个人资料

tobacco

访问： 2545982次

积分： 18042

等级： BLOG 7

排名： 第268名

原创： 150篇 转载： 4篇

译文： 0篇 评论： 118条

Python爬虫之自动登录与验证码识别

标签： python 爬虫 验证码识别

2016-02-06 19:19 10217人阅读 评论(0) 举报

分类： python（21）

Python爬虫之自动登录与验证码识别

在用爬虫爬取网站数据时，有些站点的一些关键数据的获取需要使用账号登录，这里可以使用requests发送登录请求，并用Session对象来自动处理相关Cookie。

另外在登录时，有些网站有时会要求输入验证码，比较简单的验证码可以直接用pytesser来识别，复杂的验证码可以依据相应的特征自己采集数据训练分类器。

以CSDN网站的登录为例，这里用Python的requests库与pytesser库写了一个登录函数。如果需要输入验证码，函数会首先下载验证码到本地，然后用pytesser识别验证码后登录，对于CSDN登录验证码，pytesser的识别率很高。

其中的pytesser的下载地址为: pytesser下载

具体代码如下：

```
#coding:utf-8
import sys
import time
import urllib
import shutil
import pytesser
import requests

from lxml import etree

config = {'gid': 1}

def parse(s, html, idx):
    result = {}

    tree = etree.HTML(html)
    try:
        result['lt'] = tree.xpath('//input[@name="lt"]/@value')[0]
        result['execution'] = tree.xpath('//input[@name="execution"]/@value')[0]
        result['path'] = tree.xpath('//form[@id="fm1"]/@action')[0]
    except IndexError, e:
        return None

    valimg = None
    valimgs = tree.xpath('//img[@id="yanzheng"]/@src')
    if len(valimgs) > 0:
        valimg = valimgs[0]

```

编程日记 (20)
C (22)
操作系统 (14)
Android (10)
Linux (6)
Java (3)
C++ (31)
Qt (9)
java日记 (19)
Android日记 (8)
python (22)
opencv (5)
数据库 (9)

文章存档

2016年02月 (5)
2016年01月 (6)
2015年12月 (1)
2015年09月 (1)
2015年05月 (1)
展开

阅读排行

身份证号码编码方法及校 (48469)
Python获取免费的可用代 (46277)
Python搭建聊天机器人微 (43955)
Python 利用PIL将图片转 (43839)
Android手机通过socket! (32142)
Oracle11g远程连接配置 (30543)
Java中用内存映射处理大 (23596)
关于Android中的乱码 (23015)
Python搭建聊天机器人 (22054)
C++：成员函数实现在类 (19766)

评论排行

Android手机通过socket! (30)
wxBot微信机器人框架 (11)
Qt实现Windows远程控制 (11)
Android实现远程控制PC (9)
Java中用内存映射处理大 (6)
C++中内联函数inline的 (5)
Linux下Android开发手机 (5)
ffmpeg解码内存缓冲区 (5)
Python搭建聊天机器人 (4)
C++中的explicit关键字 (3)

最新评论

wxBot微信机器人框架
tobacco: @shqlsl:嗯，现在问题还挺多的。目前在忙于其它事情，后期会抽时间慢慢优化~

wxBot微信机器人框架
shqlsl: @tobacco5648:问题几个1。 的名称一定要在先保存通迅录才能读取？2。 里的陌生人，可以显...

wxBot微信机器人框架
tobacco: @shqlsl:没多少人，不必建群的。修改代码可以直接在

```
29      validateCode = None
30      if valimg:
31          fname = 'img/' + str(idx) + '_' + str(config['gid']) + '.jpg'
32          config['gid'] = config['gid'] + 1
33          ri = s.get("https://passport.csdn.net" + valimg)
34          with open(fname, 'wb') as f:
35              for chk in ri:
36                  f.write(chk)
37              f.close()
38          validateCode = pytesser.image_file_to_string(fname)
39          validateCode = validateCode.strip()
40          validateCode = validateCode.replace(' ', '')
41          validateCode = validateCode.replace('\n', '')
42          result['validateCode'] = validateCode
43
44      return result
45
46  def login(usr, pwd, idx):
47      s = requests.Session()
48
49      r = s.get('https://passport.csdn.net/account/login',
50      headers={'User-Agent': 'Mozilla/5.0 (Windows NT 6.1; WOW64; rv:41.0) Gecko/20100101 F
51
52      while True:
53          res = parse(s, r.text, idx)
54          if res == None:
55              return False
56          url = 'https://passport.csdn.net' + res['path']
57          form = {'username': usr, 'password':pwd, '_eventId':'submit', 'execution':res['ex
58          if res.has_key('validateCode'):
59              form['validateCode'] = res['validateCode']
60          s.headers.update({
61              'User-Agent': 'Mozilla/5.0 (Windows NT 6.1; WOW64; rv:41.0) Gecko/20100101 Fi
62              'Accept-Language': 'zh-CN,zh;q=0.8,en-US;q=0.6,en;q=0.4',
63              'Content-Type': 'application/x-www-form-urlencoded',
64              'Host': 'passport.csdn.net',
65              'Origin': 'https://passport.csdn.net',
66              'Referer': 'https://passport.csdn.net/account/login',
67              'Upgrade-Insecure-Requests': 1,
68              })
69          r = s.post(url, data=form)
70
71          tree = etree.HTML(r.text)
72          err_strs = tree.xpath('//span[@id="error-message"]/text()')
73          if len(err_strs) == 0:
74              return True
75          err_str = err_strs[0]
76          print err_str
77          err = err_str.encode('utf8')
78
79          validate_code_err = '验证码错误'
80          usr_pass_err = '帐户名或登录密码不正确，请重新输入'
81          try_later_err = '登录失败连续超过5次，请10分钟后再试'
82
83          if err[:5] == validate_code_err[:5]:
84              pass
85          elif err[:5] == usr_pass_err[:5]:
86              return False
87          elif err[:5] == try_later_err[:5]:
88              return False
89          else:
90              return True
91
92  if __name__ == '__main__':
93      main(sys.argv[1], sys.argv[2], 0)
```

github上发Pull Request的~

wxBot微信机器人框架
shqlsl: 可以拉一个QQ群，大伙一起完善好吗？github 别人修改的也可以提交上去吗？

wxBot微信机器人框架
tobacco: @tobacco5648:出现这个是因为收到腾讯新闻的推送消息了，腾讯新闻的消息我还没解析过。

wxBot微信机器人框架
tobacco: @haotian2546:appid是指程序微信客户端的id，wxBot用的是Web微信的接口，所以...

wxBot微信机器人框架
tobacco: @haotian2546:pycharm报红但是能运行吗

wxBot微信机器人框架
haotian2546: pm = re.search(r'window.synccheck={retcode:"(\d+)"...

wxBot微信机器人框架
haotian2546: 谢谢了啊，不过还是 stranger ....，只是这次没有出现这个错误“ This user do...

wxBot微信机器人框架
tobacco: @haotian2546:我在github上更新了下，你重新clone下再试试~

上一篇 Python获取免费的可用代理
下一篇 Python包装网页微信API并实现简单自动回复

我的同类文章

python（21）

主题推荐

python 函数 session cookie html 验证码识别 爬虫 验证码 数据

猜你在找

查看评论

暂无评论

您还没有登录,请[登录]或[注册]

* 以上用户言论只代表其个人观点，不代表CSDN网站的观点或立场

核心技术类目

全部主题 Hadoop AWS 移动游戏 Java Android iOS Swift 智能硬件 Docker OpenStack
VPN Spark ERP IE10 Eclipse CRM JavaScript 数据库 Ubuntu NFC WAP jQuery
BI HTML5 Spring Apache .NET API HTML SDK IIS Fedora XML LBS Unity
Splashtop UML components Windows Mobile Rails QEMU KDE Cassandra CloudStack
FTC coremail OPhone CouchBase 云计算 iOS6 Rackspace Web App SpringSide
Maemo Compuware 大数据 aptech Perl Tornado Ruby Hibernate ThinkPHP HBase
Pure Solr Angular Cloud Foundry Redis Scala Django Bootstrap