# 项目总结报告

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

[项目总结报告 1](#_Toc17289904)

[1. 项目基本信息 2](#_Toc17289905)

[2. 项目成果 2](#_Toc17289906)

[3. 开发工具与环境 2](#_Toc17289907)

[4. 项目需求变更 3](#_Toc17289908)

[5. 项目计划和进度实施 3](#_Toc17289909)

[6. 经验总结 3](#_Toc17289910)

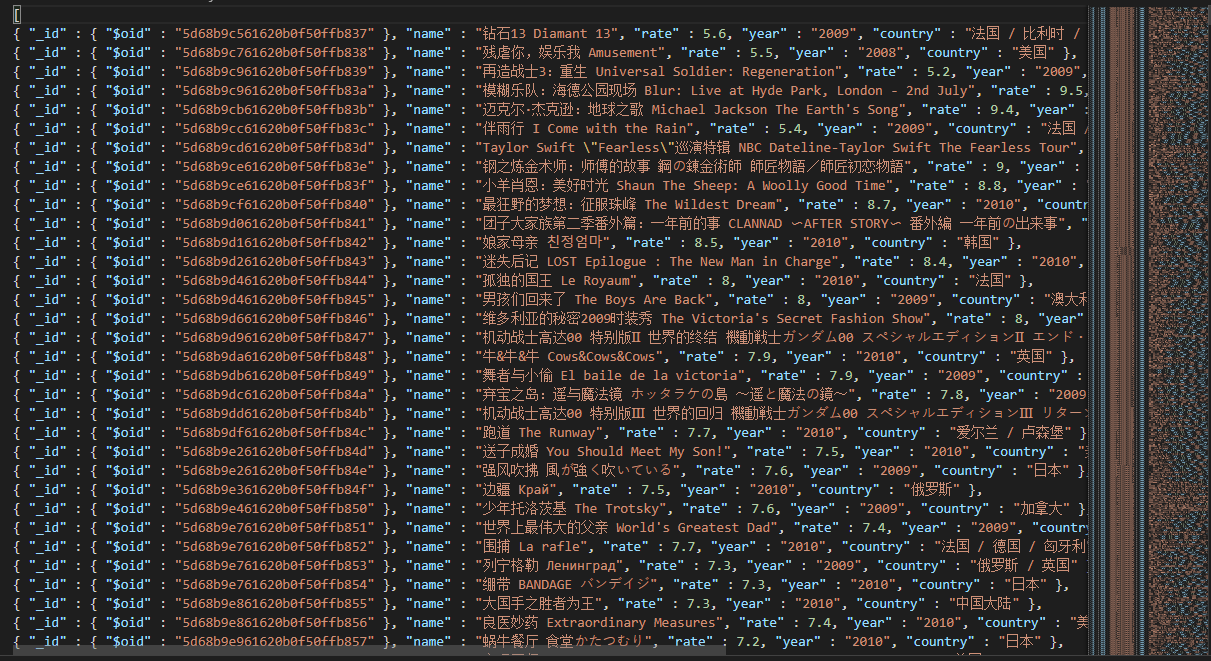
[7. 教训总结 4](#_Toc17289911)

[8. 其他改进意见 4](#_Toc17289912)

# 项目基本信息

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。

# 项目成果

主要负责2009-2010的电影数据爬取（包括爬虫代码自己完成，最后只需要提交个数据文件）

# 开发工具与环境

Node.js：主要开发语言

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）

mongoo：对数据存入，取出进行操作

# 项目需求变更

(主要说明在项目开发过程中项目出现的变更需求，出现变更的原因，以及变更导致的结果和影响等)

1.需要爬取的数据需求变更：

爬取名字、评分、国家、年份、评论人数 => 爬取名字、评分、国家、年份、导演

=> 爬取名字、评分、国家、年份

变更原因：电影子页面没有评论人数、图表中没有需要导演数据的内容

导致的结果和影响：修改爬虫代码

2.需要存储的数据需求变更：

存储所有电影 => 只存储评分、年份不为空的电影数据

变更原因：存储空评分或年份不利于数据统计

导致的结果和影响：修改爬虫代码cheerio部分及存储数据库部分

3.导出文件的变更：

首先要求导出文件必须为数组+对象格式，中间必须有’，’隔开 => 导出格式文只存在对象，不用‘,’隔开。

变更原因：部分组员直接代码导出格式不一样

导致结果和影响：另一部分组员重新导出一遍。

# 项目计划和进度实施

一共十年数据，每个人爬两年数据，剩余一人完成数据可视化部分内容

9月2日 已完成

# 经验总结

1. 对项目实施的分配，制作，以及结果有了清晰的把握。
2. 对爬虫代码，以及es6的一些解决异步的函数更为了解
3. 对数据更加敏感。

# 教训总结

1. 由于单词拼写错误，需要重新修改一遍数据库，不过作业量小。不存在不能按时提交的情况
2. 代码不断更新，因为豆瓣网有反爬机制，所以只能不断重爬，但是到了后面意识到，爬到什么位置，接下来便从那重新开始的重要性。

# 其他改进意见

建议别爬取容易被封ip的网站，导致个人账号，以及本地连接豆瓣直接403....