****

高级语言程序设计

大作业报告

 

# 题目名称: 西电教师主页爬取

1. 题目要求

**通过编写python爬虫程序，读取以下列表，电院教师的主页基本信息**

**电子工程学院教师列表-西安电子科技大学教师主页** [**https://faculty.xidian.edu.cn/xyjslb.jsp?urltype=tsites.CollegeTeacherList&wbtreeid=1001&st=0&id=1583&lang=zh\_CN#collegeteacher**](https://faculty.xidian.edu.cn/xyjslb.jsp?urltype=tsites.CollegeTeacherList&wbtreeid=1001&st=0&id=1583&lang=zh_CN)

**或者** [**https://web.xidian.edu.cn**](https://web.xidian.edu.cn/) **网站**

**将教师的 姓名，职称，性别，毕业院校，学历，学位，单位，院系，学科,,邮箱，点击次数等信息 读取并保存为csv格式文件。遍历全院300多位老师主页，存储300多位老师的基本信息。**

**如果能完成学校个人主页全部教师信息采集，更佳！**

**给出程序的主要流程，模块函数说明，执行输入输出截图。并附上源代码。所使用的python库的名字 以及功能说明**

**提交大作业报告和源代码,以及可执行程序,采集到的CSV文件，给出点击次数最多的10位老师。录屏程序运行视频.mp4。建议采用python installer打包成独立运行的程序.exe文件**

1. 设计思路

使用requests库对网址解析，由于操作过程中发现有些网址老师的网站是自己设计的所以有时候进行re和lxml采集进行会有问题

后来我转变思路直接从数据库中进行获取数据

把UI换成手机版发现手机版每次都是从数据库进行获取数据一下子就简单了很多

url = "https://faculty.xidian.edu.cn/system/resource/tsites/getsitelistcontent.jsp?collegeId=1583&disciplineId=0&honorId=0&pinYin=&requestUrl=http%3A%2F%2Ffaculty.xidian.edu.cn%2Fxyjslb.m.jsp&comType=collegeTeacher&treeid=1001&lang=zh\_CN&viewmode=8&viewid=70564&siteOwner=1438110714&start={a}&end={b}&startnum={a}&endnum={b}&viewUniqueId=u7".format(*a*=k,*b*=k+6)

这个代码就代表着我从数据库中直接获取信息

怕数据进行丢失操作我采用每次获得六个老师的信息

然后采用循环进行遍历整个数据库

当然要加上header的头

获取到result我们就解析

解析过程很简单就是对json字典进行获取

最后我们使用csv库对csv进行文件逐行读写就完成了

三、程序结果展示、

双击可以点开

四、课程收获建议

通过这次对西电教师主页的爬取我体会到了网络爬虫的便捷性，我可以轻松的获取大量数据，从而进行系统化的分析和处理，得到自己想要的结果。

对request lxml 等有了一定的了解，为今后需要进行数据处理的情况打下了坚实的基础，python确实是一种胶水语言，他富含各种库可供开发者进行使用，这是开源的非常大的亮点，让python生态越来越完善

五、参考文献

Request

Re

Csv

Lxml

六、项目代码

# 记录一下神奇的代码编写环节：

#     首先我准备使用电脑对西电教师主页进行爬取

#     然后步骤分解了得到每个老师的网址

#     然后解析每个老师网址

#     结果错误百出，解析经常出现问题

#     然后我就一不小心点了手机的UI

#     发现反馈的包中包括了所有的老师地址

#     只需要解析那个json文件就可以了

#     然后csv保存

import requests

from lxml import etree

import csv

with open('result.csv', 'w+', *encoding*='utf-8', *newline*='') as f:

    s = ["姓名","职称","性别","毕业院校","学历","学位","单位","学科","邮箱","联系方式","位置","点击数"]

    writer = csv.writer(f)

    writer.writerow(s)

f.close()

headers = {

    "Accept": "application/json, text/javascript, \*/\*; q=0.01",

    "Accept-Encoding": "gzip, deflate, br",

    "Accept-Language": "zh-CN,zh;q=0.9,en;q=0.8,en-GB;q=0.7,en-US;q=0.6",

    "Connection": "keep-alive",

    "Cookie": "JSESSIONID=0E9F54468974A09E8DDA378DF18A1AF9; UqZBpD3n3iPIDwJU=v1K6FbQwSDYkW",

    "Host": "faculty.xidian.edu.cn",

    "sec-ch-ua-mobile": "?1",

    "sec-ch-ua-platform": "Android",

    "Sec-Fetch-Dest": "empty",

    "Sec-Fetch-Mode": "cors",

    "Sec-Fetch-Site": "same-origin",

    "User-Agent": "Mozilla/5.0 (Linux; Android 6.0; Nexus 5 Build/MRA58N) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/110.0.0.0 Mobile Safari/537.36 Edg/110.0.1587.56",

    "X-Requested-With": "XMLHttpRequest",

}

# def getsid():

#     pass

# def gethomeid(): 放弃了这个步骤，换用每次请求url根据lxml查询获得

#     pass

*def* getclick(*url*):

    headers = {

        "sec-ch-ua-mobile": "?1",

        "sec-ch-ua-platform": "Android",

        "Sec-Fetch-Dest": "empty",

        "User-Agent": "Mozilla/5.0 (Linux; Android 6.0; Nexus 5 Build/MRA58N) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/110.0.0.0 Mobile Safari/537.36 Edg/110.0.1587.69",

    }

    respp = requests.get(*url*,*headers*=headers)

    pass

*def* getdata():

    num = 0

    n = 0

    for k in range(0,400,6):

        url = "https://faculty.xidian.edu.cn/system/resource/tsites/getsitelistcontent.jsp?collegeId=1583&disciplineId=0&honorId=0&pinYin=&requestUrl=http%3A%2F%2Ffaculty.xidian.edu.cn%2Fxyjslb.m.jsp&comType=collegeTeacher&treeid=1001&lang=zh\_CN&viewmode=8&viewid=70564&siteOwner=1438110714&start={a}&end={b}&startnum={a}&endnum={b}&viewUniqueId=u7".format(*a*=k,*b*=k+6)

        resp = requests.get(*url*=url,*headers*=headers)

        result = resp.json()

        w = 0

        for i in result:

            n+=1

            w+=1

            #将教师的 姓名，职称，性别，毕业院校，学历，学位，单位，院系，点击次数等信息

            name = i["name"]

            prorank = i["prorank"]

            sex = i["sex"]

            graduatedUniversity = i["graduatedUniversity"]

            education = i["education"]

            degree = i["degree"]

            unit = i["unit"]

            email = i["email"]

            discipline = i["discipline"]

            contact = i["contact"]

            officeLocation = i["officeLocation"]

            url = i["url"]

            num+=1

            a = [name,prorank,sex,graduatedUniversity,education,degree,unit,discipline,email,contact,officeLocation,url]

            # print("正在打印",n)

            save(a)

        print("------------------------------",w,k)

*def* save(*a*):

    s = ["姓名","职称","性别","毕业院校","学历","学位","单位","学科","邮箱","联系方式","位置","网址"]

    with open('result.csv', 'a+', *encoding*='utf-8', *newline*='') as f:

        writer = csv.writer(f)

        writer.writerow(*a*)

    f.close()

*def* main():

    getdata()

if \_\_name\_\_ == "\_\_main\_\_":

    main()