山东大学网络空间安全学院

Python高级编程 课程实验报告

|  |  |  |  |
| --- | --- | --- | --- |
| 学号：202100450069 | 姓名：孟庆丹 | | 班级：网安21.2 |
| 实验题目： | | | |
| 实验学时：2 | | 实验日期：2022.11.15 | |
| 实验目的：熟悉urllib模块、requests模块+BeautifulSoup模块使用、Feapder框架的用法，练习编写爬虫。 | | | |
| 硬件环境：  处理器 AMD Ryzen 5 5600H with Radeon Graphics 3.30 GHz  机带 RAM 16.0 GB (13.9 GB 可用) | | | |
| 软件环境：  Windows 11 家庭中文版 21H2  Python 3.10.7 | | | |
| 实验步骤与内容：   1. 使用urllib模块或request模块读取网页内容，并利用BeautifulSoup模块进行内容解析，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。   bs解析html内容比起正则表达式算是相当舒服了，introLink = i\_bs.a['href']就可以直接获取获取院士对应的个人链接，再用introInfo = newContents.find('div', attrs={'class': 'intro'}).find\_all('p')获取院士的个人信息，最后写入本地即可。  代码实现：  import requests  from bs4 import BeautifulSoup  link = "https://www.cae.cn"  link1 = link + "/cae/html/main/col48/column\_48\_1.html"  headers = {      "User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/107.0.0.0 Safari/537.36"}  reqst = requests.get(link1, headers=headers)  contents = BeautifulSoup(reqst.text, 'lxml')  li = contents.find\_all('li', attrs={'class': 'name\_list'})  for i in li:      i\_bs = BeautifulSoup(str(i), 'lxml')      name = i\_bs.a.string      introLink = i\_bs.a['href']      newLink = link+introLink      newReqst = requests.get(newLink, headers=headers)      newContents = BeautifulSoup(newReqst.text, 'lxml')      introInfo = newContents.find('div', attrs={'class': 'intro'}).find\_all('p')      introInfo1 = str(introInfo[0].string +                       introInfo[1].string+introInfo[2].string)      print('已获取'+name+'院士信息')      f = open('./info/'+name+'.txt', 'w', encoding='utf-8')      f.write(introInfo1)      f.close()  实验结果：     1. 使用urllib模块或request模块读取网页内容，并利用BeautifulSoup模块进行内容解析，编写爬虫从<https://www.biqukan.com/2_2671>爬取小说《余罪》的内容。要求最终爬取的内容从第一章开始，且不存在大量空格等非必要字符。   由于爬虫爬取过快，也许触发了网站的网站的反爬虫机制，导致一段时间内抓取不到信息，于是换了一个相似的网站http://www.aidusk.com/t/19597/爬取小说《余罪》  代码实现：  import requests  from bs4 import BeautifulSoup  import re  import time  url = 'http://www.aidusk.com/t/19597/'  headers = {      "User-Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/107.0.0.0 Safari/537.36"}  reqst = requests.get(url, headers=headers)  reqst.encoding = 'utf-8'  li = BeautifulSoup(str(BeautifulSoup(reqst.text, 'lxml').find\_all(  'div', attrs={'class': 'book\_con\_list'})[1]), 'lxml').find\_all('li')  for i in li:      i\_bs = BeautifulSoup(str(i), 'lxml')      name = i\_bs.a.string  link = i\_bs.a['href']      newReqst = requests.get(url+link, headers=headers)      newReqst.encoding = 'utf-8'      chapterContent = BeautifulSoup(newReqst.text, 'lxml').find(          'div', attrs={'id': 'content'})      print("正在获取"+name)      f = open('./余罪/'+name+'.txt', 'w', encoding='utf-8')      chapterContent = str(chapterContent)      chapterContent = re.sub('<.\*?>', '', chapterContent)      f.write(chapterContent)      f.close()      time.sleep(5)  # 防止触发反爬  爬取结果：     1. 学习使用Feapder框架，编写爬虫，爬取中国工程院院士信息。   首先在列表页获取每个院士的介绍链接，然后将采集到的url作为新任务派发  yield feapder.Request(link, callback=self.parse\_detail, name=name)，默认回调函数应该是self.paese(),这里我们更改为自定义的获取详情的parse\_detail().  代码实现：  import feapder  class TophubSpider(feapder.AirSpider):      def start\_requests(self):          yield feapder.Request("https://www.cae.cn/cae/html/main/col48/column\_48\_1.html")      def parse(self, request, response):          lists = response.xpath("//li[@class='name\_list']")          for list in lists:              name = list.xpath("./a/text()").extract\_first()              link = list.xpath("./a/@href").extract\_first()              yield feapder.Request(link, callback=self.parse\_detail, name=name)      def parse\_detail(self, request, response):          name = request.name          content = response.xpath(              'string(//div[@class="intro"])').extract\_first()          print(name)          print(content)          f = open('./contents/'+name+'.txt', 'w', encoding='utf-8')          f.write(content)          f.close  if \_\_name\_\_ == "\_\_main\_\_":      TophubSpider().start()  运行结果： | | | |
| 结论分析与体会：  通过本次实验，熟悉了urllib模块、requests模块+BeautifulSoup模块使用、Feapder框架的用法，练习了编写爬虫。 | | | |