

【实战】免费代理！

原创：27315



引言

作为一个爬虫开发，最苦恼的事之一肯定是代理ip的问题。今天我们就自己动手来做一个可用的代理IP池。

需求分析

爬取西刺代理网站中可用的高匿代理。

需要源码的同学可以关注公众号，回复“西刺代理”获取源码。

知识点

爬取数据：Requests

数据筛选：Beautifulsoup

数据库：Mongo

主要代码

网站内容很简单，这里就不做过多的解析了。直接放出部分代码

发送requests请求：

```
def get_response(self):  
  
    headers = {  
  
        "User-  
Agent": "Mozilla/5.0 (Windows NT 10.0; Win64; x64) AppleWebKit/537.36 (KHTML, like Gecko) Chrome/68.0.3440.106  
  
    }  
  
    self.response = requests.get(self.url, headers=headers).text
```

提取目标内容：

```
def get_ip_info_list(self):  
  
    soup = BeautifulSoup(self.response, "lxml")
```

```

ip_list = (soup.find(id="ip_list"))

ip_detail = ip_list.find_all(name="tr")

ip_detail = ip_detail[1:]

for ip in ip_detail:

    item = {}

    item['ip'] = ip.find_all(name = "td")[1].string

    item['port'] = ip.find_all(name = 'td')[2].string

    try:

        item['location'] = ip.find_all(name = 'td')[3].find(name = "a").string

    except:

        item['location'] = ip.find_all(name='td')[3].string.strip()

    item['anonymous'] = ip.find_all(name = 'td')[4].string

    item['type'] = ip.find_all(name = 'td')[5].string

    item['speed'] = ip.find_all(name = 'td')[6].find(class_ = 'bar').attrs['title']

    item['connect_time'] = ip.find_all(name = 'td')[7].find(class_ = 'bar').attrs['title']

    item['alive_time'] = ip.find_all(name = 'td')[8].string

    item['verify_time'] = ip.find_all(name = 'td')[9].string

    if self.check_verify_time("20"+item['verify_time'].split(" ")[0]):

        if not check_proxy_duplicate(item):

            yield item

    else:

        return

```

检查ip是否重复:

```

def check_proxy_duplicate(proxy):

    ip = proxy['ip']

    curr = pymongo.MongoClient()

    db = curr['proxy']

    collection = db['proxy']

```

```

ip_exist = collection.find({"ip":ip})

ip_exist_list = []

for i in ip_exist:

    ip_exist_list.append(i)

if ip_exist_list :

    print("%s已经存在"%ip)

    return True

else:

    return False

```

存储至数据库:

```

def save_mongo(item):

    curr = pymongo.MongoClient()

    db = curr['proxy']

    collection = db['proxy']

    collection.insert_one(item)

    curr.close()

    print("%s 存储完成"%(item['ip']))

```

检查代理是否可用:

```

def check_proxy_enable(proxy):

    proxy_string = proxy['type'] + "://" + proxy['ip'] + ":" + proxy['port']

    if proxy['type'] == "HTTP":

        proxy_for_check = {'HTTP':proxy_string}

    elif proxy['type'] == 'HTTPS':

        proxy_for_check = {'HTTPS': proxy_string}

    try:

        requests.get("http://www.sina.com.cn",proxies=proxy_for_check)

    except:

        del_proxy_from_mongo(proxy)

```

```
else:
```

```
update_proxy_from_mongo(proxy)
```

爬取结果

[illegible]

从mongoDB中查看:

```

: "80", "location": "吉林长春", "anonymous": "高匿", "type": "HTTP", "speed":
: "0.153秒", "connect_time": "0.03秒", "alive_time": "29天", "verify_time":
"18-10-29 19:54", "status": "1" }
< "id": ObjectId("5bd6f8fe3e04c71a44356abc"), "ip": "123.134.92.240", "port":
: "53579", "location": "山东莱芜", "anonymous": "高匿", "type": "HTTPS", "spe
ed": "3.32秒", "connect_time": "0.664秒", "alive_time": "27天", "verify_time":
: "18-10-29 19:47", "status": "1" }
< "id": ObjectId("5bd6f8fe3e04c71a44356abf"), "ip": "59.172.27.6", "port": "
53281", "location": "湖北武汉", "anonymous": "高匿", "type": "HTTPS", "speed":
: "1.786秒", "connect_time": "0.357秒", "alive_time": "439天", "verify_time":
: "18-10-29 19:45", "status": "1" }
< "id": ObjectId("5bd6f8fe3e04c71a44356ac2"), "ip": "27.24.215.49", "port":
: "57248", "location": "湖北", "anonymous": "高匿", "type": "HTTPS", "speed":
: "0.14秒", "connect_time": "0.028秒", "alive_time": "20天", "verify_time": "18
-10-29 19:44", "status": "1" }
< "id": ObjectId("5bd6f8fe3e04c71a44356ac5"), "ip": "122.237.104.9", "port":
: "80", "location": "浙江绍兴", "anonymous": "高匿", "type": "HTTP", "speed":
: "0.17秒", "connect_time": "0.034秒", "alive_time": "359天", "verify_time": "
18-10-29 19:40", "status": "1" }
< "id": ObjectId("5bd6f8fe3e04c71a44356ac8"), "ip": "112.85.87.159", "port":
: "53128", "location": "江苏", "anonymous": "高匿", "type": "HTTP", "speed":
: "0.189秒", "connect_time": "0.037秒", "alive_time": "1分钟", "verify_time": "
18-10-29 19:30", "status": "1" }
Type "it" for more
>

```

源码

需要源码的同学可以关注公众号，回复“西刺代理”获取源码。

往期内容推荐

【实战】 下载歌曲只能开绿钻？NoNoNo, Python爬虫，无所不能。

【实战】爬取中国证监会指定信息披露网站——巨潮资讯网

【实战】Ajax数据爬取——头条文章图片

【实战】不知道给女朋友买什么？让爬虫告诉你！

[【福利】美女照片第二弹，Scrapy框架实战应用](#)

[这一定是学爬虫的你脑中闪现过的第一道邪念！](#)

[如果你想学习Python，这些资料你一定要储备](#)

原创不易，需要您的鼓励。如果您觉得这篇文章对您有帮助，请分享给其他需要的小伙伴们。



长按关注公众号

个人创建了一个QQ群，面向群体为爬虫学习，群里许多志同道合的朋友，大家一起进步，可以扫描二维码申请进入，申请备注：爬虫。



群名称：从零开始学爬虫

群 号：780070494

 从零开始学爬虫

最后耽误您几秒钟的时间，点下广告支持下小编。感谢！

- End -

