# 武汉大学国际软件学院 实验报告

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 课程名称 | 面向对象程序设计实验 | | | | | | | | | |
| 专业年级 | 2015 | | | | | | | | | |
| 姓名 | 马永成 | | | | | | | | | |
| 学号 | 2015302580305 | | | | | | | | | |
| 协作者 | 无 | | | | | | | | | |
| 实验学期 | 2016 | | | 学年 | 第一 | | | | 学期 | |
| 课堂时数 |  | | | 课外时数 | | |  | | | |
| 填写时间 | 2017 | 年 | 1 | | | 月 | | 1 | | 日 |

|  |  |  |  |
| --- | --- | --- | --- |
| 实验概述 | | | |
| 【实验项目名称】 | | | |
| Implementing a Bilibili video webcrawler | | | |
| 【实验目的】 | | | |
| 1. Understand HTTP requests, responses, and redirections. 2. Use apache-httpclient to make up requests and decompose responses. 3. Use jsoup to traverse between and inside HTML tags. 4. Know SQL grouping and sorting. 5. Use Maven to manage project dependencies. | | | |
| 【实验环境（使用的软件）】 | | | |
| Fiddler，IntelliJ IDEA | | | |
| 【参考资料】 | | | |
|  | | | |
| 实验内容 | | | |
| 【实验方案设计】 | | | |
| 1. 爬取网页信息   打开Fiddler,并打开[www.bilibili.com](http://www.bilibili.com)，点击任一版块，爬取信息    找到视频列表所对应的的包，是一个json文件，找到他的Headers,用于模拟连接。    之后可以看到这个json对应的内容，包括视频的av号，标题，作者，所属版块号，版块名称，硬币数，收藏数等，弄清楚这个json对象的格式，用于使用java HttpClient 模拟访问时解析json对象。    2.java爬取网页信息  新建项目homework\_5，使用maven导入Apache的HttpClient包，建立项目依赖。  根据助教的教程，使用HttpClient模拟浏览器发送get请求，根据之前获取到的Headers建立模拟请求。网页响应后可以使用HttpEntity得到一串字符串，这个字符串中包含我们需要解析的json对象。    在这里我当时没有找到解析json的JSONObject方法，所以这里我使用了Google的Gson包。这个包虽然可以解析json，但只能解析标准格式的json对象，并且需要一个实体类来映射解析结果，而我们获取到的不是标准的json对象，在这里我使用了切割字符串的方法，将从网页上爬取到的字符串切割成一个个json对象，之后使用Gson进行映射解析。      3、数据存入数据库  解析完之后，将数据存入数据库。数据库中包括版块号tid，版块名tname，av号，标题title，作者author，硬币数coin，收藏数favorite。  在这里我在存入数据库之前没有筛选每个版块的前三名，而是将爬取到的数据全部存入数据库，在数据库中根据收藏数进行分组排序，选出每个版块的前三名。    数据库中存了78万条记录。      4、下载视频  得到视频的av号后，就可以从网页上下载视频了。  这里使用[www.ibilibili.com](http://www.ibilibili.com)/video/av.....地址，根据av号查找到ibilibili的网页，使用jsoup解析该网页，得到下载链接。    此为下载链接，找到这个链接之后，使用java的io包从对应的链接下载视频。    查看本地，已下载完成。 | | | |
| 【结果】   1. 使用HttpClient爬取网页成功，每个版块爬取了1000页，数据库中一共存储了78万条记录。 2. 从数据库中排序选出前三名，根据视频的av号，使用jsoup获取下载链接成功，调用java的IO可以下载到本地。 3. 用maven添加依赖也基本掌握。 | | | |
|  | | | |
| 【小结】 | | | |
| 这学期的JAVA实验真的收获很大。  这次实验简单了解了网页爬虫的原理，也算是入了爬虫的门，虽然只是很简单的基础。  之前实验中也学习了数据库的基础知识，也算是有了一点数据库的使用能力，在这次实验中也有相应的收获，当数据库中的数据非常多时，怎样可以提高查询速度也是需要思考的问题，我自己也算是想到了一些方法。  总之，这次的实验收获很大。 | | | |
| 指导教师评语及成绩 | | | |
| 【评语】 | | | |
|  | | | |
| 成绩： |  | 指导教师签名： |  |
| 批阅日期： | | | |