创作**python获取代理(终极版本)**

[友情链接：python获取代理IP](http://blog.csdn.net/HelloHaibo/article/details/77513731)

首先感谢各个提供免费代理IP的网站，虽然IP质量不能保证，但是科技改变生活，让我们开始用程序来“淘金”吧。

之前做过一次获取代理IP（网址可以参考顶部友情链接）由于刚开始玩这些，有些缺点没发现，运行几次后发现了问题，在此花了一下午加熬夜到凌晨3点，把程序修改到第三版，效果目前来说还算满意。

1、**第一版**是从代理网站过去一次IP，经检查后，重复使用，直到一个IP两次都是无效的，则从列表删除

优点：IP利用率高，访问代理网站次数少

缺点：直接使用本机IP访问，速度慢，一次一个

2、**第二版**从代理网站获取一次IP，然后全部使用一次，全部丢弃，再重新访问

优点：速度快

缺点：IP利用率不高，本机IP容易被网站（抓取代理IP的网站）服务器加入黑名单，返回503拒绝服务错误，本人亲身经历

3、**第三版**从网站获取代理IP后，再反向使用获取的代理IP爬取代理网站

优点：**多线程，速度快，不会被服务器拉黑名单，自动检测IP缓冲区，低于下限制，开启获取，高于上限制，停止获取**

缺点：第一次的获取“启动IP”是关键

getProxyIP\_V3.py

1. *#!/usr/bin/python*
2. *#-\*- coding:utf-8 -\*-*
4. *#author:dasuda*
5. *#CSDN blog:HelloHaibo*
6. *#data:2017.8.25*

import urllib2

import re

import socket

import threading

import global\_para

import targetURLs

import random

import datetime

import time

import user\_agents

import get\_first\_ip

import os

*#thread lock*

lock = threading.Lock()

*#get ip and port from html*

def html\_to\_ip(html):

a = re.compile(r'(?<=<td>)[\d]{1,3}\.[\d]{1,3}\.[\d]{1,3}\.[\d]{1,3}')

b = re.compile(r'(?<=<td>)[\d]{2,5}(?=</td>)')

html\_to\_ip\_ip\_table=[]

findIP = re.findall(a, str(html))

findPORT = re.findall(b, str(html))

for i in range(len(findIP)):

temp = findIP[i] + ":" + findPORT[i]

html\_to\_ip\_ip\_table.append(temp)

*# html\_to\_ip\_ip\_table is a list*

return html\_to\_ip\_ip\_table

*#thread function*

*#para1: 0 -> get 1->attack*

def get\_one(url\_get,good\_ip,cnt,para1,para2):

*#open url timeout*

socket.setdefaulttimeout(5)

try:

*#assemble ip like {'http':'1.1.1.1:8888'}*

aseemble\_ip = {'http': good\_ip}

proxy\_support = urllib2.ProxyHandler(aseemble\_ip)

openercheck = urllib2.build\_opener(proxy\_support)

urllib2.install\_opener(openercheck)

if para1 == 0:

temp\_url\_get = url\_get+str(random.randint(1,3))

elif para1==1:

temp\_url\_get = url\_get

request = urllib2.Request(temp\_url\_get)

temp\_agent = random.choice(user\_agents.user\_agents)

request.add\_header('User-Agent',temp\_agent)

content = urllib2.urlopen(request).read()

if para1 == 0:

*# print content*

tt = html\_to\_ip(content)

print 'this one is ok ',temp\_url\_get

lock.acquire()

*#print(global\_para.IP\_data[i],'is OK')*

if len(global\_para.IP\_data) <= cnt:

global\_para.IP\_data.extend(tt)

temp\_tt = {}.fromkeys(global\_para.IP\_data).keys()

print 'temp\_tt:',len(temp\_tt)

global\_para.IP\_data= temp\_tt

lock.release()

openercheck.close()

elif para1 == 1:

openercheck.close()

except Exception as e:

lock.acquire()

if para1 == 0:

good\_ip = good\_ip

elif para1 ==1:

if len(global\_para.IP\_data) >= para2:

if good\_ip in global\_para.IP\_error\_list:

global\_para.IP\_error\_list.remove(good\_ip)

if good\_ip in global\_para.IP\_data:

global\_para.IP\_data.remove(good\_ip)

print 'now drop ip:%s' %good\_ip

print 'proxy ip left:%d' %len(global\_para.IP\_data)

else:

global\_para.IP\_error\_list.append(good\_ip)

print 'ip:%s go to IP\_error\_list' %good\_ip

print 'IP\_error\_list:',len(global\_para.IP\_error\_list)

*#print('error')*

lock.release()

def mul\_thread\_get(url\_mul\_get,get\_counter,get\_mode,go\_refresh):

threads = []

for i in range(len(global\_para.IP\_data)):

thread = threading.Thread(target=get\_one, args=[url\_mul\_get,global\_para.IP\_data[i],get\_counter,get\_mode,go\_refresh,])

threads.append(thread)

*# thread.start()*

*#print "new thread start",i*

for thr in threads:

thr.start()

for thread in threads:

thread.join()

if get\_mode==0:

if len(global\_para.IP\_data) >= get\_counter:

print 'ok,get ip done'

return 1

else:

print 'getting ip...'

return 0

elif get\_mode == 1:

if len(global\_para.IP\_data) <= go\_refresh:

return 0

else:

return 1

global\_para.py文件：

1. *#getProxyIP.py ,this file is about global para*

IP\_data = []

IP\_data\_temp = []

IP\_data\_checked = []

findIP = []

findPORT = []

available\_table = []

IP\_error\_list = []

csdn\_url\_cnt = 0

error\_cnt = 0

start\_time = 0L

user\_agents.py文件：保存访问头，每次访问随机选取一个

*#!/usr/bin/python*

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

user\_agents = [

'Opera/9.25 (Windows NT 5.1; U; en)',

'Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.1.4322; .NET CLR 2.0.50727)',

'Mozilla/5.0 (compatible; Konqueror/3.5; Linux) KHTML/3.5.5 (like Gecko) (Kubuntu)',

'Mozilla/5.0 (X11; U; linux i686; en-US; rv:1.8.0.12) Gecko/20070731 Ubuntu/dapper-security Firefox/1.5.0.12',

'Lynx/2.8.5rel.1 libwww-FM/2.14 SSL-MM/1.4.1 GNUTLS/1.2.9'

"Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; AcooBrowser; .NET CLR 1.1.4322; .NET CLR 2.0.50727)",

"Mozilla/4.0 (compatible; MSIE 7.0; AOL 9.5; AOLBuild 4337.35; Windows NT 5.1; .NET CLR 1.1.4322; .NET CLR 2.0.50727)",

]

mul\_thread\_get()函数介绍

*#开启获取模式，当获取到不重复的IP大于等于400个时候，停止获取*

*#特点：多线程获取，获取的IP不重复，不进行可用性检验(没有必要)*

*#返回值：是否完成，当前个数大于等于400，返回1，反之，返回0*

rez = mul\_thread\_get（'代理网站的代理IP页面',400,0,0）

不过需要先获取至少一个能用的代理IP，让程序“起火”，哈哈，不过你也可以将程序稍加修改，提供思路：先用本机IP访问代理网站，获取“启动IP”后，调用mul\_thread\_get即可使global\_para.IP\_data的IP在低于指定值时开启获取，这样在一般情况下，IP是源源不断的，而且不会被网站服务器拉黑。  
使用实例：

while True:

rez = mul\_thread\_get('http://www.\*\*\*.com/page/',400,0,0)*#网址自行脑补，免费的网站就那几个*

if rez == 0:

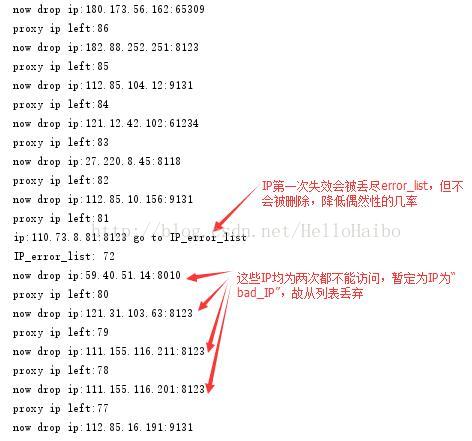
time.sleep(1)

continue

else:

print 'get over!!!!!!!!!'*#此时global\_para.IP\_data里面为获取到的IP*

下面是运行后的输出：

  
当当当当，效果是不是很赞，see u next time~