# 项目技术说明书

豆瓣电影数据可视化

小组名称： 第五组

学员姓名： 周召敬

目录

[项目技术说明书 1](#_Toc17289747)

[1. 项目介绍 2](#_Toc17289748)

[2. 需求分析 2](#_Toc17289749)

[3. 技术选型 3](#_Toc17289750)

[4. 爬虫开发和优化 3](#_Toc17289751)

[5. 数据清洗和存储 4](#_Toc17289752)

[6. 数据查询和可视化展示 5](#_Toc17289753)

[7. 开发环境 5](#_Toc17289754)

[8. 产品原型 6](#_Toc17289755)

# 项目介绍

豆瓣电影数据可视化项目通过对大量电影关键数据进行收集与分析，利用可视化的方式为用户呈现了各种统计结果。

# 需求分析

1. 提现准备好单个电影url存放地址：[https://movie.douban.com/j/new\_search\_subjects?tags=电影&start=20&year\_range=2009,2019](https://movie.douban.com/j/new_search_subjects?tags=%E7%94%B5%E5%BD%B1&start=20&year_range=2009,2019)
2. MongoDB 数据库存放数据，npm下载mongoose框架，便于操作数据库
3. 构建爬虫运行服务器，以及各种所需要的框架
4. 爬虫脚本代码。

# 技术选型

Node.js：前端应用开发方法

Cheerio：实现HTML数据解析

MongoDB：对需要的数据进行存储

eCharts：数据可视化展示

superagent：客户端请求代理模块（支持HTTPS）

# 爬虫开发和优化

1. 首先应该了解获取数据的各个url。
2. 然后对每个url地址的页面信息进行分析，
3. 连接数据库并建立model，确定json文件格式
4. 编写爬虫代码，首先通过豆瓣数据库网页，一组获取20个电影的url
5. 然后对每个url进行页面信息爬取
6. 使用cheerio解析html，配合正则获取数据信息

# 数据清洗和存储

1.对信息进行清洗，正则表达式匹配数据，拆分获取

2.如果存在重复不存入数据库，如果存在数据项为空的不存入数据库

# 数据查询和可视化展示

(根据不同条件在mongodb数据库查询对应的数据集，然后使用echarts进行可视化展示，生成<产品原型>部分的可视化图表)

# 开发环境

开发环境：Windows 10 专业版

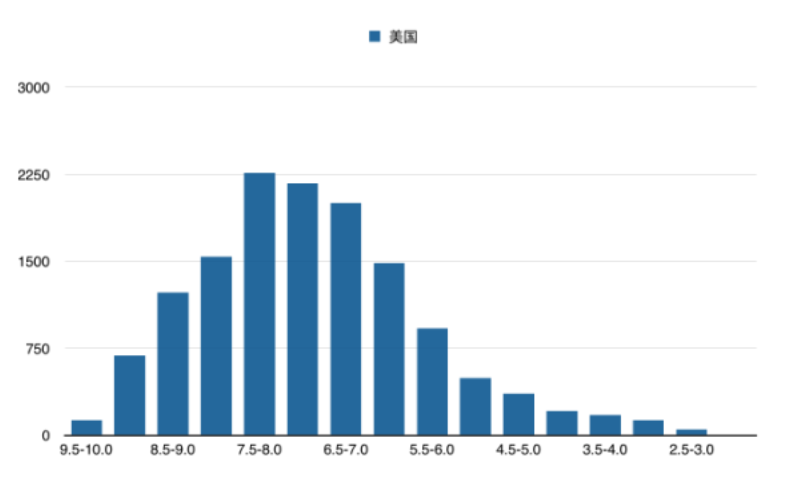
IDE：VS Code （使用免费，集成各种插件，已使用很久，很熟悉。写代码更方便，调试代码更方便）

# 产品原型

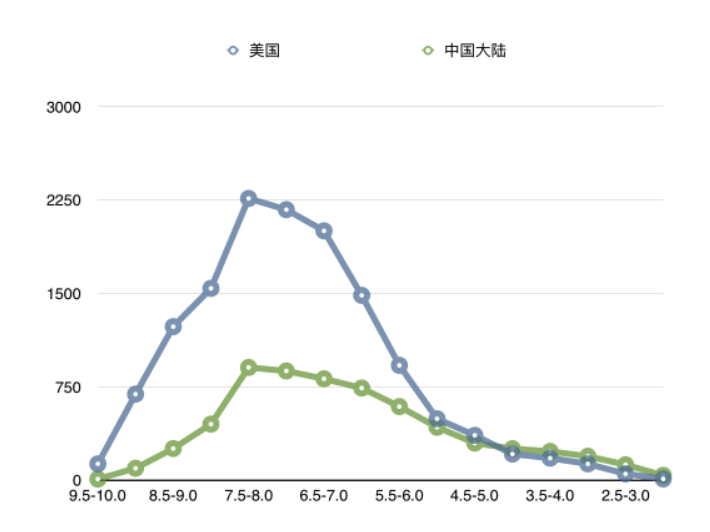
(主要说明自己所负责的的模块最终的呈现效果)

|  |  |
| --- | --- |
| 豆瓣电影数据可视化分析平台V1.0 | |
| 影评分和电影数量分析 | echarts展示 |
| 影评分和电影数量对比分析 |
| 近十年每年发行的电影分析 |
| 2019各个国家发行的电影数量分析 |
| 近十年中国发行的电影评分 |
| 近十年中国发行的电影每一年的数量分布图 |
| 2019年评分在5分以下的电影名称 |
| 自己自由发挥 |

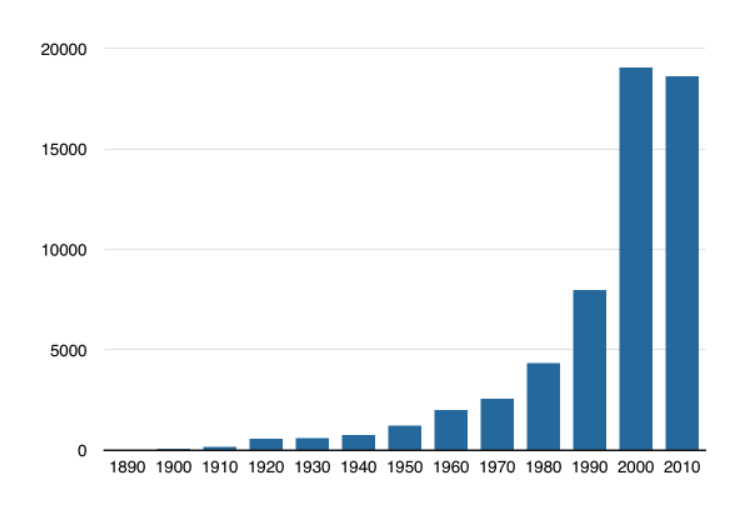
1. 根据国家生成电影评分和电影数量的柱状图



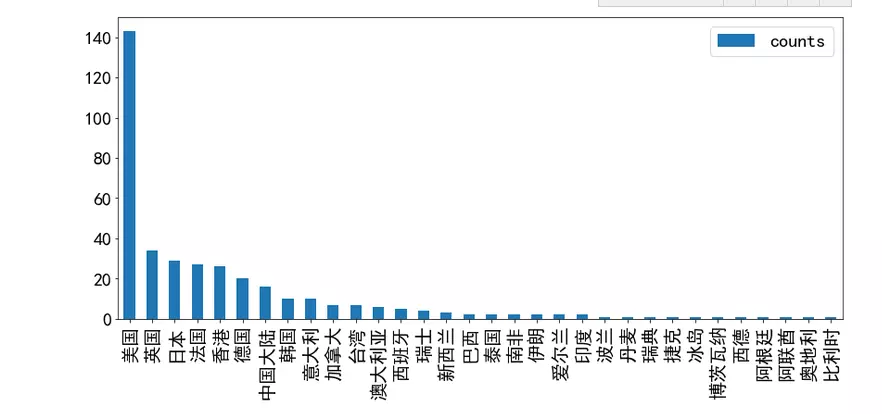
1. 对比任意两个国家电影评分和电影数量的折线图



1. 近十年每年发行的电影数量分布柱状图



1. 2019年截至到今天为止各个国家发行的电影数量分布图



1. 近十年中国发行的电影评分(0~3、3~5、5~7、7~10)的范围的数量分布图
2. 近十年中国发行的电影每一年的数量分布图
3. 2019年截至到今天为止评分在5分以下的电影名称
4. 自己自由发挥…(加分项)