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：202100460069 | 姓名：孟庆丹 | | 班级：网安21.2 |
| 实验题目： urllib模块使用 | | | |
| 实验学时：2 | | 实验日期：2022.11.10 | |
| 实验目的：熟悉模块的用法，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. 网页内容读取，读取指定URL的网页内容   利用urllib模块爬取指定网页的内容并写入本地文件即可。  **代码实现：**   import urllib.request  url = input()  link = "https://"+url  fp = urllib.request.urlopen(link)  contents = fp.read().decode('utf-8')  f = open(str(url[0:3])+".html", 'w', encoding='utf-8')  f.write(str(contents))  f.close  爬取效果：     1. 利用urllib模块+正则表达式爬取<http://tieba.baidu.com/p/2460150866里的的所>有图片，保存到本地。   遍历爬取帖子中的每一页，使用正则表达式<img.\*?class="BDE\_Image".\*?>提取回帖中的图片信息，然后利用正则表达式(?<=src=").\*?(?=")提取从img标签中获得图片的地址，并从图片地址爬取图片写入本地指定位置。  **代码实现：**  import urllib.request  import re  link = "https://tieba.baidu.com/p/2460150866?pn="  pattern1 = re.compile('<img.\*?class="BDE\_Image".\*?>')  # 提取img标签  pattern2 = re.compile('(?<=src=").\*?(?=")')  # 从img标签中提取图片地址  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",      "Accept": "text/html,application/xhtml+xml,application/xml;q=0.9,image/avif,image/webp,image/apng,\*/\*;q=0.8,application/signed-exchange;v=b3;q=0.9"}  for i in range(1, 61):      print("爬取第", i, "页中的图片...")      req = urllib.request.Request(url=link+str(1), headers=headers)      fp = urllib.request.urlopen(req)      contents = fp.read()      pics = pattern1.findall(str(contents))      pics\_src = list()      for j in pics:          pics\_src.append(pattern2.findall(j)[0])      for j in pics\_src:          if 'https' in j:              fpic = urllib.request.urlopen(j)              f = open("./pics/"+j[-15:], 'wb')              f.write(fpic.read())              f.close()  爬取效果：     1. 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   利用正则表达式<li class="name\_list">.\*?</li>和(?<=<li class="name\_list"><a href=").\*?(?=")在院士列表页面中提取所有的院士个人链接利用正则表达式(?<=target="\_blank">).\*?(?=</a></li>)提取院士的姓名，匹配院士个人链接里院士的个人信息及照片链接（正则表达式在下面代码中），在本地创建文件夹并写入个人信息以及图片链接对应的照片。  **代码实现：**  import urllib.request  import re  import os  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"}  pattern1 = re.compile('<li class="name\_list">.\*?</li>')  pattern2 = re.compile(      '(?<=<li class="name\_list"><a href=").\*?(?=")')  # 匹配院士个人链接  pattern3 = re.compile('(?<=target="\_blank">).\*?(?=</a></li>)')  # 匹配院士姓名  # 匹配院士照片链接  pattern4 = re.compile('<div class="info\_img">.\*?</div>')  pattern5 = re.compile('(?<=src=").\*?(?=")')  # 匹配院士介绍  pattern6 = re.compile('<div class="intro">.\*?</div>')  pattern7 = re.compile('(?<=&ensp;&ensp;&ensp;&ensp;).\*?(?=</p>)')  # 中文解码  def cnDecode(codes):      j = 0      for i in codes:          i = eval(repr(i).replace('\\\\', '\\'))          i = i.encode('raw\_unicode\_escape').decode()          codes[j] = i          j += 1  links = list()  # 院士个人链接  names = list()  # 院士姓名  req = urllib.request.Request(url=link1, headers=headers)  fp = urllib.request.urlopen(req)  contents = fp.read()  info = pattern1.findall(str(contents))  for i in info:      links.append(pattern2.findall(i)[0])      names.append(pattern3.findall(i)[0])  cnDecode(names)  pwd = os.getcwd()  k = 0  for i in links:      req = urllib.request.Request(url=link+i, headers=headers)      fp1 = urllib.request.urlopen(req)      contents1 = fp1.read()      picinfo = pattern4.findall(str(contents1))      introinfo = pattern6.findall(str(contents1))      piclink = link + pattern5.findall(picinfo[0])[0]      maininfo = pattern7.findall(introinfo[0])      # 创建院士个人文件夹      os.mkdir(pwd+"/info/"+names[k])      # 写院士信息      fh = open("./info/"+names[k]+"/"+names[k]+".txt", "w")      cnDecode(maininfo)      for j in maininfo:          fh.write(j+'\n')      fh.close()      # 写院士照片      fh = open("./info/"+names[k]+"/"+names[k]+".jpg", "wb")      for j in piclink:          req = urllib.request.Request(url=piclink, headers=headers)          fp2 = urllib.request.urlopen(req)          contents2 = fp2.read()          fh.write(contents2)      fh.close()      k += 1  爬虫效果： | | | |
| 结论分析与体会：  通过本次实验，熟悉了熟悉模块的用法，练习爬虫的编写。 | | | |