链滴

爬虫代理

作者：AlwaysBeFriday
原文链接：https://ld246.com/article/1573727694164
来源网站：链滴

为什么需要代理

之所以使用代理，可能是因为：1.加速访问速度；2.隐藏主机真实ip

简单的说，网络通信需要ip地址，ip地址提供主机在网络中的位置，而公共网络ip地址具有唯一性。i可以理解为主机的门牌号，要保证网络信息的快递顺利送达，必须保证门牌号的唯一性。而就因为是有唯一性的，所以直接访问对方站点，对方会被明确的告知(写在请求头中)源地址。

网络爬虫往往并不能告知对方自己的ip，因为站点会在一定程度上容忍爬虫的抓取，只要爬虫保持低不对服务器造成影响，并且不去抓取敏感信息。但实际上，爬虫的速度往往不会很低，并且会收集一敏感信息或者站点不希望被收集的信息，所以反爬一定存在，而最简单的反爬手段，就是封ip，检测法是站点检测到短时间内某ip的访问次数是否超过限制。

当然，真正合理的抓取，是速度不会对服务器造成压力，避免抓取某些信息并使用这些信息获利。

付费代理和免费代理

代理可已购买，也可在网上抓取一些免费代理。相对而言，购买代理质量较高。根据测试(抓取某一目网站年)，买代理的请求成功率的在80%左右，免费代理的请求成功率在40%～50%，当然对于不同站成功率不同，这取决于很多因素，当然最主要的因素是该目标网站被抓取的量级，对于被大量抓取站点，当然会有很多ip被封。

关于代理池的构建和维护

代理池的架构很简单，功能只需要实现：1.抓代理，2.测代理。所以两个功能模块即可。功能模块分，互不影响。

抓代理，分为免费代理网站获取和从付费代理网站获取代理。前者需要找到代理网站，分析页面及接，处理反爬，抓取信息，解析，入库。后者只需要在付费后，通过站点提供的接口，按照一定速度，隔一定时间获取一次并入库即可。抓取代理可能需要从多个代理站点抓取， 可使用多个线程，分别设间隔时间，保持缓慢的速度获取代理

测代理，对代理库中的代理进行定时检测。测试的主要目的不是为了测试代理是否可用，是为了测试理对目标网站是否可用，所以测试需要使用将要抓取的目标网站url。单位时间内测试代理个数需要更需求确定，所以使用单进程异步较为合适，控制好并发数量即可。

问题

python的异步库aiohttp对https的处理有问题，暂时不可用，目标只能处理http请求。当测试量不大，可以使用多线程或线程池。

分布是爬虫可能会需要对目标站点的多个cdn进行抓取，所以如果只检测代理对单个域名ip的可用性而抓取目标站点的cdn，自然会出现问题。所以如果是分布是爬虫，需要对代理池的检测策率进行修。简单的思路是，添加对目标网站的cdn站点ip的检测模块，然后针对每个cdn站点的ip进行检测。