山东大学 软件 学院

Web数据管理 课程实验报告

|  |  |  |  |
| --- | --- | --- | --- |
| 学号：202000300125 | 姓名：贾星宇 | | 班级：2020级5班 |
| 实验题目：简单爬虫 | | | |
| 实验学时：6 | | 实验日期：2023年4月16日 | |
| 实验目的：  1、选择熟悉的语言和开发环境：c语言，java语言，python语言都可以  2、选择合适的爬虫框架，合适的模块，开源工具（不能是APP）  3、选择合适的网站作为爬取目标 | | | |
| 硬件环境：  Intel(R) Core(TM) i5-10210U CPU @ 1.60GHz 2.11 GHz | | | |
| 软件环境：  Intellj IDEA 2022.2 | | | |
| 实验步骤与内容：   1. 目标网站分析：   实验一的主要内容是使用python对几个小说网站进行爬取。因此选择了三部小说对应的网站：   |  |  | | --- | --- | | 小说 | 网站 | | 《山羊不吃天堂草》 | https://www.99csw.com/book/9665/index.htm | | 《草房子》 | https://www.xyyuedu.com/etdw/caofangzi/ | | 《小王子》 | http://www.newxue.com/mingzhu/xiaowangzi/ |   可hi发现，三种网站的url都是比较规范的url，也存在编号或者拼音来指示相应的内容。   1. 使用的开发工具：   本次爬虫使用python语言，主要使用urllib库进行请求的发起，随后使用beautifulsoup对网页进行分析清理。   1. 操作过程： 2. 网页分析：三类网站的基本结构相似，因此我们选取一个进行示例分析：对于网站99csw，我们希望爬取《山羊不吃天堂草》的所有章节内容，因此我们进入到网页首页：     可以发现，我们需要的内容存在于目录中的每一章。点击每一章节，页面会编写url并跳转。对应于每一章的url存在规律：  第一章：https://www.99csw.com/book/9665/347118.htm  第二章：https://www.99csw.com/book/9665/347119.htm  ……  可以发现，只需要不断增加最后的编号即可实现对每一个章节页面的选择。   1. 编写代码：根据对网页元素的分析，我们使用beautifulsoup进行网页内容的提取：     # 导包,发起请求使用urllib库的request请求模块  import urllib.request  from bs4 import BeautifulSoup  for i in range(26):      # urlopen()向URL发请求,返回响应对象,注意url必须完整      index = str(191658+i)      response = urllib.request.urlopen('https://www.nunubook.com/xiandaixiaoshuo/4176/'+index+'.html')      html = response.read().decode('utf8')      soup = BeautifulSoup(html, 'html.parser')      print('-----------------------------------------------------------------------------')      fileWriteName = 'Chapter0' + str(i + 1) + '.txt'      file = open(fileWriteName, 'w', encoding='utf8')      file.write(soup.find(id='text').get\_text())      file.close()  其中，for循环用于迭代每一个网页，index表明了不同章节页面的变量，随后使用urlopen进入该页面，并选择id=text的部分，可以发现此页面已经被爬取到了重点的内容，随后使用文件的写方法写入文件，生成的文件如下：    相似的，另外两本小说的代码如下：  《草房子》：  #导包,发起请求使用urllib库的request请求模块  import urllib.request  from bs4 import BeautifulSoup  # urlopen()向URL发请求,返回响应对象,注意url必须完整  response=urllib.request.urlopen('https://www.xyyuedu.com/etdw/caofangzi/')  print(response)  #提取响应内容  html = response.read().decode('gbk')  #打印响应内容  print(html)  soup = BeautifulSoup(html,'html.parser')  print(soup.prettify())  print(soup.title)  print(soup.get\_text())  print(soup.findAll('a'))  《小王子》：  # 导包,发起请求使用urllib库的request请求模块  import urllib.request  from bs4 import BeautifulSoup  for i in range(26):      index = str(6992 - i)      response = urllib.request.urlopen('https://www.ppzuowen.com/book/xiaowangzi/' + index + '.html')      html = response.read()      soup = BeautifulSoup(html, 'html.parser')      print('-----------------------------------------------------------------------------')      fileWriteName = 'Chapter0' + str(i + 1) + '.txt'      print(soup.select("[class~=articleContent]")[0].get\_text())      file = open(fileWriteName, 'w', encoding='utf8')      file.write(str(soup.select("[class~=articleContent]")[0].get\_text()))      file.close()     1. 爬取数据：   如上所提，所有数据都以python写的形式，使用UTF-8的格式写入到了txt文件中，每一个文件代表了一章节。经过统计，数据量达到MB级别。  附件：  Lab1-data：显示了每一部小说的数据  Lab1-code：每一部小说对应的代码 | | | |
| 结论分析与体会：  爬虫就是自动获取网页内容的程序，例如搜索引擎，Google，Baidu 等，每天都运行着庞大的爬虫系统，从全世界的网站中爬虫数据，供用户检索时使用。  爬虫流程其实把网络爬虫抽象开来看，它无外乎包含如下几个步骤  模拟请求网页。模拟浏览器，打开目标网站。  获取数据。打开网站之后，就可以自动化的获取我们所需要的网站数据。  保存数据。拿到数据之后，需要持久化到本地文件或者数据库等存储设备中。 | | | |