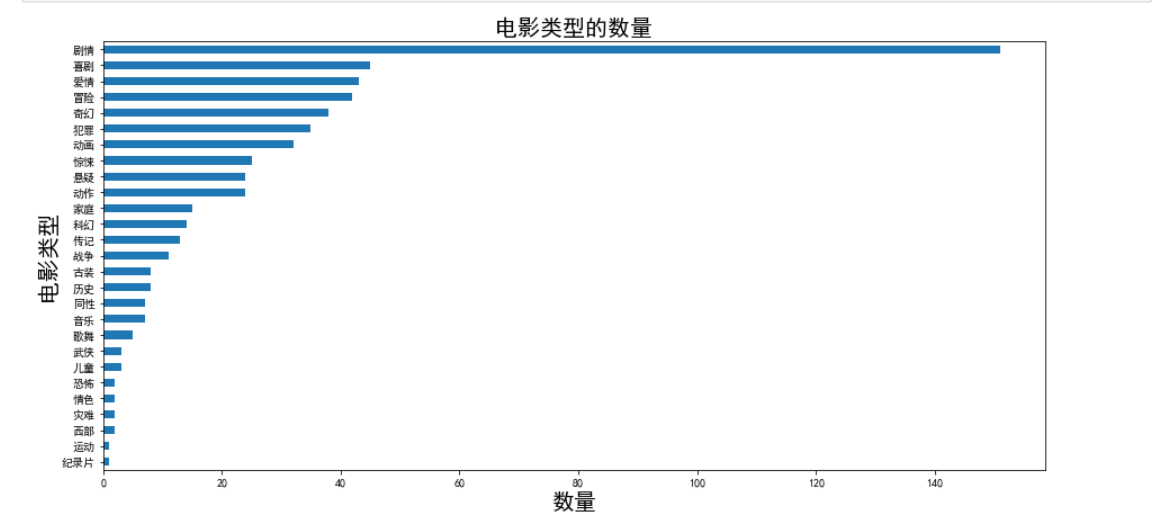
可以看出评分与评论人数的整体趋势是较为一致的。

8、这里用matplotlib做了电影类型随年份的变化的折线图：



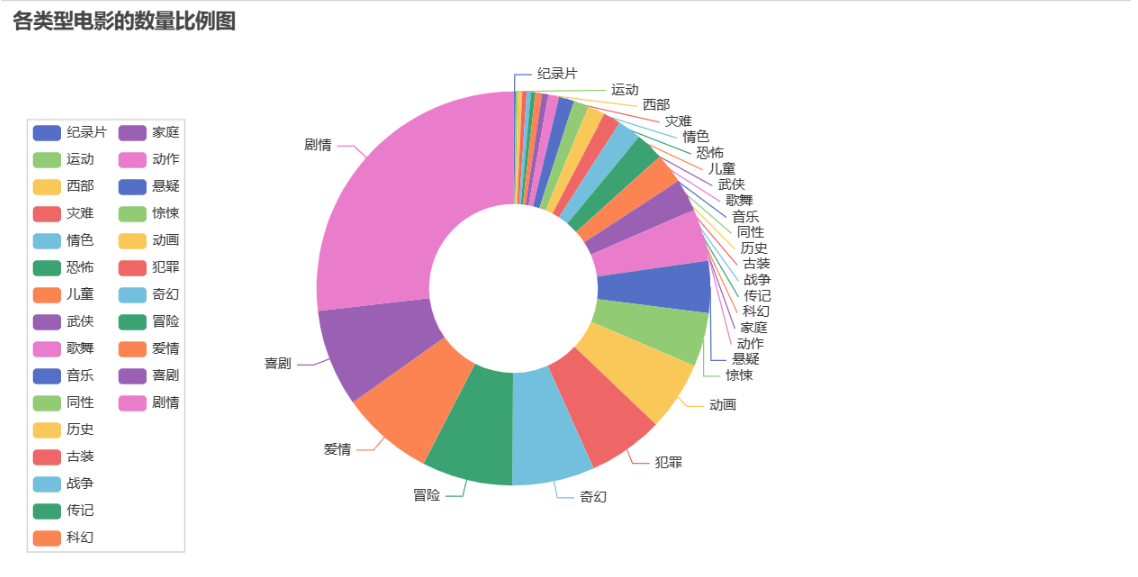
从这张图可以看出各类型的电影量随年份而变化的情况，从中可以较为明显的看到其中有一两种类型的电影在某段时期数量特别突出。

9、这里是用了matplotlib做的一个电影类型的数量的横向柱形图：



从柱形图中分析出，剧情电影的数量是最多的，而纪录片与之却差了很远，这也是大众喜好的缘故，由于剧情片更为有趣，有连贯性，因此被更多人喜爱，那么出的数量自然也就多了。

1. 这里是用了pyecharts做的一个各类型电影的数量比例图：



这是一个非常直观的饼图，且观赏性很强，可以明确看出剧情类型的电影数量是最多的。并且此饼图是从多到少排开的，让人容易理解。

1. 这里是用了pyecharts做的一个豆瓣电影Top200评论人数的词云图：



这个词云图可以直接看到前豆瓣Top200电影的名字，并且鼠标移动到相对应的电影名字时可以显示出对应的评论人数。

1. 结论与收获

结论：从电影类型来说，在豆瓣前200的电影中，最受欢迎的电影类型是剧情、喜剧和爱情。这说明人们更喜欢情感化的电影内容，而不是过于商业化的电影。因此给各大导演证明了人们更喜欢看情感细腻的电影。从豆瓣前十电影可以看出，

美国和英国是豆瓣前200电影的主要制片国家。这也表明了这两个国家在电影制作方面的重要性。而豆瓣前200电影的平均评分为8.6分，这表明这些电影都是优秀的作品，受到了观众的高度评价。有些类型的电影数量在不同时期波动较大，这可能跟社会的潮流存在一定的联系。

从以上结论中，我们可以看出人们更喜欢情感化的电影内容，同时美国和英国在电影制作方面具有重要性，而豆瓣前200的电影评分都较高并且是由民众选出的，因此都是可以推荐给各大观众。希望这此的作业分析可以帮助电影事业的现状，这些结论对于我们在电影制作或推广方面有一定的参考意义。

收获：这次的作业让我收获颇多，学习到了许多不熟悉的爬虫知识，并且把之前的知识进行了加固。知道了网页爬取前要先确定是否可以爬取，爬取后数据是否有效。而此次作业也结合了数据分析内容，虽然在分析过程中遇到了不少困难，但是通过不断解决问题也让我获得了成就感和学习满足感，并且对以前学的数据分析有了更为深入的了解和学习，使得数据分析的能力更上一层，这次的期末大作业让我在遇到困难时学会自己搜寻办法；自己学习；自己解决，使我受益匪浅。谢谢老师！