

浙江大学实验报告

专业：_____

姓名：商凯_____

学号：3180102259_____

日期：_____

地点：_____

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 HttpClient + 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 HttpClient 和 Jsoup 的 jar 包

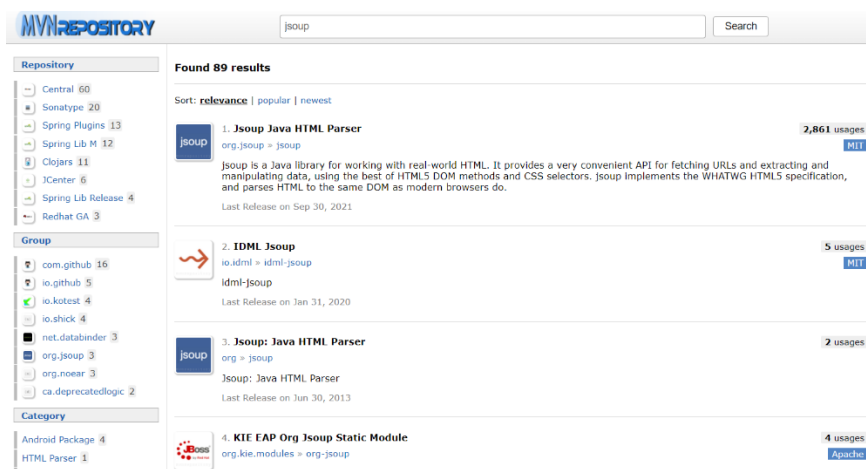
2 利用 HttpClient 工具，发送请求，类似模仿人的点击游览，可以获取页面数据

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 HttpClient 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

```

<!--
<ol class="headline">第1章 山边小村</ol>
<div class="pager"><a href="http://m.soshow.org/junWeiZongCaiDiTieShenKuangYi/" target="_blank">上一章</a> <a href="/book/junWeiZongCaiDiTieShenKuangYi.html">书 页</a> <a href="/junWeiZongCaiDiTieShenKuangYi/856823.html">下一章</a>
<div class="content">
<p>二愣子静坐着发呆，直直望着草车和思绪模糊的果密特，身上盖着的旧棉裤，已呈深黄色，看不出原来的本来面目，还若有若无的散发霉淡淡的霉味。</p>
<p>在他身边张罗着另一人，是二愣子韩铁，酣睡的十分香甜，从他身上不时传来轻重不一的阵阵打呼声。</p>
<p>离床大约半丈远的地方，是一堵黄泥糊成的土墙，因为时间久远，墙壁上爬了几根不起眼的细长口子，从这些裂缝中，隐隐约约的传来韩铁母唠叨的埋怨声，偶尔还掺杂着韩铁，抽旱烟杆的“吧嗒”“吧嗒”吸允声。</p>
<p>二愣子烦躁的闭上已有些发涩的双眼，即便自己已经进入了深深的睡梦中，他心还是异常清楚，再不能安心入睡的话，明天就无法早起赶了，也就无法和其他约好的同伴一起进山弄干柴。</p>
<p>二愣子姓韩名立，这么像稀稀的名字，他父母叫不起来，这是韩父拿用两个粗糙制成的夹头，手心里老张的给起的名字。</p>
<p>若张姓年轻时候，曾跟随城里的有钱人当过几年的传读书童，是村里唯一认识几个字的读书人，村里小孩子的名字，倒有一多半是他给起的。</p>
<p>韩立被村里人叫作“二愣子”，可不是说他真傻，而是村中很怪一指的傻愣愣子，但就像其他村中的孩子一样，除了家里人外，他就很少听到有人正式叫他名字“韩立”，倒是“二愣子”“二愣子”的称呼一直伴随至今。</p>
<p>南之所以被人起了个“二愣子”的绰号，也只是因为村里已有一个叫“愣子”的孩子。</p>
<p>这也不奇怪，村里的其他孩子也是“陈姓”，“吕家”之类的被人一直称呼着，这些名字也不见得比“二愣子”好听了哪里去。</p>
<p>因此，韩立虽然并不喜欢这个称呼，但也只能这样一直的自欺欺人罢了。</p>
<p>韩立外表长得瘦不起，皮肤黑黑的，就是个普通的农家小孩模样，但他的内心深处，却比同龄人早熟了许多，他从小就向往外面世界的繁华繁华，梦想有一天，他能走出这个巴掌大的村子，去看看老张叔经常所说的外面世界。</p>
<p>当韩立的这个想法，一直说给别人听起时，否则，一定会使村里人感到愕然，一个乳臭未干的小孩，竟然会有这么一个大人也不忍轻信的念头，要知道，其他同韩立是不大的小孩，都还只会调皮的追鸡摸狗，更别说会有离开故土，这么一个古怪的念头。</p>
<p>韩立一家四口人，有两个兄长，一个姐姐，还有一个小孩，他比张叔还小，今年刚十岁，家里的生活很清苦，一年也吃不上几顿荤腥的饭菜，全家人一直在温饱线上得苦苦。</p>
<p>此刻的韩立，正处于迷迷糊糊，睡梦未醒之间，脑中还一直浮现着这样的念头，上山时，一定要帮他最亲爱的妹妹，多弄些他最喜欢的红苕果。</p>
<p>第二天中午时分，当韩立顶着火辣辣的太阳，背着半人高的木柴堆，怀里还揣着满满一布袋苕果，从山里往家里赶的时候，并不知道家中已养了一位，会改变他一生命运的客人。</p>
<p>这些声音，是随着血腥味靠近一位至亲，他的母亲三娘。</p>
<p>听说，在附近一个小镇的酒楼，给人当大掌柜，是他父母口中能人，韩家近百年来，可能就出了三叔这么一位有点身份的亲戚。</p>
<p>韩立是在很小的时候，见过这位三叔几次，他大曾在家里给一位老铁匠当学徒的工作，就是这位三叔给介绍的，这位三叔还经常托人给他父母捎带一些吃的用的东西，因此韩立对这位三叔的印象也很好，知道父母虽然嘴上不说，心里也是很想念。</p>
<p>大掌柜是家里人的亲戚，听说当铁匠的活，不能随便干，一个县还有二十多个铁匠，等到正式出师给人家用时，挣的钱可就更多了。</p>
<p>每当父母一提起大掌柜，就神采飞扬，像换了一个人一样，韩立年龄虽小，也羡慕不已，心目最好的工作也早早就有了，就是给小镇里的哪位手艺师傅当学徒，做铁匠活，从此变成手艺吃饭的体面人。</p>
<p>所以当年韩立见到穿着身崭新的蓝布褂子衣服，胖胖的圆脸，留着一撮小胡子的三叔时，心里兴奋极了。</p>
<p>把木柴在里后放好，便回到屋里睡熟的三叔见了小立，高兴的叫了声，“三叔好”，就老实实在的坐在一边，听父母同三叔聊天。</p>
<p>三叔笑眯眯的望着韩立，打量着他一番，嘴里夸了他几句“听话”“懂事”之类的话，然后就转过头，和他父母说起这次的来意。</p>
<p>韩立虽然年龄尚小，不能完全听懂三叔的话，但也听明白了大概的意思。</p>
<p>原来三叔工作的地方，是一个叫“七玄门”的江湖门派所有，这个门派有外门和内门之分，而前不久，三叔才正式成为了这个门派的外门弟子，能够推荐7岁到12岁的孩童去参加七玄门招收内门弟子的考验。</p>
<p>五年一次的“七玄门”招收内门弟子测验，上个月就开始了，这位有着几分精明助自己尚无子女的二叔，自然想到了远亲的韩立。</p>
<p>一向忠实的韩父，听到“江湖”“门派”之类的话，心里有些犹豫不决拿不定主意，便一把拿起旱烟杆，“吧嗒”“吧嗒”的狠狠抽了几口，就坐在那里，一声不吭。</p>
<p>在三叔嘴里，“七玄门”自然是为数不多的门派，第一二的大门派。</p>
<p>只要成为内门弟子，不但以后可以免费习武吃喝不愁，每月还能有一两多的散银子零花，而且参加测验的人，即使未能入选也有机会成为像三叔一样的外门人员，专门替“七玄门”打理门外的生意。</p>
<p>当然到有可能每月有一两银子可拿，还有机会成为和三叔一样的体面人，韩父经手拿定了主意，答应了下来。</p>
<p>三叔见到韩父答应了，心里很高兴，只留了几两银子，第二天月后来带韩立走，在这期间给韩立多准备些好吃的，给他补补身子，好应付考验，随后三叔和韩父打声招呼，摸了摸韩立的头，出门回城了。</p>
<p>韩立虽然不全明白三叔所说的话，但可以进能进大钱还是明白的。</p>
<p>一直以来的愿望，眼看就有可能实现，他一连好几个晚上兴奋的睡不着觉。</p>
<p>三叔在一个多月后，很时的到来中，韩家立走了，他告诉韩父及复嘱咐韩立，做人要老实，遇事要忍让，别和其他人起争执，而韩母则要他多注意身体，要吃好睡好。</p>
<p>在马车上一路，随着父母渐渐远去的身影，韩立咬紧了嘴唇，强忍着不让自己眼眶中的泪珠流出来。</p>
<p>他虽然从小就比其他孩子成熟的多，但毕竟还是个十岁的小孩，第一次出远门让他心里有点伤感和彷徨，他年幼的心里暗暗下定了决心，等挣到了大钱就马上赶回来，和父母再也不分开。</p>
<p>韩立从那时起，此次去江湖的多少对他已失去了意义，他毅然踏上了一条与众不同的修仙之路。</p>
<p>喜欢凡人修仙传免费阅读全文请大家收藏：(m.soshow.com)凡人修仙传免费阅读全文电子书阅读网更新速度更快。</p>
</div>
<div class="pager"><a href="/junWeiZongCaiDiTieShenKuangYi/">上一章</a> <a href="/junWeiZongCaiDiTieShenKuangYi/">目录</a> <a href="https://m.soshow.org/junWeiZongCaiDiTieShenKuangYi/856823.html" target="_blank">下一章</a> <a id=
```

4. 将读取到的数据写入 novel.txt 文件中

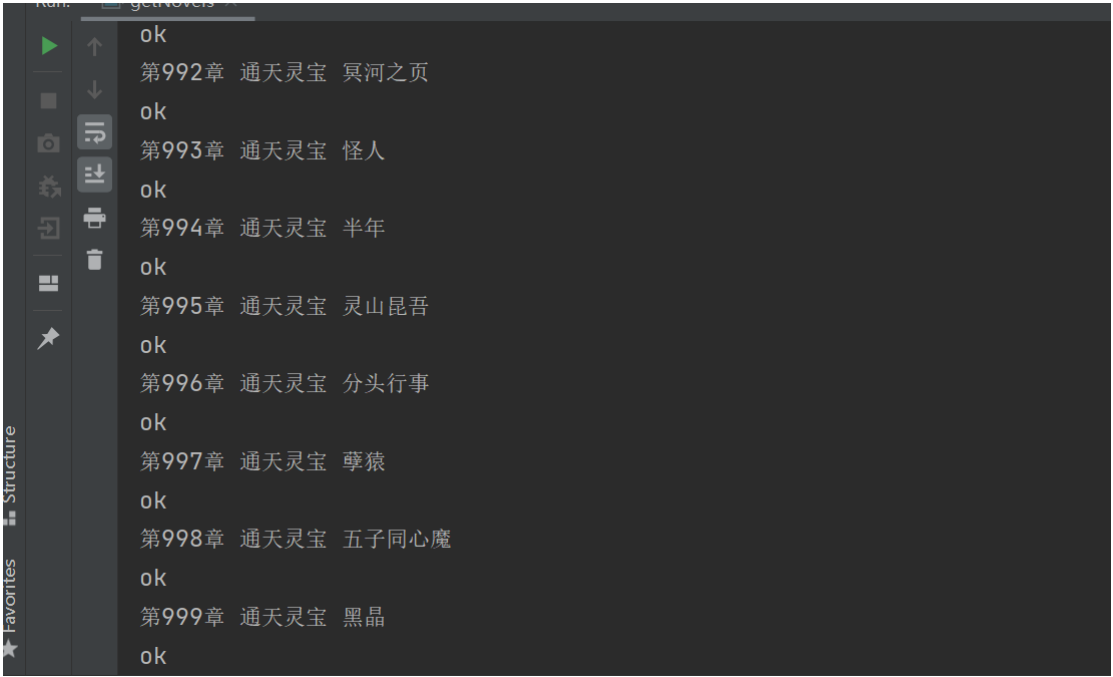
```

// 写入text文件中
FileWriter fileWriter = new FileWriter( fileName: "C:\\Users\\shangkai\\Deskto
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

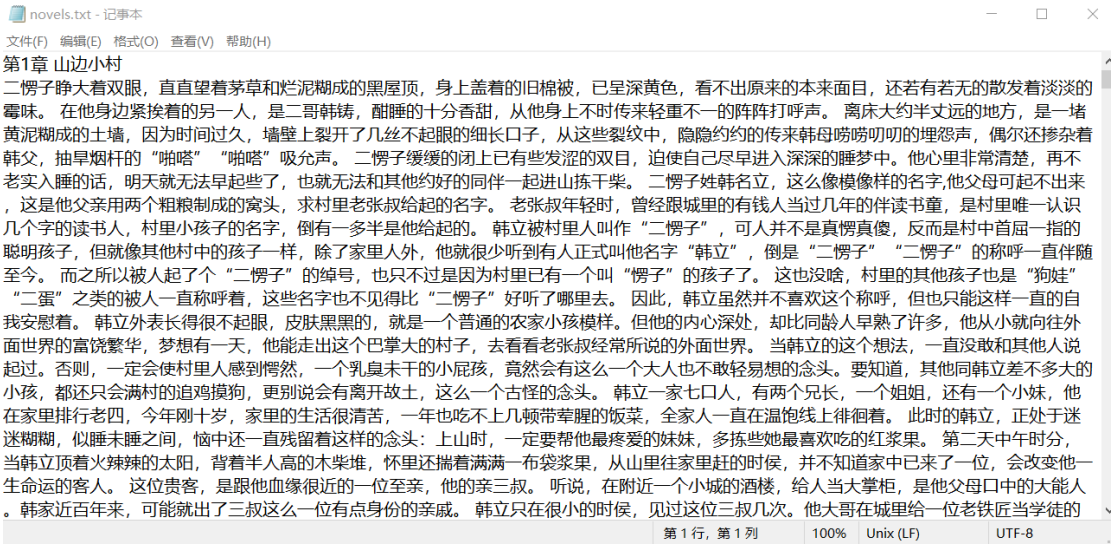
fileWriter.close();
```

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"; // 需要
```

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。