1. 简介

豆瓣用户每天都在对“看过”的电影进行“很差”到“力荐”的评价，豆瓣根据每部影片看过的人数以及该影片所得的评价等综合数据，通过算法分析产生豆瓣电影 Top 250。本文对于"好评电影"（豆瓣电影top250）的影片评分，上映时间，评论人数，电影类型进行分析，期望为渴望寻找 优秀影片的观影者 以及 电影制片方 对电影的选择提供若干参考建议。

为了完成对豆瓣电影Top250的数据进行爬虫，并对爬取的数据进行数据分析，本文主要完成 了一下工作：（1）使用request请求技术对豆瓣电影Top250进行数据抓取，并使用lxml提取技术对数据进行处理，完成前期的数据爬取和整理工作。（2）使用matplotlib以图标的展现形式对爬取的数据进行分析并归纳总结。

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成

1. 数据爬取

本文进行数据分析的主要目的是为了获取豆瓣电影Top250的信息，首先需明确目标信息的爬取路径。具体爬虫步骤如下：

1. 明确爬虫路径：<https://movie.douban.com/top250>
2. 请求库与提取：requests、lxml

文本

描述已自动生成

文本

描述已自动生成

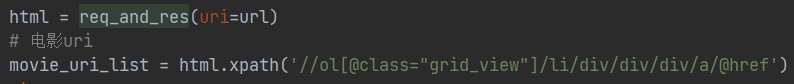
1. 分析翻页：start为偏移量，步长为25



1. 分析详情页字段数据xpath路径

图形用户界面, 文本, 应用程序, 电子邮件

描述已自动生成



文本

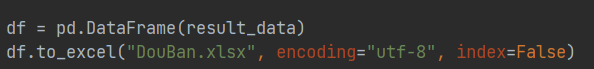
描述已自动生成

1. 添加随机User-Agent与请求间隔避免反爬

文本

描述已自动生成

1. 从网页中提取出top250电影的数据存储excel：pandas



1. 根据爬取结果，进行简单的可视化分析

图形用户界面, 表格, Excel

描述已自动生成

1. 数据分析

俗话说“文不如表，表不如图”。下面根据excel的数据结果，进行简单的数据可视化分析，并用图表呈现。

1. 各年份进入top250的电影数量统计

站在漫长的百年电影史的时间角度上，分析一下哪些年份"贡献了"最多的电影数量，也可以说是"电影大年"。

可以看出，可以看到，250部电影来自56个年份。其中1994、2004、2010年上榜电影数量最多，达到13部；其次是2013年的11部。20世纪90年代开始电影评分成上升趋势。

再往后看，发现遥远的1931和1936年也有电影上榜。那会儿应该还是黑白电影时代吧，看来电影的口碑好坏跟外在的技术没有绝对的关系，质量才是王道。

图表, 直方图

描述已自动生成

文本

描述已自动生成

1. 各类型top250电影分析

根据统计直方图分析得出，剧情、爱情、喜剧、冒险、犯罪题材类型的电影数量最多，其中，剧情剧高居首位，其次爱情剧紧随其后，喜剧也名列前三。恐怖、色情、运动、灾难这些较为小众的题材则占比较低。

图表

中度可信度描述已自动生成

文本

描述已自动生成

1. 评分top30电影分析

想看一看评分最高的前30部电影是哪些？

可以看到：排名最高的是9.7的"肖申克的救赎"，其他还包括"美丽人生"、"霸王别姬"等也有9.6的评分。

图表, 折线图

描述已自动生成

文本

描述已自动生成

1. 总结

豆瓣电影Top250是豆瓣根据每部影片看过的人数以及该影片所得的评价等综合数据，通过算法分析产生的排行榜。能在一定程度度上反映各部电影的影响力、关注度和观众喜爱程度。

以下将从个方面进行总结：

1. 总体：豆瓣电影Top250评分平均值为8.868（十分制），最高评分9.7（肖申克的救赎），最低评分8.3（驴得水）。最多电影评分8.8，有47部。
2. Top30：从豆瓣电影Top250的评分与热评人数前十来看，不少都是广大群众心中的经典之作。
3. 逐年变化趋势：总的来看，1931-1983年都有稳步产出1-2部优质电影，1983年以后呈增长趋势，其中1994-2015年是高产出期，直到2015年开始有下滑迹象。中国在1961年开始出现第一部上榜优质电影（大闹天宫），1979年开始有优质电影陆续产出，但数量不多。美国电影制作较为成熟，希望中国能稳住电影质量，继续加油产出更多优质电影。