

**信 息 工 程 系**

**2018年暑期实训报告**

|  |  |
| --- | --- |
| **班 级：** | 计算机科学与技术2班 |
| **学 号：** | 20170180 |
| **姓 名：** | 文岚清 |
| **指导老师：** | 何青 |

二零一八年七月

**目 录**

[一、实训任务书 1](#_Toc524516833)

[二、实训日志 2](#_Toc524516834)-7

[三、项目报告 8-](#_Toc524516835)10

[四、实训心得 11](#_Toc524516836)

[五、实训评价 12](#_Toc524516837)

# 

# 一、实训任务书

**1、实训目的**

本学年我们学习了Python语言程序开发课程，学习、掌握了Python语言的基础语法。通过学习我们了解到Python语言常用的场景包括：科学计算、web项目开发、人工智能、网络爬虫。为了巩固Python基础知识，扩展Python编程技能，提高学生发现问题、解决问题、独立思考的能力，以Python网络爬虫的设计与实现为题开展为期10天的暑期实训，介绍网络爬虫的原理与实现方法、技术，最终由学生独立设计实现自己的爬虫软件。

目标：

1. 培养学生运用所学课程Python语言程序设计的理论知识和技能，分析解决计算机实际应用中的问题的能力；
2. 培养学生在Python语言程序设计的基础上，开发和实现网络爬虫项目的能力；
3. 培养学生调查研究、查阅技术文献、资料手册以及编写技术文献的能力。

**2、实训时间**

2018年7月5日-2018年7月14日

上午：8:00 - 11:30

下午：14:00 - 17:30

**3、实训地点**

计算机专业实训室1,2,3

**4、实训主要内容**

|  |  |
| --- | --- |
| 7月5日-6日 | Python面向对象编程 |
| 7月7日 | Requests模拟网络请求 |
| 7月8日 | Beautifulsoup解析 |
| 7月9日 | Python操作数据库 |
| 7月10日 | 爬虫项目实现 |
| 7月11日-7月13日 | 小组项目实施 |
| 7月14日 | 小组项目展示 |

# 二、实训日志

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **学号** | | 20170180 | | **姓名** | 文岚清 | | **班级** | 计算机科学与技术2班 | |
| **实训题目** | | | 基于Python的网络爬虫设计与实现 | | | | | | |
| **实习天数** | | **第 1天** | | **日期** | | | **2018.7.5** | | |
| **日报内容** | | | | | | | | | |
| **工作**  **记录（**  **完成的任务、遇到的问题，解决的办法）** | **收获：**  **1.在老师的帮助下，装了requests库，并学习了requests库的7个主要方法：requests.requst(),requests.get(),requests.post(),requests.head(),requests.put(),requests.patch(),requests.delete()**  **2.学会了http响应正文，以及用用二进制、json响应正文**  **3.理解了requests异常意思**  **4.运用这些知识，并在爬虫框架的学习之上，爬取了百度搜素首页的html代码**  **5.ip138案例爬取百度代码**  **6.用requests库模拟登陆，了解登陆机制**  **困难：**   1. **在视频的基础上学习比较吃力，要照着视频的代码写。** 2. **关于会话对象不能理解** 3. **写更复杂的POST请求比较吃力** | | | | | | | | |
| 评定成绩 |  | | | | | 指导老师签字 | | |  |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **学号** | | 20170180 | | **姓名** | 文岚清 | | **班级** | 计算机科学与技术2班 | |
| **实训题目** | | | 基于Python的网络爬虫设计与实现 | | | | | | |
| **实习天数** | | **第 2天** | | **日期** | | | **2018.7.6** | | |
| **日报内容** | | | | | | | | | |
| **工作**  **记录（**  **完成的任务、遇到的问题，解决的办法）** | **收获：**   1. **学习selenium了并尝试运用它进行模拟登陆** 2. **学习库，并了解BeautifulSoup解析器，以及BeautifulSoup类的基础元素：tag、name、attributes、navigablestring、comment** 3. **学会了获取标签的方法：from bs4 import BeautifulSoup** 4. **安装driver程序**   **困难：**   1. **自己学习进度跟不上视频讲解的速度** 2. **运用section模拟爬虫非常困难** 3. **不熟悉driver相关程序** | | | | | | | | |
| 评定成绩 |  | | | | | 指导老师签字 | | |  |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **学号** | | 20170180 | | **姓名** | 文岚清 | | **班级** | 计算机科学与技术2班 | |
| **实训题目** | | | 基于Python的网络爬虫设计与实现 | | | | | | |
| **实习天数** | | **第 3天** | | **日期** | | | **2018.7.7** | | |
| **日报内容** | | | | | | | | | |
| **工作**  **记录（**  **完成的任务、遇到的问题，解决的办法）** | **收获：**   1. **复习并加深BeautifulSoup库的使用并找标签的名字，进行QQ空间留言板的查询** 2. **观看selenium注意事项的视频基本解决代码运行卡住的问题** 3. **进行更深度的学习，进行教务处查询的信息爬取**   **困难：**   1. **th的文本取出还是不能熟练运用** 2. **td.string的找出long结果** 3. **爬虫知识点比较熟，实际操作有困难** | | | | | | | | |
| 评定成绩 |  | | | | | 指导老师签字 | | |  |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **学号** | | 20170180 | | **姓名** | 文岚清 | | **班级** | 计算机科学与技术2班 | |
| **实训题目** | | | 基于Python的网络爬虫设计与实现 | | | | | | |
| **实习天数** | | **第 4 天** | | **日期** | | | **2018.7.8** | | |
| **日报内容** | | | | | | | | | |
| **工作**  **记录（**  **完成的任务、遇到的问题，解决的办法）** | **收获：**   1. **开始接触数据库，并选择了mysql作为实行的数据库，并操作redis。** 2. **了解connection对象、cursor对象和事务概念进行介绍** 3. **用python和数据库连接，尝试进行爬取。**     **困难；**   1. **刚开始接触数据库很陌生，很多方面不了解，基础知识不知道** 2. **跟着视频做很迷茫，很多专业词，不理解，必须去百度** 3. **学习python操作数据库过程缓慢，今天学习完了还有很多不懂得** | | | | | | | | |
| 评定成绩 |  | | | | | 指导老师签字 | | |  |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **学号** | | 20170180 | | **姓名** | 文岚清 | | **班级** | 计算机科学与技术2班 | |
| **实训题目** | | | 基于Python的网络爬虫设计与实现 | | | | | | |
| **实习天数** | | **第 5 天** | | **日期** | | | **2018.7.6** | | |
| **日报内容** | | | | | | | | | |
| **工作**  **记录（**  **完成的任务、遇到的问题，解决的办法）** | **收获：**   1. **终于结合数据库以及python进行爬取，可以开始做一个项目了，开始做程序爬虫框架，抓取大多数情况属于get请求，即直接从对方服务器上获取数据，** 2. **进行多线程的爬取** 3. **爬取也在当中找出数据库与python连接出现问题解决**   **困难：**   1. **制作项目时不知道如何下手，项目时什么也比较难想** 2. **关于多线程并发抓取出现许多问题待解决** | | | | | | | | |
| 评定成绩 |  | | | | | 指导老师签字 | | |  |

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **学号** | | 20170180 | | **姓名** | 文岚清 | | **班级** | 计算机科学与技术2班 | |
| **实训题目** | | | 基于Python的网络爬虫设计与实现 | | | | | | |
| **实习天数** | | **第 6 天** | | **日期** | | | **2018.7.7** | | |
| **日报内容** | | | | | | | | | |
| **工作**  **记录（**  **完成的任务、遇到的问题，解决的办法）** | **我组准备爬取京东用晒单图片，其中遇到多次的京东的反爬取，导致我们无法爬取到图片，我们于是用伪登陆进行登陆爬取，登陆用户账户密码，输入关键词进行关键词晒单图片爬取，可是图片还是无法爬取，可见京东对用户的数据还是非常重视** | | | | | | | | |
| 评定成绩 |  | | | | | 指导老师签字 | | |  |

# 三、项目报告

|  |  |
| --- | --- |
| **项目名称** | 爬取京东物品信息 |
| **项目组成员** | 20172400蒋光超，20170180文岚清 |
| **项目背景** | |
| 爬取这个是因为我们现在生活中经常网购，我们又经常有选择恐惧症，所以爬取这个可以给我们提供一些参考数据列如：该产品是否好用，是否符合商家描述，价格是否合理等；我们可以根据这些来决定是否加入购物车；方便人们购物的同时又可以了解物品。我们比别人加了一些小细节列如：去除广告和计算晒图数等。 | |
| **作品描述** | |
| 实现了提供给客户更好的购物信息。帮助客户进行筛选和查找物品，运用了requests和Scrapy库和一些函数和图片保存和下载和商品连接函数等等。 | |
| **作品截图** | |
|  | |
| **项目总结** | |
| 我在项目中去爬取京东项目中商品的url链接，访问每个类别的url就可以获取得到产品列表，找到产品的URL，进入详情页面抓取产品的详情。产品的基本信息在详情页面基本可以获取，但是有些信息，比如：价格、优惠政策等信息，是需要动态获取的。在每个产品的详情页面都可以直接找到店铺id和vender id。商品的url需要动态获取。 | |

# 四、实训心得

|  |
| --- |
| 通过这次的python爬虫实训的学习，让我有增长了一门学问，并且意识到python爬虫有好处也有局限性。先从局限性来说吧，1.访问局限性。比如我访问一个网站，要获取网站的连接。如果我们抓取的数据无法访问其类容，或者无法转到下一个页面，爬虫无法解决这类问题。 2.针对性太专一。爬虫去爬取下载网页可以的到一个统一的模板，但是关于网页的解析是不行的，网站的格式及模板却不是统一的。解析操作是在我们下载的网页中提取我们需要的内容，网站的设计不同，提取的内容不同，相关的操作也存在着很大的差异。3. .时效性：作为一个网络爬虫开发者，我们应该随时关注我们正在爬取的网站的更新动态，一旦网页的结构发生了变化，我们的解析规则可能也需要变化，及时更新爬虫的解析规则，以确保爬虫的正常运行。  随着不断的学习使用，渐渐的接受了python的编码方式，python简单的语法的确带来的极大的便利。我们们不需要去纠结变量的类型问题，了解了python的基本4大组件之后就可以开始我们自己的编程之路。特殊的语法规则大大的减少了代码的数量，严格的缩进要求使得代码的格式更加统一，交接更加便捷。针对于轻量级的任务而言，代码透明影响不大，网络上提供了大量的开源包，使得编程者的工作量骤减。我认为好的爬虫的应该具备以下的特点：1.友好性。爬虫开发者与网站开发者有永无止境的“战争”，爬虫开发者应当充分考虑网站服务器的压力，有好的去反问网站。2. .高性能：每秒钟能够抓取的网页数量是衡量爬虫性能的重要指标。这也是爬虫开发者需要好好考虑的一个问题。如和在保证友好性的同时尽可能的提高自身的效率，还需集体问题具体分析。3.拓展性。.扩展性：爬虫需要抓取的网页数最巨大，即使单个爬虫的性能很高，要将所有网页都下载到本地，仍然需要相当长的时间周期，为了能够尽可能缩短抓取周期，爬虫系统应该有很好的可扩展性，即很容易通过增加抓取服务器和爬虫数量来达到此目的。  学生签字：  日期： |

# 五、实训评价

|  |
| --- |
| **实训成绩：** |
| **教师评语：**  教师签字：  日期： |