专业：
姓名：商凯
学号：3180102259
日期：
地点：

浙江大学实验报告

课程名称：java 应用技术基础　　指导老师：　　成绩：
实验名称：　　实验类型：　　同组学生姓名：

Java 网络爬虫————简单爬取一篇读过的小说

一、实验原理及设计思路

结合课上老师所讲内容以及网上搜索到的资料，决定利用 HttpClients + Jsoup 工具来完成本次网络小说爬取工作。

自己初步是想直接利用 java 本身自带的 net 来实现的，但发现后面数据解析比较麻烦，直接使用现成的工具包更为方便。

实验步骤：

1 在 IDEA 上建立一个 Maven 工程，可以直接在 pom.xml 配置文件中导入 HttpClients 和 Jsoup 的 jar 包

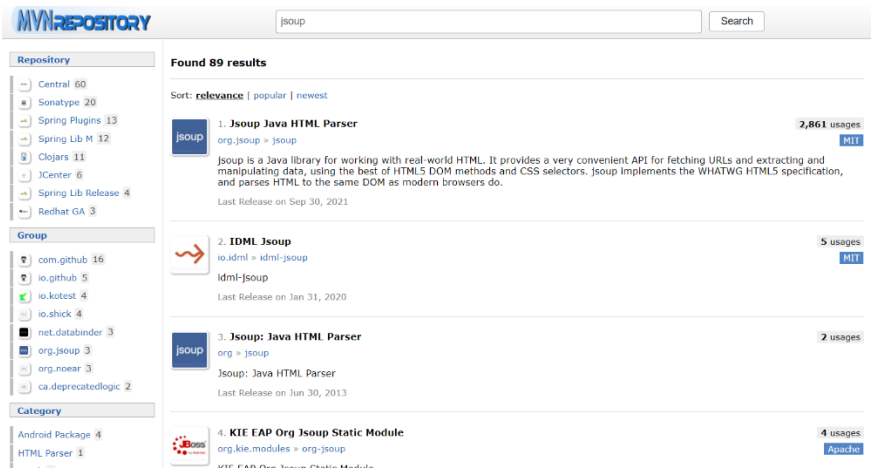
2 利用 HttpClients 工具，发送请求，类似模仿人的点击游览，可以获取页面数据

3 利用 Jsoup 工具从宽泛的页面数据中来解析我们所需要的部分

4 将解析到的数据下载到我们的 txt 文件中

二、使用方法及运行实例

1. 添加相应的 jar 包



2 使用 HttpClients 工具 jar 包模仿人访问页面

3.使用 Jsoup 解析页面

Jsoup 在学习前端的稍微一起学习了一点可以很方便来解析页面的数据，可以轻松将我们所需要的小说和标题资源存储到 String 中

4．将读取到的数据写入 novel.txt 文件中

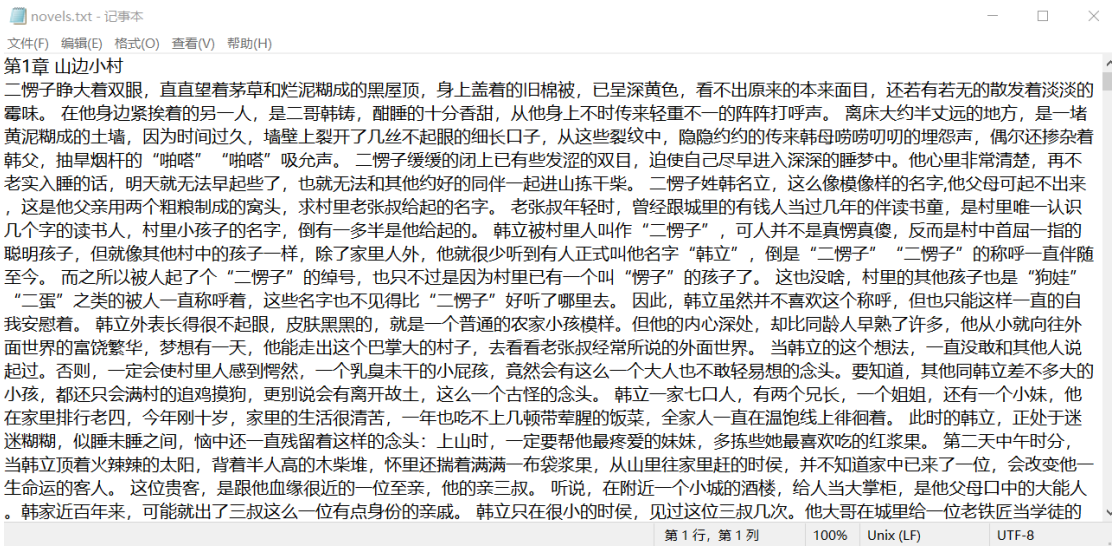
```
// 写入text文件中
FileWriter fileWriter = new FileWriter( fileName: "C:\\Users\\shangkai\\Deskt
fileWriter.write(title);
fileWriter.write( str: "\n");
fileWriter.write(mainText);
fileWriter.write( str: "\n\n"); // 在章和标题间留一些空格便于阅读

fileWriter.close();
```

完成

为正确判别我们的小说是否下载成功，我们要在最后 idea 中打出每一章的标题以及每一章是否成功存入 txt 文件中

这里鉴于该本小说章节过多，我们就只存取前 1000 章，可以发现每一章都成功答应出来



Txt 中内容也没有问题，能够正常阅读。

三、改进与不足之处

为了便于阅读小说的话，应该每 50 章创建一个 txt 文件，将所有章节小说存取在一个 txt 文件中，不便于定位上次阅读地点。

我们应该也可以直接在 html 页面中解析出下一章内容的 uri 地址

```
String uri = "https://m.soshuw.com/JueMeiZongCaiDiTieShenKuangYi/"+(856822 + i)+".html";// 需
```

而不是通过找规律直接输入下一章的 uri 地址。

本次爬取小说步骤比较简单，如果是复杂功能则需要分模块编写。