总

结

文

档

|  |  |  |  |
| --- | --- | --- | --- |
| 版本 | 作者 | 版本描述 | 完成日期 |
| 1.1 |  | 实现腾讯房产数据抓取 | 2017.05.10 |

* + 1. 网页分析

腾讯房产（北京）频道涉及上万条房产数据，既包括在售、待售、在租的房产项目，同时也包括已经售罄的项目；单个项目所含的信息量包括楼盘地址、所属区县、所属商圈、开盘时间、物业类型、产权年限、建筑类别、开发商、预售证、售楼地址、物业公司、物业类别、容积率、绿化率、物业费、停车位等各类信息。这些信息的获得将有利于建立以城市小区生活圈为核心的大数据分析模型。



图 1.1 腾讯房产分类图

针对不同的销售状态,发现其后台通过js提交并返回json数据,请求网址为: <http://db.house.qq.com/index.php>, 所需提交数据如下：

{'mod':'search',

'act':'newsearch',

'city':'bj',

'showtype':showtype,

'unit':'1',

'all':'',

'page\_no':1,

'ND':1}

其中，ND可取1,2,3,4分别代表在售、待售、再租、售罄；page\_No代表返回的页面中页码数，city项可以根据不同的城市，这里bj代表北京市。json返回的数据中可以获得本页中所有房产项目的url，再请求此url获得房产项目自身的详情信息。

在版本1.1中爬取的是总共的楼盘数据，直接从开始的<http://db.house.qq.com/index.php> 这个网站中没有分类直接按页数来爬取，所以不需要那些ND的参数了，其他参数都一样

代理设置

* 1. 代理设置逻辑

有些网站为了防止爬虫或者DDos攻击等，会记录每个IP的访问次数以及访问频率，比如允许一个IP在1s内访问10次等，对于这样一些反爬机制较强以及受法律保护的网站，往往需要通过代理IP的设置来对本地IP进行一定程度的隐匿。简单的说，代理就是换个身份去访问目标网站，而网络身份就是IP地址。

那么，这些代理从何而来呢？对于企业来讲可以直接购买代理IP，但是对于个人而言，这可能会造成资源浪费。好在网上有很多提供免费代理的IP网站，如<http://www.xicidaili.com/nn/1>，我们可以编写一个简单的爬虫先从这些网站上爬IP，然后再用这些代理去爬我们的目标网站。

1. 代理IP获取

本例中采用<http://gatherproxy.com/>作为代理获取网站，该网站还提供了可下载的程序来手动获取代理IP，链接<http://pan.baidu.com/s/1c1JFY9a>，该程序对于免费用户只提供了最大30个线程，大多数时候代理获取速度也是可以接受到。

下面以另一种方法——python脚本的形式获取此网站代理IP，代码见下网站：https://github.com/jerryan999/dianping/tree/master/dianping/utils

代码说明：

a).这里我们使用requests模块来请求url，它是python的一个非官方http库，其API明显优于urllib2，因此具有广泛的运用。

b).由于透明代理以及普通匿名代理会将本地的IP也发送出去，所以，为了最大限度保持匿名性，我们只获取高度匿名的代理IP

c). 使用网站<http://myip.dnsdynamic.org对代理IP>进行一定成功次数测试，如果成功通过测试就证明该代理可用度相对较高。

1. 爬虫代理中间件逻辑

很多时候，代理可能即使通过测试网站的测试，当真正部署到目标网站上却会出现问题；可能测试的时间段内某代理有效，而正使用它的时候却各种错误频出。为了高效的运行爬虫程序，使用代理中间件对目标网站一边爬取页面信息一边进行统计分析。

下面通过引进两个IP优质程度评价指标：绝对净成功次数和相对成功百分率，提供一种代理中间件设置策略。首先，很自然的我们可以想到一个评价指标——相对成功百分比，我们把一个IP比作是一位蝙蝠侠，如果他战斗过的每次战役（访问过的每个url）都可以赢得比赛（返回成功），那么他的成功百分比就是100%；而如果他偶尔输了某几次比赛（非成功返回如网络异常等），那么他的成功百分比必然就要小于100%。所以，我们可以设定当这一百分比低于某一值时（比如80%）时，这一蝙蝠侠需要退场休息（这一代理IP暂时停止使用）。

这一指标的计算方程如下：

式中：

—total request：某一IP总url请求次数

—net success request: 某一IP净请求成功次数

实际操作中，当我们获得一个新IP后，通过它对目标网站的请求成功与否进行统计就可以得到这一百分比。然而，这一算法存在一个较为明显的问题，假设某一IP是高质量的，但是如果第一次请求就失败了的话，其百分比就为0/1=0%,那么它立马就被出局了，而这样高质量的代理就被埋没了。

“为了解决这一问题，我们也许不应该从0开始计算净赢场次”，是的！这里，我们需要进行一个假设：获取来的代理是经过n次测试，测试成功百分比是100%，这样，当发生上述failed情况的时候，百分比就成了(。

下面以n=100来进行详细阐述某代理IP的运行期间各参数情况。

图2.1 某代理IP运行参数

运行期间可以分为三个阶段：

第一阶段，净成功返回请求计数和总请求数同步增加，两者的比值恒定为100%，体现优质代理连续成功返回的特性。

第二阶段，总请求数虽然增加，但是净成功返回请求计数由于这一阶段发生的失败请求而降低，百分比也降低

第三阶段，净成功返回请求计数达到下限（用户自己设定，比如100）或者百分比达到下限（比如80%），该代理暂时停止使用，即需要换别的代理继续请求目标网站。

补充一项重定向的中间件，把所有的重定向URL写入到一个文件中，格式包含从哪个URL重定向到哪个URL中去，同时还包含此次访问用到的IP是什么。

* 1. 异常错误码处理

1. 307异常

使用代理进行爬虫项目的时候，有时候会遇到重定向的问题，正常的重定向会将网址引导到新的url上，而出现问题的代理重定向则引导到一个莫名其妙的网址，比如图1.4显示的某个代理就重定向了一个陌生网页，而发生这一情况的时候，爬虫代理中间件并不认为该代理有问题，依然使用这一代理进行不休止的请求。

解决办法：将307错误排除在爬虫可允许遇到的返回码之外，即无论如何，当返回码为307时将告之此代理失效。

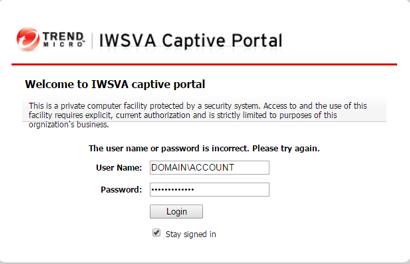


图2.2代理307错误.

1. 404，500，502异常

正常的404错误页面代表客户端在浏览网页时，所访问的页面不存在，即Not Found错误。在使用代理之后，有时候返回404时并非一定是页面没有找到，也有可能是页面正常存在而代理异常返回。如果不对这一情况进行考虑，必然导致目标网站很多url丢失。与此类似，500,502等服务器端发生的错误也存在如上假性错误的现象，也不能单纯的以为返回这些错误就一定是目标网站服务器的问题。

解决方法：针对以上两类问题， 我们增加一个对这些错误的真假判断：如果字符串“腾讯房产”在返回的response中存在即真错误，反之为假错误，此时就需要切换另一代理。

3）403异常

403的处理比较简单，遇到此异常代表所用代理IP被目标网站服务器暂时禁止访问。

解决方法：立即切换代理IP

4）302异常

302重定向URL的处理，会重定向到一个移动端页面，所以就此情况选择从移动端页面中获取数据（个别数据是不完整的，不全）。