目录

**[一、课程设计目的 1](#_Toc8249)**

**[二、设计任务及要求 1](#_Toc29683)**

[（一）设计任务 1](#_Toc18934)

[（二）设计要求 1](#_Toc11271)

**[三、概要设计 2](#_Toc4728)**

[（一）功能实现 2](#_Toc20074)

[（二）开发环境 2](#_Toc19763)

[（三）开源框架 2](#_Toc30633)

[（四）类描述 3](#_Toc1989)

[（五）类关系图 5](#_Toc23843)

**[四、系统测试 5](#_Toc28516)**

[（一） IDE内测试 6](#_Toc22340)

[（二）jar运行测试 7](#_Toc683)

**[五、总结 10](#_Toc23186)**

# 一、课程设计目的

该项目的主要目的是：通过实现网络爬虫来获取需要的信息，将抓取的信息分类整理后存储到txt文档中，在之后的项目实践中调用所需信息。例如，在深度学习的模型训练中需要大量的图片训练数据才能得到一个准确可靠的深度学习模型，通过提取并下载txt文档中的图片链接就可以获得模型训练的训练数据集、校验数据集等内容。

# 二、设计任务及要求

## （一）设计任务

抓取豆瓣电影和豆瓣读书的评分榜单前250名的信息，如电影名称、电影评分、一句话精彩影评、电影的海报链接和电影详细页面信息等内容，并通过将信息存储到txt文档中实现信息的查询和提取。例如通过查询文档中电影海报或图书封面的链接下载图片，同时查询电影或图书的名称来赋予图片相匹配的名称，实现将图片存储到本地。

## （二）设计要求

通过设计一个网络爬虫来下载网页文本信息，分解网页，遍历网络，存储主要信息。

利用httpcomponent通过HTTP协议来访问网络资源，下载网页的文本信息，将文本信息html的字符流转换为字节流存储到Java的String数据类型中。只获取文本信息，并不直接下载图片、视频、音频等。

利用Jsoup来解析html的信息，通过分析html的标签、类名、CSS或jQuery样式来提取所需信息。

将提取到的信息写入本地磁盘内的txt文档保存。

# 三、概要设计

## （一）功能实现

1 与URL建立链接

此部分通过httpcomponent（httpclient）实现一个客户端，能够向URL发送GET、POST等请求，并获取回应Response。

2 将字节流转换为字符流

BufferedWriter提供通用的缓冲方式文本读取，readLine读取一个文本行，从字符输入流中读取文本，缓冲各个字符，从而提供字符、数组和行的高效读取。BufferedWriter 由Writer 类扩展而来，提供通用的缓冲方式文本写入， newLine使用平台自己的行分隔符，将文本写入字符输出流，缓冲各个字符，从而提供单个字符、数组和字符串的高效写入。

3 解析html文本信息

通过Jsoup解析字符流形式的html文本信息，通过html的标签、类名或者css样式、jQuery来解析文本，获取所需信息。

4 反爬虫

通过设置user-agent的方式，在post或者get的时候添加请求头，可以使用requests库来添加，在post，get或者其他方法是加上headers参数就可以了。

5 写入文件

通过FileWriter 类构造相关对象，将解析后的html信息写入本地磁盘中的txt文档。

## （二）开发环境

1 开发平台：Windows10

2 集成开发环境：IntelliJ IDEA 2020.2.3

3 JDK版本：JDK1.8

## （三）开源框架

### 1 httpcomponents-client-4.5.13

httpcompoent 是Apache Jakarta Common 下的子项目，可以用来提供高效的、最新的、功能丰富的支持 HTTP 协议的客户端编程工具包，并且它支持 HTTP 协议最新的版本和建议。使用 HttpClient 需要以下 6 个步骤：

①创建 HttpClient 的实例

②创建某种连接方法的实例，在这里是GetMethod。在 GetMethod 的构造函数中传入待连接的地址

③调用第一步中创建好的实例的 execute 方法来执行第二步中创建好的 method 实例

④读 response

⑤释放连接。无论执行方法是否成功，都必须释放连接

⑥对得到后的内容进行处理

### 2 jsoup-1.13.1

Jsoup 是一款Java 的HTML解析器，可直接解析某个URL地址、HTML文本内容。它提供了一套非常省力的API，可通过DOM，CSS以及类似于jQuery的操作方法来取出和操作数据。使用Jsoup需要以下3个步骤：

①从一个URL，文件或字符串中解析HTML

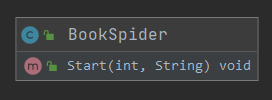
② 使用DOM或CSS选择器来查找、取出数据

③ 可操作HTML元素、属性、文本

## （四）类描述

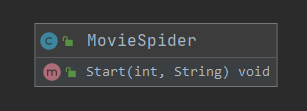
### 1 BookSpider.class

豆瓣书籍top250榜单爬虫核心功能类。通过Jsoup解析字符流形式的书籍榜html文本信息，通过html的标签、类名或者css样式、jQuery来解析文本，获取所需信息。并将获取的信息写入txt文档。



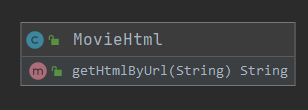
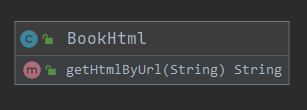
### 2 MovieSpider.class

豆瓣电影top250榜单爬虫核心功能类。通过Jsoup解析字符流形式的电影榜html文本信息，通过html的标签、类名或者css样式、jQuery来解析文本，获取所需信息。并将获取的信息写入txt文档。



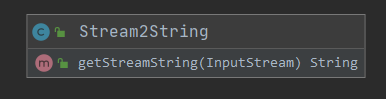
### 3 BookHtml.class/MovieHtml.class

书籍/电影HTML页面获取类。根据传入的url参数获取相关的HTML页面。



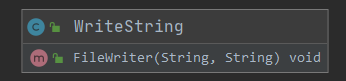
### 4 Stream2String.class

输入流转换类。将抓取到的HTML输入流转化为string类型数据。



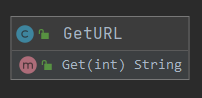
### 5 WriteString.class

文件写入类。实现将string类型数据写入指定位置文件夹内的txt文档。



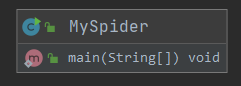
### 6 GetURL.class

Url获取类。实现对top250榜单url的自动获取和跳转。

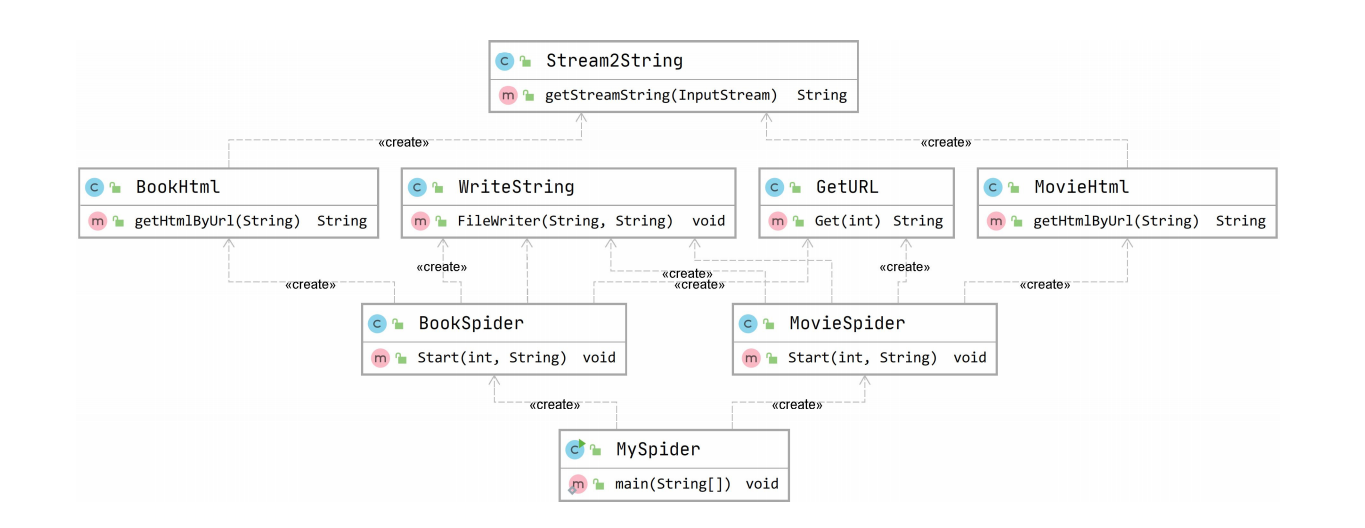


### 7 MySpider.class

程序控制类。控制书籍和电影爬虫的运行，提供程序主函数入口。

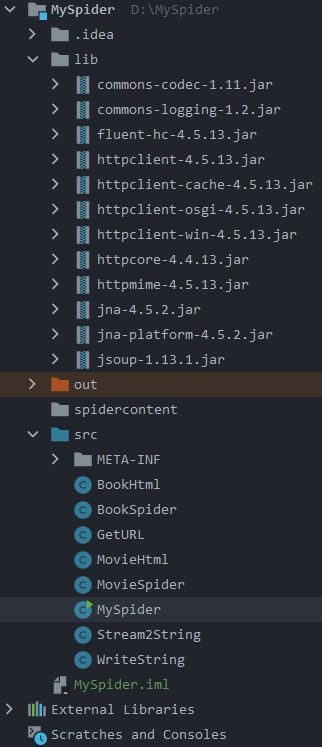


## （五）类关系图



# 四、系统测试

本工程目录结构如下：

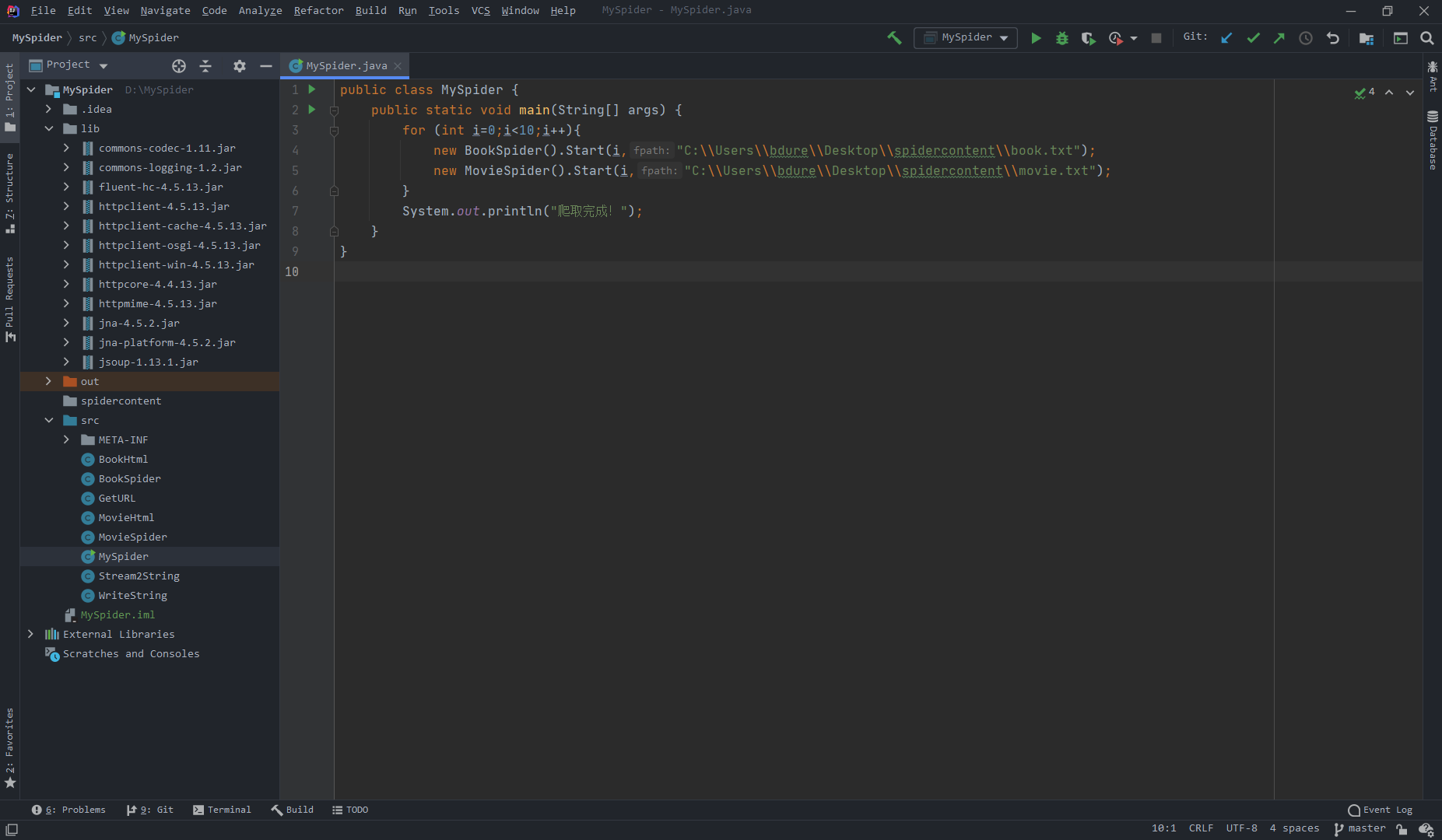


src文件夹内存放各类源码；lib文件夹内存放第三方开源框架；out文件夹内存放编译后文件以及工程打包后的jar文件。

## IDE内测试

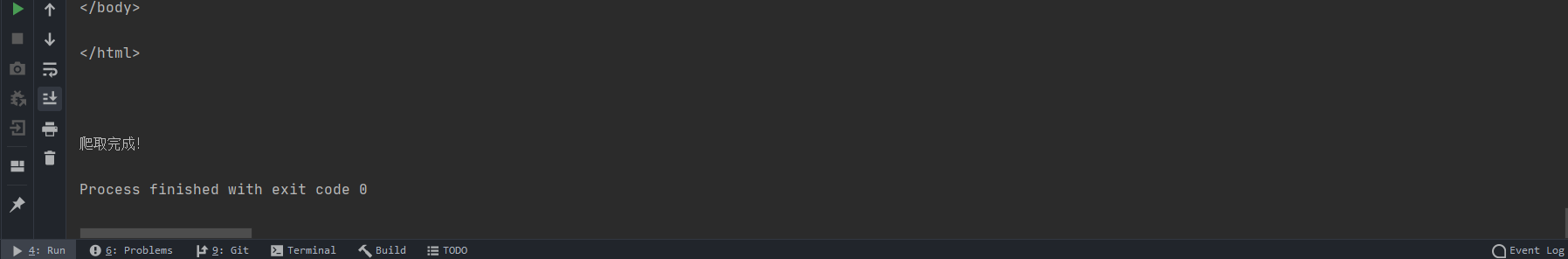
测试开始前在桌面新建文件夹命名为spidercontent，存放写入的txt文档。

①使用idea打开工程文件MySpider；



②点击RUN键或使用快捷键shift+F10，程序运行；

③程序运行结束，运行窗口显示“爬取完成！”；



④打开桌面spidercontent文件夹，文件夹内存在两个txt文档：book.txt、movie.txt。







## （二）jar运行测试

为了方便测试和使用，本工程已被打包成jar文件，爬虫程序运行无需使用特定IDE，可以在windows平台上使用cmd命令运行。Jar文件存放于MySpider\out\artifacts。

测试开始前在桌面新建文件夹命名为spidercontent，存放写入的txt文档。

①在jar文件目录下打开Windows power shell；

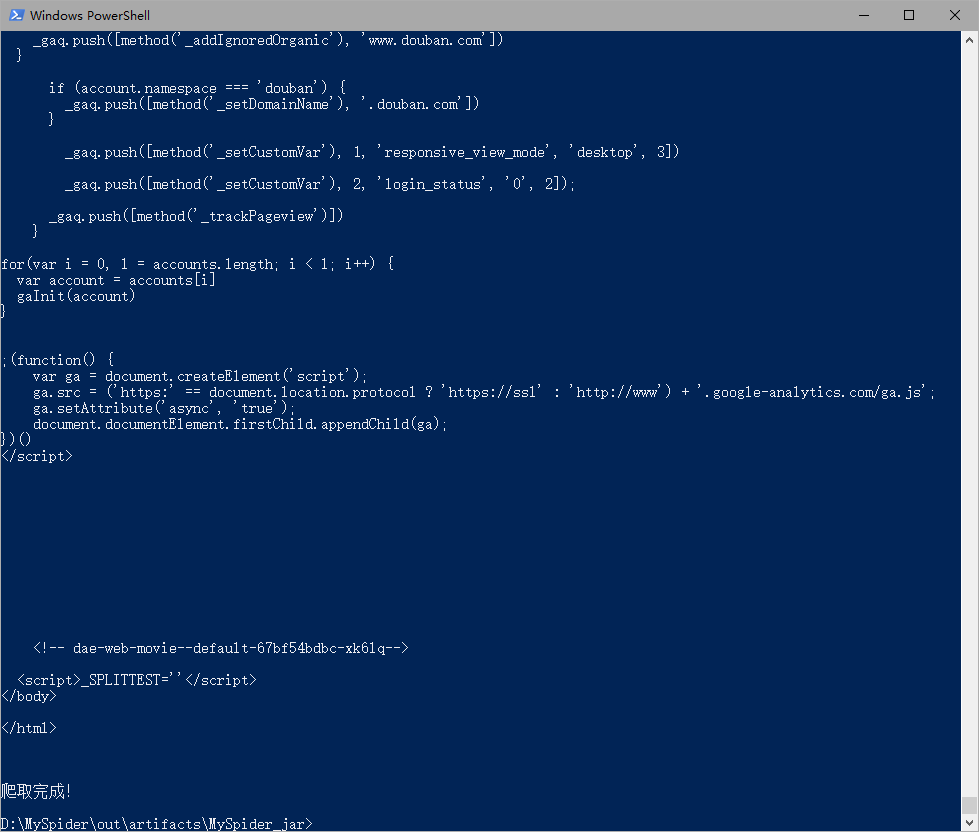


②输入命令：cmd；



③输入jar运行命令：java -Dfile.encoding=utf-8 -jar MySpider.jar；

④运行结束Windows power shell窗口显示“爬取完成！”；



⑤打开桌面spidercontent文件夹，文件夹内存在两个txt文档：book.txt、movie.txt。







# 五、总结

通过此次实现了网络项目，掌握了Java的基本语法，掌握了开源库httpclient、Jsoup的基本应用方法。了解了爬虫的基本原理，能够独立完成一个爬虫项目，能够根据所需信息选择性地抓取文本信息。

由于时间和精力上的限制，没能更进一步实现与数据库的连接以及图形化控制界面。

在此次项目实现的过程中，能够快速地了解所需要的技术，并快速地学习技术，学习实例，并能通过使用来掌握相关技术知识。