1. 基本名称解释：

url：URL就是网站的网址，是这个网站的简单门牌号。用非人话来说，统一资源定位系统（uniform resource locator;URL）是因特网的万维网服务程序上用于指定信息位置的表示方法。其实是通过网站的ip地址在dns解析之后得到的结果

小拓展：

1.IP地址:IP地址是用来唯一标识互联网上计算机的逻辑地址，让电脑之间可以相互通信. 每台连网计算机都依靠IP地址来互相区分、相互联系

2.域名:由于IP地址是数字标识，使用时难以记忆和书写，因此在IP地址的基础上又发展出一种符号化的地址方案，来代替数字型的IP地址。每一个符号化的地址都与特定的IP地址对应，这样网络上的资源访问起来就容易得多了。这个与网络上的数字型IP地址相对应的字符型地址，就被称为域名。

3.DNS:在Internet上域名与IP地址之间是一对一（或者多对一）的，域名虽然便于人们记忆，但机器之间只能互相认识IP地址，它们之间的转换工作称为域名解析，域名解析需要由专门的域名解析服务器来完成，DNS就是进行域名解析的服务器。域名的最终指向是IP

4.网址:统一资源定位符（URL，英语UniformResourceLocator的缩写）也被称为网址，网址格式为：<协议>://<域名或IP>:<端口>/<路径>

cookie：

Cookie是保存在客户端的纯文本文件，比如txt文件，所谓的客户端就是我们自己的本地电脑，当我们使用自己的电脑通过浏览器进行访问网页的时候，服务器就会生成一个证书并返回给我的浏览器并写入我们的本地电脑，这个证书就是cookie。一般来说cookie都是服务器端写入客户端的纯文本文件。cookies就好像你的一张身份证，你电脑上的cookies和其他电脑上的cookies是不一样的;cookies不能被视作代码执行，也不能成为病毒，所以它对你基本无害。

cookies的作用主要是，当你访问了某些网页，并且对网页的一些设置进行修改，cookies就能跟踪并记录到这些修改，当你下一次访问这个网页的时候，这个网页会分析你电脑上的cookies，进而采取措施像你返回更符合你个性化的网页;当然，目前大部分广告的定位基础也是基于cookies的，比如你此前访问了大量的健身类网站，cookies记录了你的访问行为，广告主就能够根据你的访问行为，向你推送健身类的广告。

当你浏览某网站时，由Web服务器置于你硬盘上的一个非常小的文本文件，它可以记录你的用户ID、密码、浏览过的网页、停留的时间等信息。当你再次来到该网站时，网站通过读取Cookie，得知你的相关信息，就可以做出相应的动作，如在页面显示欢迎你的标语，或者让你不用输入ID、密码就直接登录等等。如果你清理了Cookie，那么你曾登录过的网站就没有你的修改过的相关信息。Cookie是非常常见的， 基本上你的浏览器中都会存储了成百上千条Cookie信息。

html：Hyper Text Markup Language，HTML的全称为超文本标记语言，是一种标记语言。它包括一系列标签．通过这些标签可以将网络上的文档格式统一，使分散的Internet资源连接为一个逻辑整体。HTML文本是由HTML命令组成的描述性文本，HTML命令可以说明文字，图形、动画、声音、表格、链接等。

正则表达式： 这是一种过滤数据的规则，用来在采集时进行数据的提取和替换等操作。其实用人脑来理解就是机器很笨不能理解人类的意图，所以需要把人脑理解的一些命名规则转换给电脑脑来使用。或者是电脑之间沟通方式如果不统一那么把他们的语言翻译让他们能相互理解，也可以用正则表达式来替换。

1. 采集模式分类：

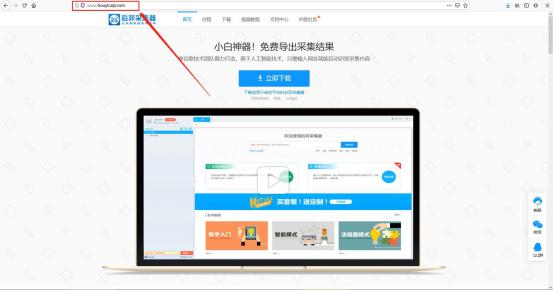
1、智能采集模式：

该模式操作极其简单，只需要输入网址就能智能识别网页中的内容，无需配置任何采集规则就能够完成数据的采集。这种采集模式适合列表类型网页、单页类型网页和列表+详情页类型网页。智能模式比较适合以上三种类型的网页，复杂的网页类型不宜使用该模式，会降低采集对象的识别准确率。

2、流程图采集模式：

完全符合人工浏览网页的思维方式，用户只需要打开被采集的网站，根据软件给出的提示，用鼠标点击几下就能自动生成复杂的数据采集规则。基本可以适用如今大多数网页的要求。

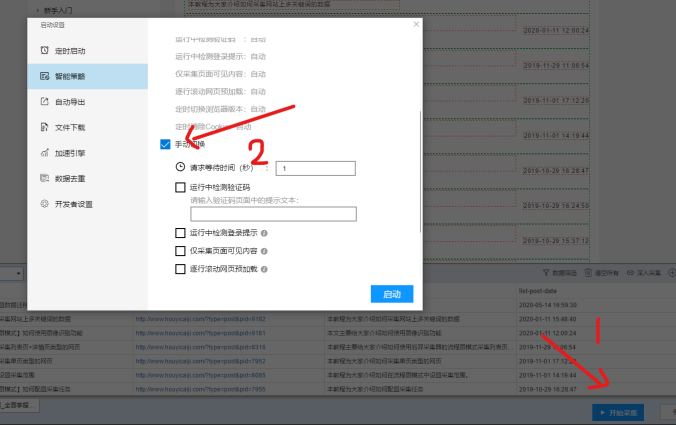
1. 注意账号注册可使你以前的工作可以在登陆其他电脑的后裔采集器的时候依然可以得到使用。（推荐注册账号）
2. 我们需要复制的是展示了采集对象的页面的网址，才能采集到具体内容。

如图：左图可以采集，右图不可。

1. 为何分批导出到多个文件采集结果数量和在浏览器中看到的数量不一致

在采集过程中，如果遇到这个问题，有以下两种可能性：

1. 第一种可能性是采集速度过快而网页加载时间过慢，从而导致无法采集到网页中的数据。遇到这种情况时请增加请求等待时间，等待时间长一点之后，就有足够的时间留给网页加载内容。请求等待时间的设置在 启动设置—>智能策略中，如下图所示：
2. 第二种可能性是采集信息错误，主要在于用户不知道自己采集了那些信息，亦或者是设置爬取的规则不同。这可以通过查看爬取信息是不是整个少了一类爬取对象，或者是看看爬取数据的过程中是不是有报错信息，可以查看爬取日志。
3. 注意网页爬取中三大关键点：

url：首先要清楚你需要的信息列表在那个网址里面体现

结束条件：什么时候结束我们的爬取

采集的内容：你需要的信息到底是标题，视频还是图片？还是文章正文？

这三个明确而且清晰，爬虫就基本不会出问题

7.最后想说的话，其实我虽然有信心教会大家快速上手网络爬取信息，但是大家课后真的需要多了解一些网络知识，特别是关于url命名规则和超文本标记语言的一些规则，这样才能真正从底层了解到这些网站的信息是如何被我们爬取到的。物理学第一原理不仅在物理上适用而且在其他地方也是真正使用的方法。类比远没有从底层理解要清晰深刻。最后祝贺大家完成这一章节的学习

https://s.weibo.com/weibo?q=%23%E4%BD%A0%E5%A5%BD%2C%E6%9D%8E%E7%84%95%E8%8B%B1%23&wvr=6&topsug=1&Refer=SWeibo\_box