# 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： 黄学伟

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

爬取近十年豆瓣的电影数据，最后用可视化插件进行可视化显示。

# 需求分析

需要爬取大量的电影数据，MongoDB能够提供我们所需求的技术，并能够通过筛选来呈现结果。

# 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

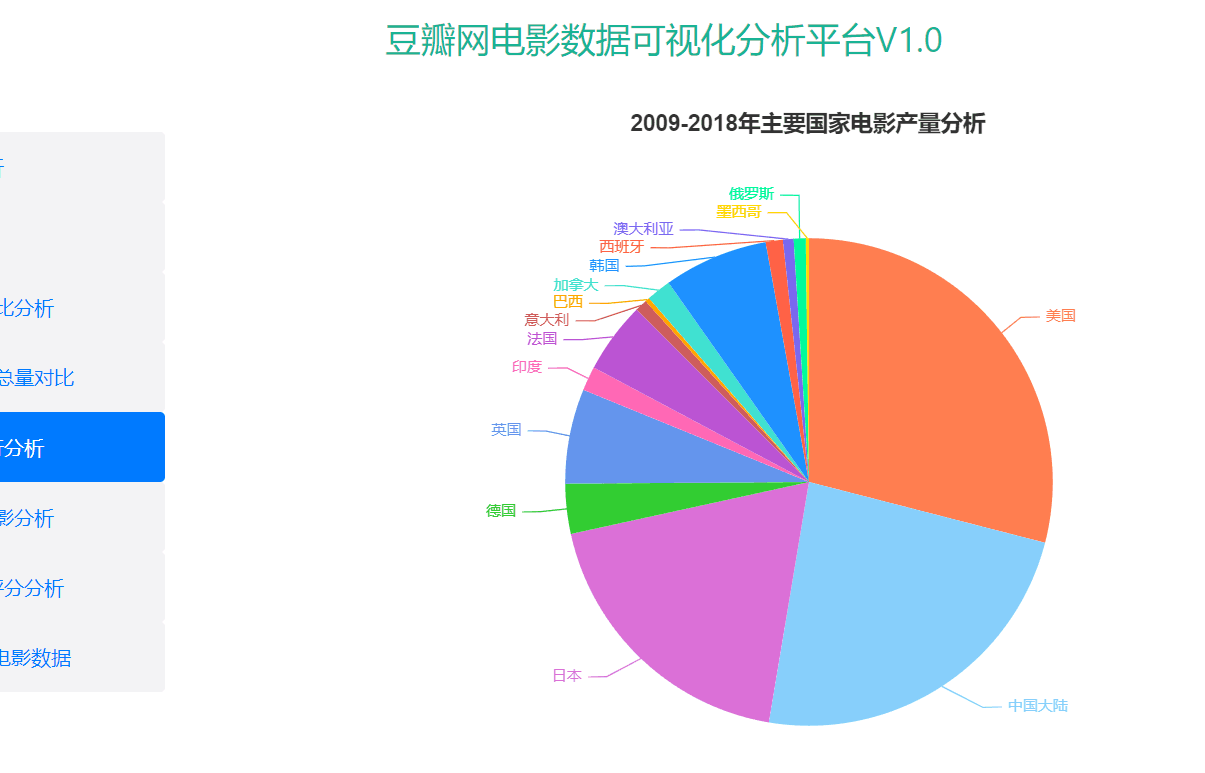
superagent：客户端请求代理模块（支持HTTPS）

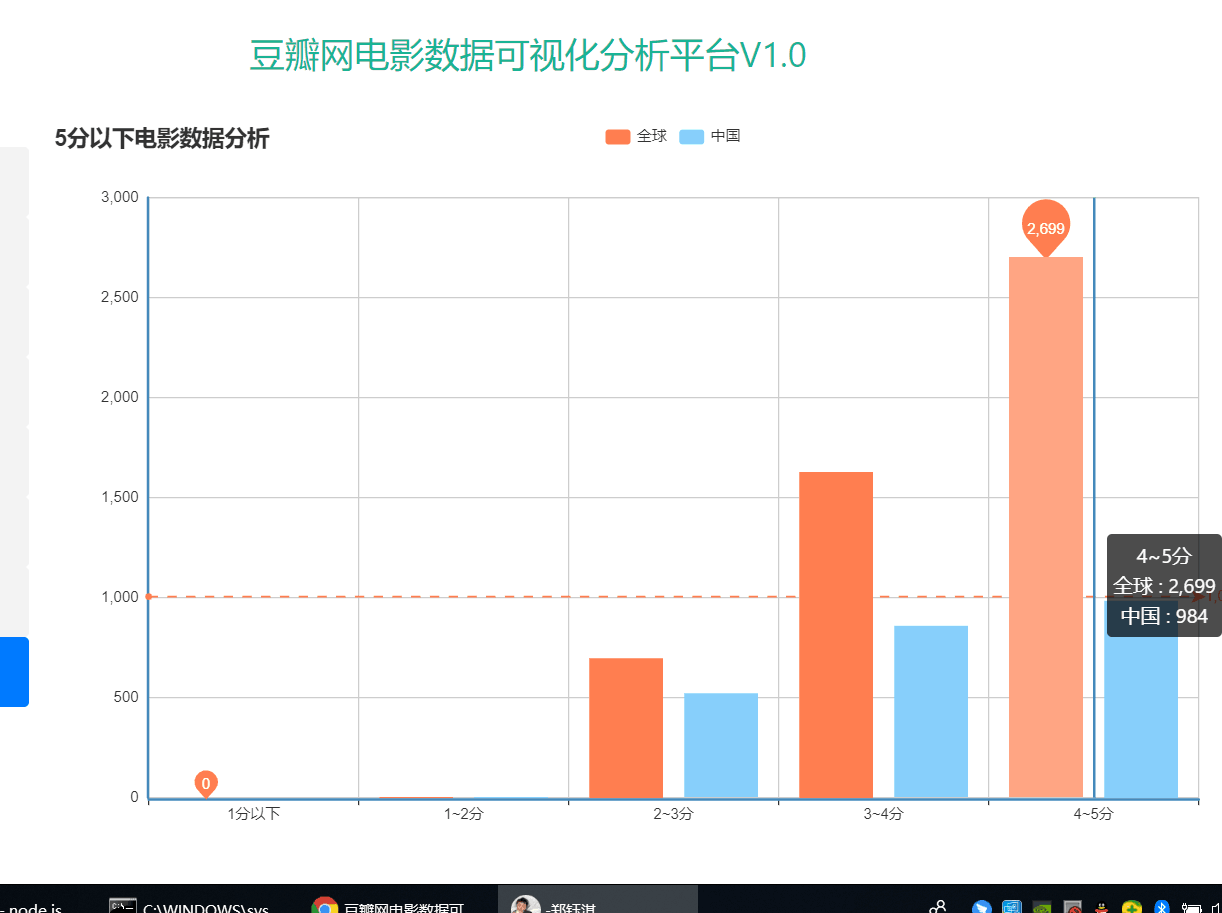
# 爬虫开发和优化

1. 连接数据库并建立model
2. 模拟豆瓣电影加载更多按钮发送请求获取一组（20个）电影对应的url
3. 遍历获取的一组url，根据url发送请求获取html
4. 使用cheerio解析HTML并配合正则表达式进行获取所需数据
5. 判断所得数据在数据库里是否已存在，是则不进行存储，否则存储数据
6. 重复2-5直至所有电影数据均被爬取
7. 如果电影没有评分则不进行爬取

# 数据查询和可视化展示

根据项目所需要求，进行数据筛选并使用echarts进行数据的显示，如图





# 开发环境

IDE：VS Code ：免费，插件丰富，调试方便

# 产品原型

|  |  |
| --- | --- |
| 豆瓣电影数据可视化分析平台V1.0 | |
| 2015-2016两年电影数据的爬取 |  |
|  |
|  |
|  |
|  |
|  |
|  |
| 自己自由发挥 |