山东大学网络空间安全学院

Python高级程序设计课程实验报告

|  |  |  |  |
| --- | --- | --- | --- |
| 学号：202100460065 | 姓名：李昕 | | 班级：21级密码2班 |
| 实验题目：实验七 urllib模块使用 | | | |
| 实验学时：2 | | 实验日期：2022/11/8 | |
| 实验目的：熟悉模块的用法，练习编写爬虫。 | | | |
| 硬件环境：  AMD Ryzen 7 5800H with Radeon Graphics 3.20 GHz  机带 RAM 16.0 GB (13.9 GB 可用) | | | |
| 软件环境：  系统：Windows11  编译器：IDLE | | | |
| 实验步骤与内容：  本次实验包括三个部分：  （1） 网页内容读取，读取指定URL的网页内容；  （2） 利用urllib模块+正则表达式爬取http://tieba.baidu.com/p/2460150866帖子里的的所有图片，保存到本地。  （3） 使用urllib模块+正则表达式，编写爬虫从http://www.cae.cn/cae/html/main/col48/column\_48\_1.html爬取中国工程院院士信息  【网页内容读取】   1. 第一题思路比较简单，urllib库里的requset.urlopen()创建请求对象 2. 使用read()方法来读取网页源代码，将其解码并打印出来即可，最后用close()关闭连接，完成操作。   **代码实现：**   1. import urllib.request 2. *#用urlopen()创建urllib连接对象* 3. fp = urllib.request.urlopen(r'http://www.baidu.com') 4. *#read()可以读取网页超文本源代码，也就是HTML代码* 5. *#参数为指定读取内容大小，无参数默认全部文本* 6. *#需要解码，否则会直接返回十六进制编码字符串* 7. print(fp.read().decode()) 8. *#关闭连接对象* 9. fp.close()   **运行结果：**    【保存图片到本地】   1. 利用urllib.request模块提供的urlretrieve()函数下载贴吧中的数据。 2. 在请求网页爬取的时候，输出的text信息中会出现抱歉，无法访问等字眼，这就是禁止爬取，需要通过**反爬机制**去解决这个问题。对headers（网页头信息）进行模拟，这时候需要使用到 urllib.request.Request 类：class urllib.request.Request(url, data=None, headers={}, origin\_req\_host=None, unverifiable=False, method=None)   3、使用re模块，构造网址的正则表达，将图片依次下载到当前文件夹中  4、编写函数打印下载进度，进度%=目前为止传递的数据块数量\*每个数据块的字节大小/远程文件的大小  **代码实现：**   1. import urllib.request 2. import re 3. *#获得目标网页全部源代码* 4. def getHtml(url): 5. *#目标头部信息* 6. headers = { 7. 'User-Agent':'Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/93.0.4577.82 Safari/537.36' 8. } 9. *#url作为Request()方法的参数，构造并返回一个符合headers要求的Request对象* 10. req = urllib.request.Request(url, headers=headers) 11. *#Request对象作为urlopen()方法的参数，发送给服务器并接收响应* 12. fp = urllib.request.urlopen(req) 13. *#读取指定网页全部源代码并解码* 14. htmlCode = fp.read().decode() 15. *#返回文本* 16. return htmlCode 18. *#获得下载地址并进行下载* 19. def getImg(htmlCode): 20. *#构建正则表达式对象* 21. reg = r'src="(.+?\.jpg)" pic\_ext' 22. imgre = re.compile(reg) 23. imglist = imgre.findall(htmlCode) *#如果正则表达式带子模式，findall返回括号中匹配的* 24. *#计数变量，用于命名* 25. x=0 26. *#图片存放的目录，默认空串为当前目录* 27. paths = "" 28. for imgurl in imglist: 29. urllib.request.urlretrieve(imgurl,'{}{}.jpg'.format(paths,x)) 30. x = x + 1 31. htmlCode = getHtml("http://tieba.baidu.com/p/2460150866") 32. getImg(htmlCode) 33. print("OVER")   运行结果：    【爬取院士信息】  1．用urllib.request.urlopen获得目标网址的全部源代码，使用正则表达式对html代码进行筛选得到院士的姓名。  2．对院士名称标签对应的网址（网页相对路径）进行记录，随后再去超链接对应的网址爬取院士的详细信息并记录在txt文件中。   1. 代码实现： import urllib.request 2. import re 3. *#写入文件函数，规定写入的格式* 4. def writer(name, filename, text): 5. with open(filename, 'a', encoding='utf-8') as f: 6. f.write(name + '\n') 7. f.writelines(text) 8. f.write('\n\n') 10. *#获得目标网页全部源代码* 11. def getHtml(url): 12. *#目标头部信息* 13. headers = { 14. 'User-Agent':'Mozilla/5.0 (Windows NT 10.0; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/93.0.4577.82 Safari/537.36' 15. } 16. *#url作为Request()方法的参数，构造并返回一个符合headers要求的Request对象* 17. req = urllib.request.Request(url, headers=headers) 18. *#Request对象作为urlopen()方法的参数，发送给服务器并接收响应* 19. fp = urllib.request.urlopen(req) 20. *#读取指定网页全部源代码并解码* 21. htmlCode = fp.read().decode() 22. *#返回文本* 23. return htmlCode 25. def getContent(html): 26. *#构建正则对象* 27. *#/s - 匹配任何空白字符，包括空格、制表符、换页符* 28. *#/S - 与\s含义相反* 29. reg = r'<a href="(.+?\.html)" target="\_blank">(.+?)</a>' 30. regintro = r'<div class="intro">([\s\S]+?)</div>' 31. regp = r'<p>([\s\S]+?)</p>' 32. htmlre = re.compile(reg) 33. htmllist = htmlre.findall(html) 34. introre = re.compile(regintro) 35. intropre = re.compile(regp) 36. *#计数变量，用于计算下载进度* 37. for htmlurl,name in htmllist: 38. *#字符串拼接出院士信息的网址* 39. htmlcontent = getHtml('http://www.cae.cn'+htmlurl) 40. intro\_withp = introre.findall(htmlcontent) 41. intro\_withoutp = intropre.findall(intro\_withp[0]) 42. *#信息* 43. txt = "" 44. for cont in intro\_withoutp: 45. txt += cont.replace('&ensp;','') 46. *#写入* 47. writer(name, '工程院士信息.txt', txt) 48. allyuanshi = len(htmllist) 49. htmlCode = getHtml("https://www.cae.cn/cae/html/main/col48/column\_48\_1.html") 50. getContent(htmlCode) 51. print("OVER")   执行结果： | | | |
| 结论分析与体会：  通过本次练习，我熟悉了解了如何爬取网页上的数据和图片，也通过网络检索学习了正则表达式的用法。 | | | |