爬虫模拟登录

模拟登录知乎：

著作权归作者所有。  
商业转载请联系作者获得授权，非商业转载请注明出处。  
作者：Jean Green  
链接：https://www.zhihu.com/question/29925879/answer/54819870  
来源：知乎

用抓包工具看到登录的url为

http://www.zhihu.com/login/email

,  
POST提交表单格式为：

{email: "", password: "" ,captcha:"", rememberme:"true", \_xsrf: ""}

,  
获取xsrf就是直接读取登录页的xsrf标签，captcha有时候会出现，可以请求

http://www.zhihu.com/captcha.gif

来获取，然后手动输入。  
返回

{

"r": 0,

"msg": "\u767b\u9646\u6210\u529f"

}

表示登录成功，对了他们写错字了，不是登陆成功。=\_=||  
  
我可以贴上代码，我用python写的，当时只是为了查看某个人的关注了哪些人

-\*- coding:utf-8 -\*-

import requests

import re

from bs4 import BeautifulSoup

def login():

url = 'http://www.zhihu.com'

login\_url = url+'/login/email'

login\_data = {

'email': 你的邮箱,

'password': 你的密码,

'rememberme': 'true',

}

headers\_base = {

'Accept': 'text/html,application/xhtml+xml,application/xml;q=0.9,image/webp,\*/\*;q=0.8',

'Accept-Encoding': 'gzip, deflate, sdch',

'Accept-Language': 'en-US,en;q=0.8,zh-CN;q=0.6,zh;q=0.4,zh-TW;q=0.2',

'Connection': 'keep-alive',

'Host': 'www.zhihu.com',

'User-Agent': 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10\_10\_3) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/43.0.2357.130 Safari/537.36',

'Referer': 'http://www.zhihu.com/',

}

s = requests.session()

def get\_xsrf(url=None):

r = s.get(url, headers=headers\_base)

xsrf = re.search(r'(?<=name="\_xsrf" value=")[^"]\*(?="/>)', r.text)

if xsrf == None:

return ''

else:

return xsrf.group(0)

xsrf = get\_xsrf(url)

login\_data['\_xsrf'] = xsrf.encode('utf-8')

captcha\_url = 'http://www.zhihu.com/captcha.gif'

captcha = s.get(captcha\_url, stream=True)

print captcha

f = open('captcha.gif', 'wb')

for line in captcha.iter\_content(10):

f.write(line)

f.close()

print u'输入验证码:'

captcha\_str = raw\_input()

login\_data['captcha'] = captcha\_str

res = s.post(login\_url, headers=headers\_base, data=login\_data)

print res.status\_code

m\_cookies = res.cookies

test\_url = 'http://www.zhihu.com/people/某人的url/followees'

res = s.get(test\_url, headers=headers\_base, cookies=m\_cookies)

def get\_users(content=None):

users = re.search(r'<a title="\*>', content)

print users.group(0)

get\_users(res.text)

ids = 'fd80e95ce7069f0f06cc0ac07be34864' #某人的member id

nums = 300

def get\_fos(nums, ids):

users = []

for index in range(20, nums+1, 20):

fo\_url = 'http://www.zhihu.com/node/ProfileFolloweesListV2'

m\_data = {

'method': 'next',

'params': '{"offset":'+ str(index) + ',"order\_by":"created","hash\_id":"'+ ids +'"}',

'\_xsrf': xsrf,

}

res = s.get(fo\_url, headers=headers\_base, data=m\_data)

soup = BeautifulSoup(res.text, 'html.parser')

for link in soup.find\_all('a', 'zm-item-link-avatar'):

users.append(link.get('href'))

return users

users = get\_fos(nums, ids)

def work():

login()

if \_\_name\_\_ == '\_\_main\_\_':

work()