天津工业大学计算机科学与技术学院

实训总结报告



　项目名称: 基于豆瓣资源的个性化信息整合库平台

　合作培养企业：天津市融创软通科技股份有限公司

　专业方向：大数据方向

　项目负责人：何晶

　专业班级: 软件Z1801

姓名：郝子成

　实训起止日期：2020.6.15-2020.7.15

**1 引言** （一级标题，小三黑体，居中）

1.1项目概述（二级标题，四号黑体，左对齐）

说明项目目的、背景、意义及参考资料（包括企业提供的素材、DEMO）等；

1.2 需求分析

为什么做这个题目，有什么实际价值

1.3 运行环境

软、硬件环境

（正文，小四号宋体，1.25倍行距）

**2 项目设计**（小三黑体，居中）（以个人主要负责内容为主）

2.1 设计思路

2.2 模块功能介绍

2.3 模块结构图

2.4 程序流程图

2.5 功能设计分工

需详细标注每部分的内容，具体任务及任务安排等。

**3 详细设计**（小三黑体，居中）

详细写明项目的设计（系统设计、数据库设计）等，并写明对所涉技术、核心功能等描述；

**4 实现方案**（小三黑体，居中）

详细写明项目功能实现描述、核心代码及实现结果等；

并对调试好的程序，从执行程序弹出界面开始，每一步操作截一个图，并附加说明（图片格式为JPG格式）。

**5 系统测试**（小三黑体，居中）

详细写明测试方法、测试描述（bug与修正方法等）、预期结果等；

# 引言

## 1.1项目概述

随着科技时代的发展，我们的社会也在飞速的发展。作为一个信息化的社会，我们的衣食住行都离不开各种各样的信息。例如我们习惯网购各式各样的衣服，在网上搜索各种各样的美食，然后网站收集你习惯搜索的信息，再为你进行个性化推荐。毫不夸张的说，我们的生活已经离不开各种各样的信息了，我们每一个人每时每刻都在产生大量的数据。而就是为了处理如此海量的数据，大数据这门研究就应运而生了。

我们这个项目叫做基于豆瓣资源的个性化信息整合库平台，其目的就是做一个可以收集大量数据将其插入数据库，并在后台提取出来对数据进行分析整合的平台。这么做的意义在于我们可以收集大量我们想要的数据，然后针对我们想要的方面进行统计分析。例如拿我们所爬取的电影数据来说，我们可以统计哪一年评分最高的前50名的电影，亦或是可以统计五星好评的人是男性多还是女性多，甚至可以精确到某一条评论来估计是正面情绪还是负面情绪。我们将结果以图表的形式展现，这样可以更直观的得到我们想要的结果。通过一系列功能的实现，用户可以更好地检索自己想要的东西，我们也可以通过海量数据更好地了解用户，更好地为用户服务。

参考资料：通过老师对各个模块的讲解，以及老师写的简单例子，加上网上资料的自学，逐步搭建起整个框架。

## 1.2需求分析

我们身处于信息化的时代，不仅我们想要的东西可以快速的检索出来，我们自己无时无刻也在产生大量数据，这就是大数据给我们生活所带来的实际影响。但是杂乱无章的数据是没有任何意义可言的，我们也无法快速找到我们想要的数据，更别提统计了。这就需要我们利用技术对数据加以整合分析，留下我们想要的数据，剔除我们不要的数据，然后将其规范化，最后可以利用得到的数据进行我们想要的分析，将结果进行可视化，这是我们做此项目的初衷。做出这个项目我们可以从任何一个网站使用爬虫爬出我们想要的数据，将其用一张规范的二维表整合进数据库，然后使用BS项目的开发结合前端的开发，将得到的数据提取出来，用前端技术做成一个平台，我们可以在这个平台上检索我们想要的数据，亦或是可以选择一些项目进行分析，用eCharts技术进行可视化，使展示更加直观、美观。这就是此项目的意义。

同时我们在此平台加入用户管理功能，使用此平台前需要登录，如果没有账户则需要注册。此外用户还分为普通用户和管理员用户，二者的权限有所差别，同时管理员用户可以管理普通用户。并且用户拥有标记自己喜欢的或者已看过的图书、电影的功能。

## 1.3运行环境

软件环境：Gitee负责版本管理，提交项目；Pycharm负责搭建scrapy框架，利用爬虫爬取数据；IDEA负责服务器端以及前端代码编写——服务器端开发框架SSM（Spring+SpringMVC+MyBatis），前端开发技术（jQuery+LayUI+eCharts）；连接远程数据库MySQL(Mariadb)；平台开发模式：基于JSON数据的前后端分离式开发。最后使用浏览器打开最终成型的平台。

硬件环境：运行内存8G的PC即可。

# 项目设计

## 2.1设计思路

一个软件开发的目的就是面对用户，去完成用户的要求，需要建立合适的思路，开发出能完成用户所需的项目。所以第一点是要明白用户需要什么样的项目，这个项目不仅要满足用户所需要的功能，也要将软件做的人性化，使得界面美观让用户愿意去使用，并将软件使用方法通俗易懂化。

①首先我们需要爬取数据，我们使用的是python相关的scrapy框架，使用的编程软件是Pycharm，我们写好爬虫，给其需要爬取的各项数据以及网页，让其自己运行，因为大部分网站设有反爬虫机制，所以要设置限制爬取请求数、延迟、模拟URL等，因为运行时间较长，考虑增加错误提醒设置以及中断重爬设置。将爬取的数据插入远程数据库，提前在数据库中建好相应的表，运行爬虫即可。

②接下来是后端配合前端的开发，后端的开发均已框架化，由于纯html和javascript技术是无法与数据库建立连接的，所以我们后端就是要干这些事情，要让我们的页面能够查询到数据库里的信息，也能让页面输入的东西传回数据库（例如注册用户等）。注意各个模块之间的联系，之间的互相调用。

前端开发主要是html、css、js等搭载layui框架所编写而成，此部分为各个页面的展示，注意各个页面之间的关系、跳转联系。

将二者结合起来便可以完成使用SQL语句从数据库里查出想要的内容的功能，简单的检索利用页面显示，复杂的使用eCharts进行可视化展示。

③在完成这些基本功能的前提下，对操作功能、页面等进行优化设计，尽量使界面更加美观，使使用更加方便。

由于此次我们为4人一组，一人负责一个模块的搭建，共同实现用户模块的

创建，在设计过程中由于使用Gitee以及搭建共通项目的原因，要注意差异的处理，将小框架一步步拼接成大框架。

## 2.2模块功能介绍

本次项目分别在两个平台上，除去PyCharm上的爬虫以外，在IDEA上再分为各个小模块。

1. Pycharm上的scrapy框架爬虫：此模块是用于爬取豆瓣所选内容的具体内容，例如每个电影的名字，再深入一层为电影的导演、类型等内容。提前在远程数据库中建好应该建好的表，将爬来的数据对应存入规范二维表中。此外，为了防止反爬虫机制的屏蔽，应该设置延迟参数等，并加入错误处理机制，以防爬虫突然的中断。
2. IDEA上的BS开发(Spring Boot框架）：此模块除了要用到爬虫爬来的存到数据库里的数据以外，应该是与第一个模块分开来的。
   * + 1. 登录页面：此界面是登录所用，该平台需要登录才可以使用，老用户输入账户密码验证码即可登录，而新用户需要注册，老用户输入的用户名密码不能为空且需要与后台数据库已经记录的某条用户信息所相匹配。该页面需要可以跳转到注册页面，此外还搭配了验证码，输入与随机验证码相匹配即可成功登录。
       2. 注册页面：此界面是注册所用，该平台需要登录才可以使用，没有用户的所以需要注册之后才可以使用。注册页面需要登记的信息（必填项）有：用户名、昵称、密码、需要再次确认密码以及验证码，注册页面可以跳回登录页面。在注册页面填写的表单数据会传回到数据中已经存在的用户表中，登录时会与用户表中数据进行匹配，相匹配的用户即可进入。
       3. 欢迎界面（即登录之后跳转到的页面）：此页面展示了此次小组的成员以及每个人爬取的数据量，右上角显示了当前登录用户的昵称以及自定义头像。在旁边设有安全退出按钮，单击之后可以退出到登录页面重新登录。在左边为导航栏，分别为四个数据模块以及用户模块的导航，单击即可进行跳转。
       4. 数据查询模块：分别为图书、音乐、电影、电视剧四大模块的数据检索，这四个模块的功能大同小异，通过与后台数据库的对接，从爬虫爬取来的数据中进行检索，通过某一特定项的检索，可以查出该条目所有详细信息。此外还可以将查询结果以表格或文档形式导出，亦或是可以连接打印机打印。
       5. 数据分析模块：对以上数据进行分析，使用eCharts功能，将某些想要知晓的内容（例如1-10得分占比），进行可视化分析，让其显得美观、清晰、简洁。