# 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（Java） | | | | | | | | | |
| 专业年级 | 软件工程大二 | | | | | | | | | |
| 姓名 | 刘璇 | | | | | | | | | |
| 学号 | 2015302580282 | | | | | | | | | |
| 协作者 | 无 | | | | | | | | | |
| 实验学期 | 2016-2017 | | | 学年 | 第一 | | | | 学期 | |
| 课堂时数 |  | | | 课外时数 | | |  | | | |
| 填写时间 | 2017 | 年 | 1 | | | 月 | | 4 | | 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）】 | | | |
| IntelliJ IDEA Telerik Fiddler Web Debugger  Maven Bilibili | | | |
| 【参考资料】 | | | |
| The instruction which is given by the assistant, Baidu , Maven repository | | | |
| 实验内容 | | | |
| 【实验方案设计】  First clear the solving ideas of this project:  Using Fiddler to know the constructor of the HTTP and the important things we should Crawler down→Create a database→Crawler the data from [www.bilibili.com →insert](http://www.bilibili.com→insert) the data into database→use SQL statement to get top 3→put top 3’s URL to array→Download the URL from array.  Understand the system framework of this project:   1. Display Layer:The videos I get.The count of data. 2. Logical Layer:Crawler.URL.DataBase. 3. Data-access Layer:Realize access,connect,insert,select,sort to database.   Start Project: | | | |
| 1. ***Analyze important requests/responses using a web debugger, i.e. an HTTP message sniffer. You should focus on av numbers, authors, categories, number of coins, number of favourites, etc.***So I use Fiddler according to assistant’s video,and register on [www.bilibili.com](http://www.bilibili.com) to see the constructor or format of HTTP message,just like this.   And I find in stat, it has coin,favorite,view,share,reply’s number.  In tags,it has tid,title and tname.  And I get av numbers from aid,author from owner’s name.    And when we Crawler it,we just get av numbers, authors, categories, number of coins, number of favorites, etc.  Then code it like this.IMG_256   1. ***Store each av number along with its important attributes in your favorite database.***   So I need to create a database and a database table.And get the table in this:create table yurui(  -> tid int(11) not null,favorite int(11) not null,coin int(11) not null,  author varchar(100) not null,tname varchar(100)not null,av int(11) not null);  And insert the av number along with its tid,favorite,coin,author and so on.just like this.  IMG_256  And I should Crawler lots of data to store it into the database.  So I get traverse around of tid from 1 to 200,and page from 1 to 1000 to get enough data.  6`7U_CFKB%64CSW66J{K%QE  And the point is IMG_256  The tid and the pg.   1. ***Find out the top 3 favorite videos for each category.***   I need to use SQL’s statement to get the top 3 LikeZTPI5EQ}_DAHX%@`UB0OO23   1. ***Get the URLs of the videos in step 3 by using apache-httpclient (perhaps jsoup is needed to parse html pages). Then, download the URLs via Java IO. You can simplify this step with www.ibilibili.com .***   The last step is downloading the top 3.  First, I need to get the way to download by URL.Like this.%1%EOU{@)P3JB~)}{0`ND[N  Then ,I should get the URL from database by this way.ZGRCT5{2OOEJY@WP]C6Z)3H  OK,This is all the main steps to finish the last homework. | | | |
| 【结果】 | | | |
| 1. I get lots of database,like this.   ]AS7~A}{D3JZ)@YK4J%D_08   1. I can store the data into database correctly,like this.OTHO_@9@(OGZ)~`$A()0WZO 2. I can get top 3 correctly,like this   .Q~AV_{7X4LVT7T7@DU$HG81  (I can get top 3 by favorite with same tid or tname.)   1. And print it out in the console.)3UNN7608]7343VH@D49WVB 2. Every JARs you need should be configured in pom.xml, which will appear in your project root directory after enabling Maven support.   First  ,4~0X@9EWI07X3$K1ZXQ{C[V.  Second,)OPR3J[YS{Q~68`JW63$3WF.  Third  6$L[@95(Y`(E1SN5915M9(H.  Last  .  Just like O]S%CF}F[~8FGY0@H%3332G   1. Download the videos of top 3 of different tname.   And the videos is OK.This video is the first of 宅舞  WN0P_OTM2K}9716[)06W[90  Of course,I can get many videos. | | | |
| 【小结】 | | | |
| It’s a big Crawler project ,and we should use Fiddler to get the constructor of HTML,so we can get information correctly from it.And we need to know how to create a database and the table in it,and the knowledge of inserting data and get the top3 of it.  The end ,we should via IO.JAVA to download videos by URL.  In this project,I also get the way how to use Maven to put the jar more convenient and safe.  Of course,I have some problems during I finish this project.First,when I insert the data into database,it just present garbage character,and use Internet to select the way which is modifying the language of SQL,from GBK to utf8.And when Crawler the Bilibili ,it always stop because the Internet is not very well,so I should put an stop timer to let it go on.  And in this project,we obviously know more about Bilibili and statistics the number of favorite,coins,reply and so on.Using this project,we can downloading the most popular videos. | | | |
| 指导教师评语及成绩 | | | |
| 【评语】 | | | |
|  | | | |
| 成绩： |  | 指导教师签名： |  |
| 批阅日期： | | | |