总

结

文

档

|  |  |  |  |
| --- | --- | --- | --- |
| 版本 | 作者 | 版本描述 | 完成日期 |
| 1.0 | 安志斌 | 天眼查公司数据抓取 | 2016.12.29 |

* + 1. 网页分析

天眼查是一款提供全面的企业信息查询、专业的企业关系挖掘的工具(图1.1)。可查询企业工商信息、法律诉讼、商标专利、对外投资、招投标、失信、经营异常、企业年报、招聘及新闻动态等。覆盖全国超8000万家企业信息，与[工商局](http://baike.baidu.com/view/815055.htm)网站同步更新。



图1.1 天眼查官方网站

天眼查公司页面的网页结构相对比较规范，数据采集相对较为容易，只需要编写好每条信息对应的css或xpath位置即可，图1.2为其公司页面。



图1.2 天眼查公司网站

* + 1. 代理设置

1. 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。

1. 1. 验证码破解尝试

天眼查网站验证码调用网易出品的极验验证码api(图2.2 验证码页面), 与以往传统验证码不同的是，极验通过分析用户完成拼图过程中的行为特征，通过数据分析来判断是人还是机器。代码编写参考如下网址http://blog.csdn.net/paololiu/article/details/52514504中代码，并做一定的修改（tianyancha/tianyancha/utils/captchaHack.py）



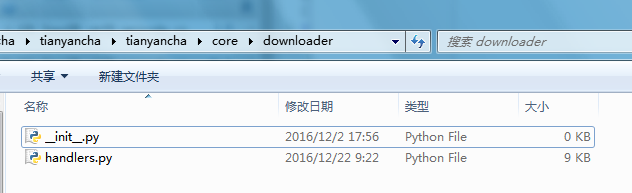
图2.2 天眼查验证码页面

遗憾的是，该破解代码虽然可以勉强通过人机验证，但是测试成功率仍然不太理想，如果需要获得较好的结果，还需要进行大量的鼠标拖动测试并分析如何模拟每次拖动点之间的时间间隔和距离间隔以获近乎真实的机器拖动效果。

* + 1. 技术难点

天眼查网站由于动态加载元素较多，如果细致地分析每个后台api必然工作量较大，而且如果若想获得每次提交所携带的验证字符串，那么从js文件中获取它们也将是不小的工作量。

针对天眼查网站这一情况，我们考虑调用phantomjs来直接执行js代码，具体代码实现见tianyancha/tianyancha/core/downloader/handlers.py



需要注意的是，phantomjs.exe需要提前复制到python/scripts文件夹中

