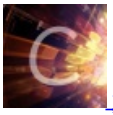
python爬取看雪论坛的所有主题帖的回复消息

原创

郭小文 于 2018-11-27 10:20:04 发布 1501 收藏 7
分类专栏： python 文章标签： 看雪
版权声明：本文为博主原创文章，遵循 CC 4.0 BY-SA 版权协议，转载请附上原文出处链接和本声明。
本文链接：https://blog.csdn.net/weixin_39641975/article/details/84561184
版权



python 专栏收录该内容
6 篇文章 0 订阅
订阅专栏

最近因为实验课题的需要，我们对看雪论坛的消息回复进行爬取，

https://bbs.pediy.com/（看雪论坛）

对于看雪论坛的消息回复查看的一般顺序为：

进入看雪论坛的主页----->选择查看的主题----->选择想要查看的话题-------->查看该话题的所有回复信息

代码主要分三个模块，首先就是对所有的主题的的链接进行爬取

然后再对每个主题里面的话题链接进行爬取，最后就是访问话题的链接，爬取回复的消息内容

```
from bs4 import BeautifulSoup
from urllib.request import urlopen
import random
import requests
import time
import thread6
import re


"""

2018-11-26
```

```
author:郭文博

"""
def get_url(url,headers):       #   首先是获取到主页面所有的主题链接网址

    Theme = {}


    """
        模拟浏览器来获取网页的html代码
        """


    timout = random.choice(range(80,100))

    request = requests.get(url,headers = headers)

    if(request.status_code!=200):

        print("获取网址失败")

    html = BeautifulSoup(request.text,"html.parser")

    theme = html.find_all("div",{"class":"card px-0"})

    for i in theme:

        themecontant = i.find_all("a")

        for j in themecontant:

            href = j['href']

            themeString = j.string

            if(themeString == None):

                continue

            themestring = themeString.strip()

            Theme[themestring] = href

    print(Theme)
    return  Theme




def get_topic_url(url,urldist,headers):    #   获取每一个主题的所有话题的URL

    themeitemurl = {}

    listurl = []

    for value in urldist.values():
```

```
        themeurl = url + value
        # print(themeurl)
        request = requests.get(themeurl,headers = headers)
        if (request.status_code != 200):
            # print("获取网址失败")
            continue
        else:
            html = BeautifulSoup(request.text, "html.parser")
            urlhtml = html.find_all("nav")
            for i in urlhtml:
                urlcontant = i.find_all("li",{"class":"page-item"})
                for j in urlcontant:
                    itemhref = j.find_all("a")
                    for j in itemhref:
                        href = j['href']
                        themeString = j.string
                        if (themeString == None):
                            continue
                        themestring = themeString.strip()
                        themeitemurl[themestring] = href
        listurl.append(themeitemurl)
    print(listurl)
    return listurl

def get_contanturl(url,listurl,headers):        #  获取每个话题的url
    contanturl = {}
    contanturllist = []
    for i in listurl:
        for values in i.values():
            URL = url + values
            request = requests.get(URL,headers = headers)
```

```
            request = requests.get(URL,headers = headers)
            if(request.status_code != 200):
                continue
            html = BeautifulSoup(request.text,"html.parser")
            htmlurl = html.find_all("tr")
            for k in htmlurl:
                htmlhref = k.find_all("div",{"class":"subject"})
                for href  in htmlhref:
                    a = href.find_all("a")
                    lena = len(a)
                    if(lena>1):
                        # print(a[1])
                        topicstring = a[1].string
                        if (topicstring == None):
                            continue
                        Topicstring = topicstring.strip()
                        contanturl[Topicstring] = a[1]['href']
                    else:
                        # print(a[0])
                        topicstring = a[0].string
                        if (topicstring == None):
                            continue
                        Topicstring = topicstring.strip()
                        contanturl[Topicstring] = a[0]['href']
                print(contanturl)
        contanturllist.append(contanturl)
        # print(contanturllist)
    print(contanturllist)
    return contanturllist

def get_contant(url,urllist,headers):         #    获取每一个话题的所有论坛回复
```

```
    contant = {}
    contantlist = []
    for i in urllist:
        for values in i.values():
            contanturl = url + values
            request = requests.get(contanturl,headers = headers)
            if(request.status_code != 200):
                continue
            html = BeautifulSoup(request.text,"html.parser")
            Name = html.find_all("tr",{"class":"post"})

            for k in Name:
                contantkey = ''
                contantvalue = ''
                td = k.find_all("td",{"class":"px-0"})
                for TD in td:
                    span = TD.find_all("span",{"class":"username font-weight-bold"})
                    Contant = TD.find_all("div",{"class":"message mt-1 break-all"})
                    for Span in span:
                        name = Span.find_all("a")
                        contantkey = name[0].string.strip()
                        contantvalue = Contant[0].string
                        contantValue = ''
                        if (contantvalue != None):
                            contantValue = contantvalue.strip()
            contant[contantkey] = contantValue
            print(contant)
        contantlist.append(contant)
    print(contantlist)
    return contantlist
if __name__ == "__main__":
```

```
if __name__ == '__main__':

    url = "https://bbs.pediy.com/"

    headers = {
        'Accept': 'text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,*/*;q=0.8',
        'Accept-Encoding': 'gzip, deflate, sdch',
        'Accept-Language': 'zh-CN,zh;q=0.8',
        'Connection': 'keep-alive',
        'User-Agent': 'Mozilla/5.0 (Windows NT 6.3; WOW64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/43
    }

    Theme = get_url(url,headers)

    topicurl = get_topic_url(url,Theme,headers)

    topiccontanturl = get_contanturl(url,topicurl,headers)

    get_contant(url,topiccontanturl,headers)
```

源码查看github：https://github.com/wenboi/Spider

个人网站：http://guoxiaowen.com