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。此处也为动态内容，随着数据库的变化而变化。
       6. 用户管理模块：分为管理员用户和非管理员用户，非管理员用户可以更改自己的个人资料（如昵称等），管理员用户除去非管理员用户的功能以外还有管理其他用户的功能，可以给予其他用户授权或是取消其他用户授权。
       7. PS：我喜欢我已看功能：每个新用户的我喜欢以及我已看均为0，用户可以对每项条目进行编辑，可以为打分评论，还可以将其添加到我已看和我喜欢，并将数据传入后台计数，体现在数据分析模块上，属于一个动态的模块。

## 2.3模块结构图

（一）

Spider爬虫文件从网站检索所选数据

pipelines文件负责连接远程数据库，把所得数据插入数据库

（二）①

新用户注册

老用户登录成功

登录页面

欢迎界面

注册页面

②

安全退出

图书

音乐

电影

电视剧

用户

欢迎界面

③

检索项

查询

重置

检索项

编辑操作

查询出的列名

④

对数据分析形成的各式各样图表

## 2.4程序流程图

豆瓣数据管理系统

登录注册

管理模块

数据总览

注册

登录

图书

音乐

电影

电视剧

普通用户

管理员用户

数据展示

对应数据查询

## 2.5功能设计分工

此次小组共四个人，在scrapy爬虫阶段每个人负责自己模块的编写，通过对不同页面的解析，爬取自己相对应的数据。各自在远程数据中建立自己的二维表，然后将爬来的数据插入二维表中，为后续奠定基础。

进入下一步编写后，仍然存在个人项目，但此时个人项目只是意义上的个人项目，大家通过版本控制实则在编写一个共通项目。需要由组长上传必要组件以及配置文件，组员通过gitee下载，然后系统自动导入所需运行文件即可。大家的文件夹目录均是一样的，在目录下创建自己的模块，自己模块包括数据的检索以及数据的展示。

再然后就是用户模块，用户模块分为前端代码以及后端代码，前端代码作为html代码以简洁、美观作为重要因素，后端代码需要实现一系列功能，如对接数据库，用户登录注册等内容。

分工完毕之后，先共同编写主界面等内容，然后再各自细化、为自己的模块添砖加瓦。总之，各个页面模块存在跳转、调用等关系，所以在编写时要注意它们之间的联系，它们既是独立存在也是紧密相联的，所以既要有独立的设计，也要存在互相联系。

# 详细设计

①首先是爬虫的设计。爬虫的设计使用的是python中的scrapy框架，在设置里需要与远程数据库对接。此次实训使用的是远程MySQL（Mariadb）数据库，所以我们只需在本地部署一个数据库管理工具即可，连接上远程数据库之后即可进行建表、查询等操作。注意在此框架里爬取数据和插入数据至数据库中是两条平行的大道，二者互不干扰，所以编写好爬取的项目之后，再分别写spider和pipelines，同时在settings里面配置相应项。在插入数据的同时应该设有更新操作以便遇到主键相同条目时候的更新而不是重复插入。

技术：sscrapy框架。功能：获取数据并将其插入数据库表中。

②检索功能的设计。检索功能的前提是数据库表中存在数据，同样需要连接后台数据库。通过前端页面以及后端框架的配合，得以实现此功能。编写需要检索的条目，框架执行语句去数据库中相关的表中匹配，将匹配项的所有内容进行输出展示。在编写前端的时候，使用了很方便实用的Layui框架。

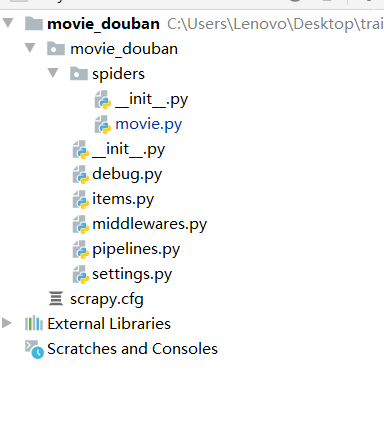
技术：Spring、Layui、html。功能：检索自己想要查询的条目。

③页面的编写。页面使用layui框架，具体如html页面一样。

技术：html、LayUI.

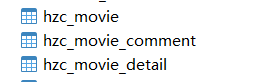
# 实现方案

首先通过scrapy的命令行创建项目，创建完成后的项目目录为

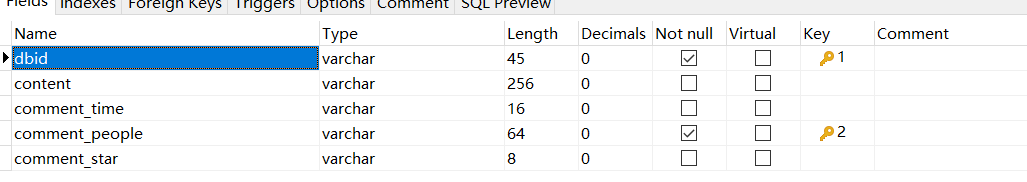


注意将项目创建在git clone的目录下，以便于项目的上传。

在数据库中创建对应的表：

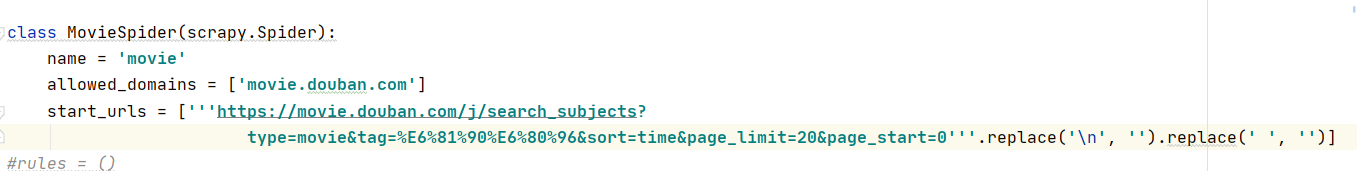


对应数据格式：



首先是spiders(虫子）下的movie爬虫文件：

设置爬取的范围：



由于电影在豆瓣上不是翻页来查看的形式，而是以加载更多的形式展现，点一次加载更多页面不会刷新，点一次加载20个新数据项，是json动态页面，以下是加载更多的设置：

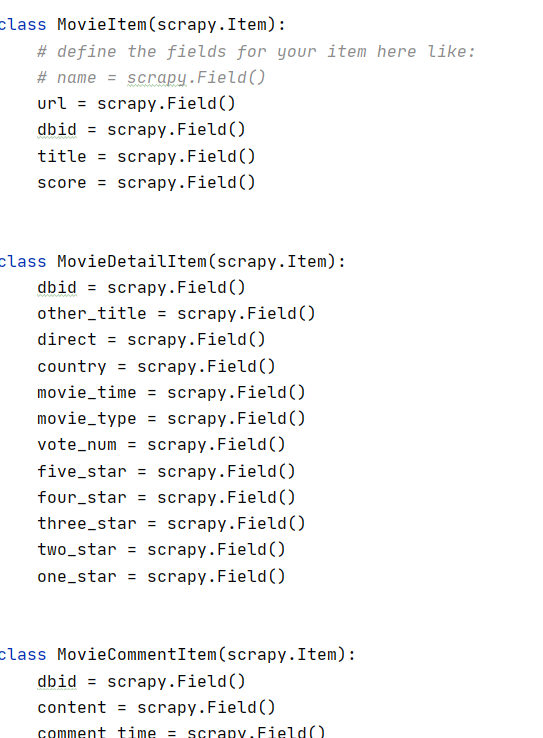


接下来便是各个方法的编写：包括爬取详情、评论等内容：

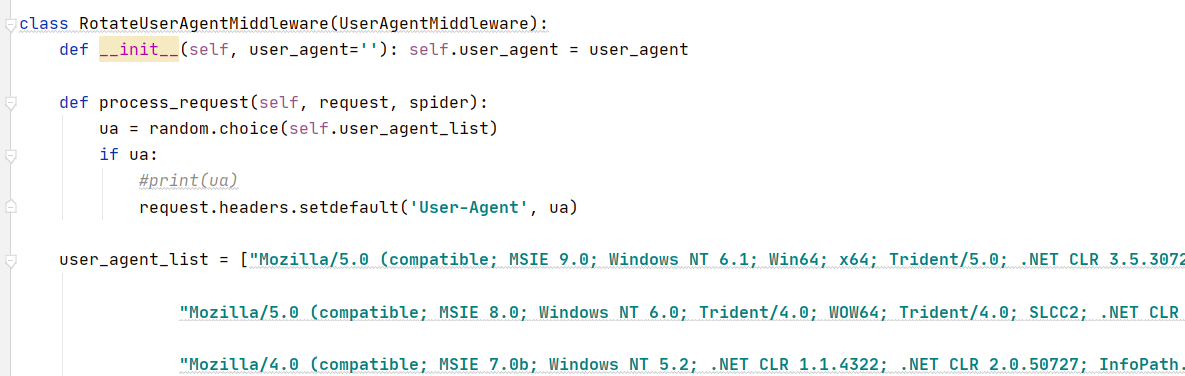


注意在上一张图对方法进行压入（yield）操作处理

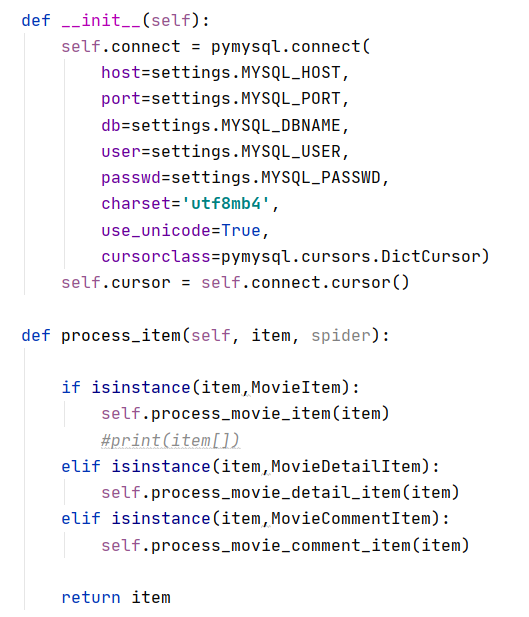
要爬取的内容的变量定义：



关于中间键的配置，这是为了模拟ID，防止爬虫被屏蔽：



Pipelines中设置对数据库的连接，用SQL语句进行插入、更新操作：

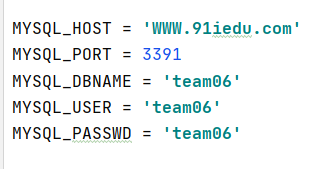


只截取一个函数插入的操作：

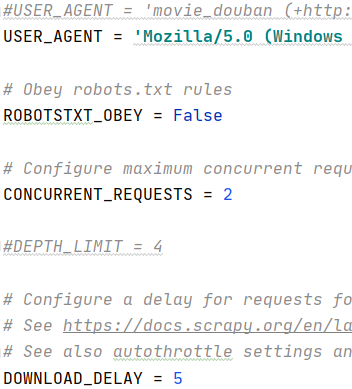


最后是settings里的设置，配置这些设置是为了完善、优化爬虫：

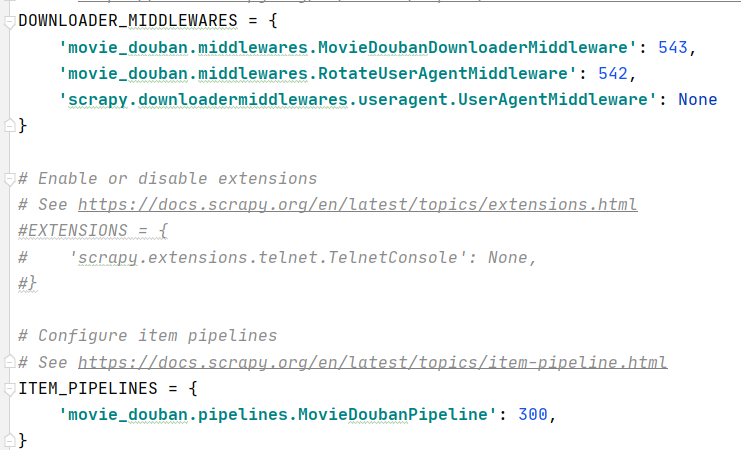
远程数据库配置：



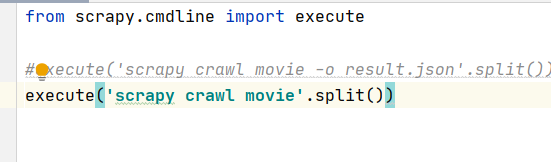
延迟、请求数等：



打开pipelines以及中间键：



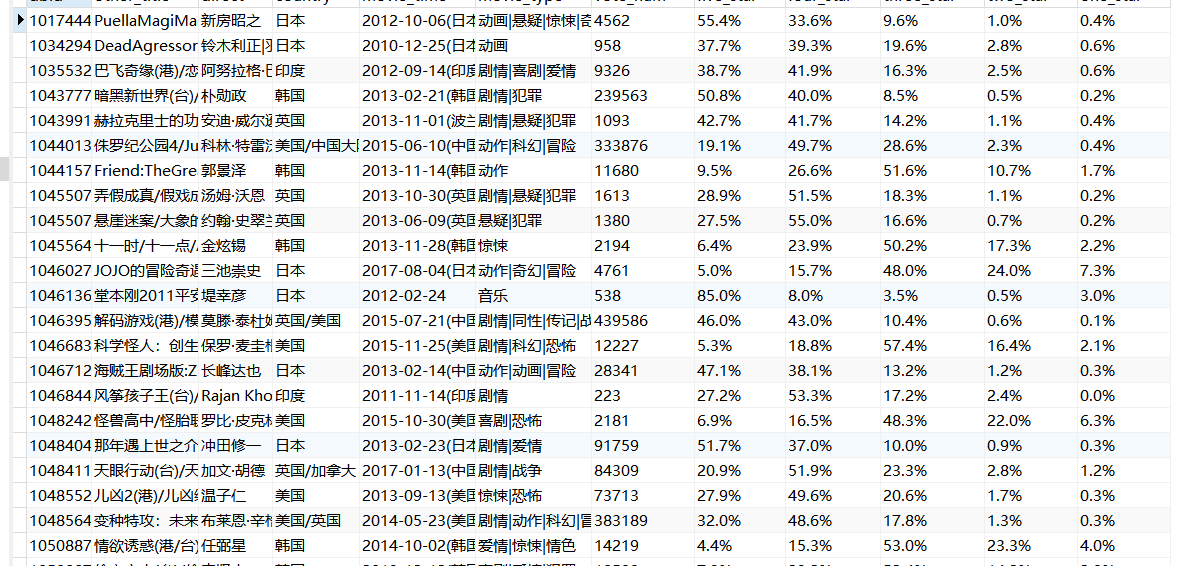
最后是运行文件：



运行结果（爬取时的状态）：

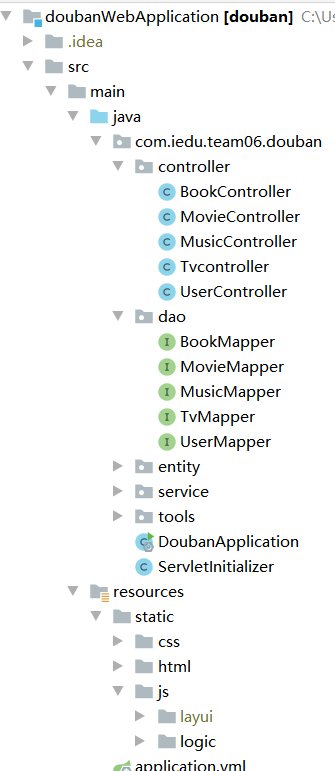


数据库展示：





IDEA：——目录结构：

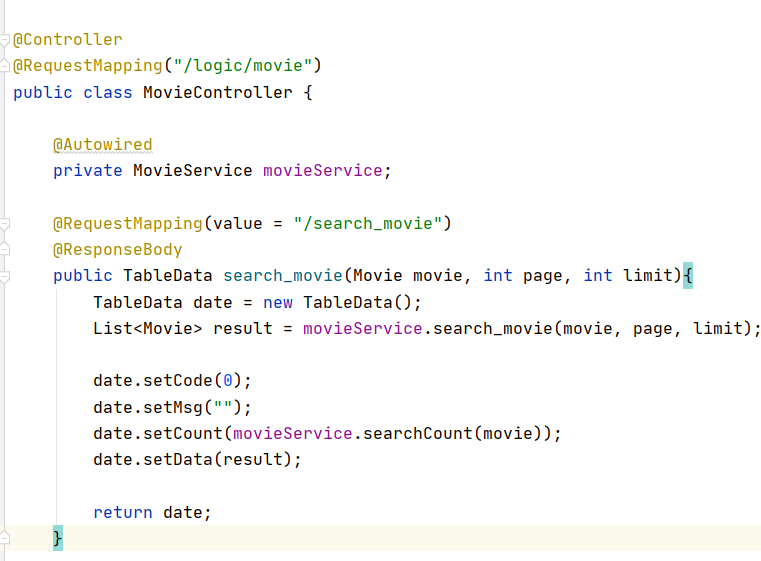


实体类里定义需要从数据库中提取的项，注意不要使用下划线命名法：（最下面关于评论的定义是错误的）



剩下的文件互相调用、关联性较强：

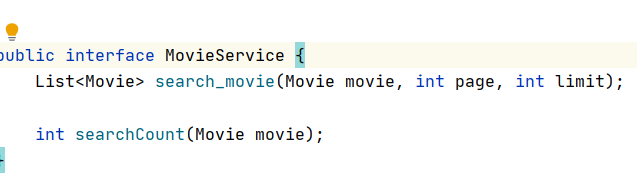
Controller层：



dao层的MovieMapper:



用来写 sql语句进行查询的



MovieServiceImpl implements MovieService



查询页面部分截图：



与其配套的JS页面：





要注意里面的变量名要与entity中的相同，且不能用下划线命名法。

导航栏：



注册页面：



运行结果：在浏览器中输入<http://127.0.0.1:8080/html/login.html>即可进入登录页面：



注册页面：



电影查询功能：





# 系统测试

当写代码的时候并没有遇到报错即编译没有错误的时候，会出现运行不了的情况。此时要在代码中每个方法前打上断点然后采用分步运行的方法进行运行下一步的操作，直到找到出现错误的方法。或是对可能出错的语句进行注释，这样可以采用逐个方法调试的方法找到出错的地方，直到找到为止。

观察是否是函数调用错误，还是传参错误等，要通过不断地写不断地调试积累找错的经验，才能做到在遇到错误的时候顺利地解决。

我们所遇到的bug有id无法寻址，某些语句未定义等问题，在自己调试以及询问老师之后逐渐完善解决了所有问题。预期结果即为实现各个功能。