链滴

# 简单百度爬虫、查询个人公司信息的实现

作者：sirwsl

原文链接：https://ld246.com/article/1590066538386

来源网站：链滴

本来对python兴趣不大，但是为了期中考，简单记录一下超级无敌简单的爬虫实现的过程

接下来开始大白模式的记录，打发我无聊的时光：

1. headers:解决request反爬虫，就是当我们访问部分网页的时候，会出现无法爬取数据、或者无法问的时候，这时候采用headers,将爬虫伪装成浏览器去访问，这样就起到解决反爬虫的作用。
2. url:需要爬取的网页连接，注意{}的使用，解决了后面自己赋值的问题
3. requests.get(url),进行服务器访问请求，由request内部生成url对象（具体可以看一下人家： https://blog.csdn.net/k_koris/article/details/82950654）
4. url.format进行之前的{}赋值
5. headers=headers设置相应request header
6. etree.HTML：采用XPath进行资源的解析修正，便于后面截取处理（详情查看其他大佬博客： https://blog.csdn.net/qq_38410428/article/details/82792730）
7. for循环：没什么好说的，就是打字浪费一下时间
8. XPath：进行对元素属性的遍历查找（参照： https://www.cnblogs.com/lei0213/p/7506130.html）
9. join：字符连接操作，这个不懂么emmmmm（孩子没救了）

10、同样的，replace、split、format，不会自己百度。。。。。

PS：这里说一下headers的获取方法，打开Edge，按F12，找到network，刷新，找到request heades。如果还不会参照人家大佬的（https://blog.csdn.net/ysblogs/article/details/88530124）

好了，废话不多说，记录下代码，hhhhh，看到的如果小白可以照葫芦画瓢，大佬绕道

```
import requests
from lxml import etree
```

```
headers = {
    'User-Agent':
    'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chr
me/67.0.3396.99 Safari/537.36'
}
url = "https://www.baidu.com/s?wd={}&rn=20&ie=utf-8&usm=4&rsv_pq=dda16fac00085a5
&rsv_t=ca17vQlKisiOERxSWewcnyg/K/0flYw9KAqdtGiqNMpwAXUTAmhv6MG/f5M"
keyword = input('请输入关键词：')
response = requests.get(url.format(keyword), headers=headers)
html = etree.HTML(response.text)

for i in range(1, 21):
    title = html.xpath(
        '/html/body/div/div[3]/div[1]/div[3]/div[{}]/h3/a//text()'.format(i))
    title_ = ''.join(''.join(title).replace(' ', '').split())
    print(title_, '\n')
```

接下来附上这次期中作业的代码。下面这个企信通查询个人名下的公司基本信息的一个实现

```
# -*- coding:utf-8 -*-
import requests
import xlwt
from lxml import etree
headers = {
    'User-Agent':
    'Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chr
me/70.0.3538.102 Safari/537.36 Edge/18.18362',
    'Accept':
    'image/png, image/svg+xml, image/*; q=0.8, */*; q=0.5',
    'Accept-Encoding': 'gzip,deflate,br',
    'Connection':'Keep-Alive',
    'Accept-Language':'zh-Hans-CN,zh-Hans;q = 0.5',
    'Host':'hm.baidu.com',
    'Referer':'https://www.tianyancha.com/?jsid=SEM-BAIDU-PZ2005-SY-000001'
}
url = "https://www.qixintong.cn/qxtsearch/?key={}&typestr=0"
keyword = input("请输入姓名：")
num = int(input("请输入查询次数："))
response = requests.get(url.format(keyword), headers=headers)
html = etree.HTML(response.text)
#print(html)查看请求成功的list

#创建表格
wb = xlwt.Workbook()
sh = wb.add_sheet("test")
#开始解析
index = '/html/body/div[3]/div/div/div/div[2]/ul/li[{}]/'
test =['h2','h2/span','span[1]','span[2]','span[3]','p']

for i in range(int(num)):
    for j in range(6):
        flag = index+test[j]+'//text()'
        name = html.xpath(flag.format(i))
```

```
        if j==0 or j==1:
            name_ = ''.join(''.join(name).replace(" "," ").split())
        else:
            name_ = ''.join(''.join(name))
        print(name_, '\n')
        sh.write(i,j,name_)
wb.save("企信通.xls")
print("提取结束，保存退出")
```