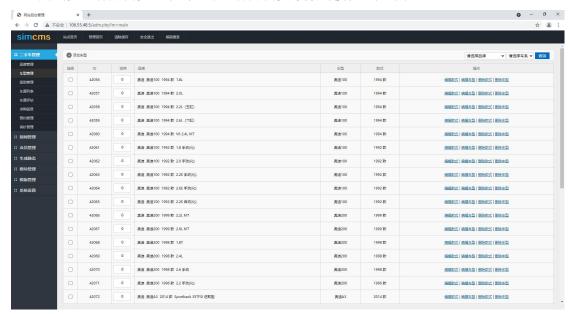
# 阿乐数据采集器 V2.0 使用说明书

# 1.1 软件简介

快速爬取 url 有规律变化的网站,保存静态的 html 文件。如果是后台管理系统或者网站中包含表格,可以自动将表格保存至 csv 文件。同时也可以爬取前端渲染类网站的接口,将 json 格式的数据,转化保存至 csv 文件。

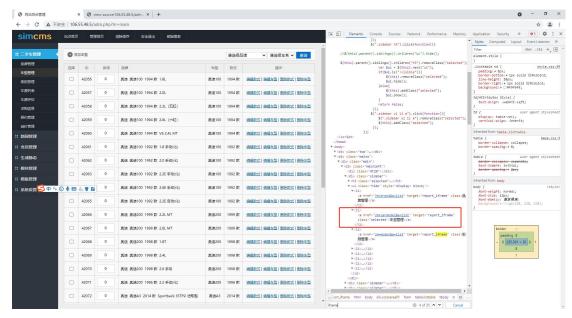
# 1.2 使用说明

一、用浏览器打开需要爬取的网站,打开需要采集的模块。如下图

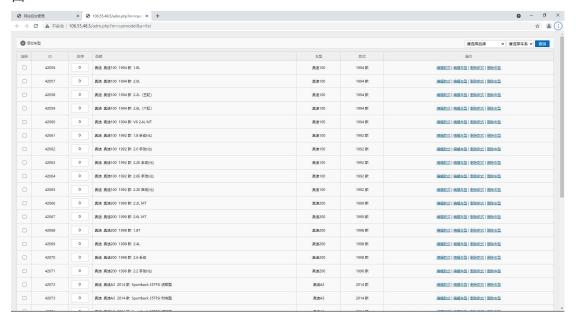


二、如果有 iframe 框架,需要去掉框架,即打开真实的链接。这里有两种方法,第一种可以通过右键表格区域,查看框架源代码,这里以软件内置的绿色版谷歌为例,如下图

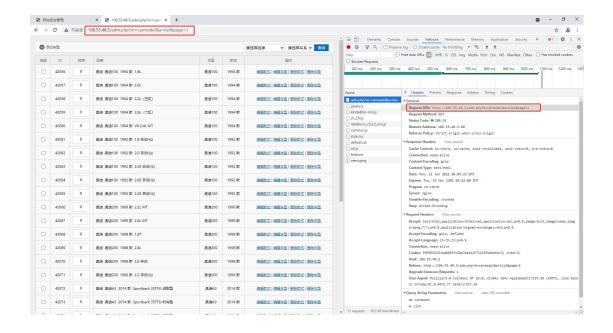
去掉地址栏中的 view-source:即可打开真实的链接。如果是火狐浏览器则更方便,右键点击表格区域,点击此框架,再点击新建标签页打开框架。另一种方式是通过 F12 去查看网页源代码,搜索 iframe,如下图所示

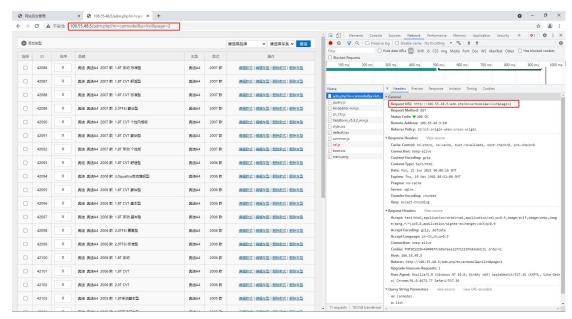


右键点击找到的链接,选择 Open in new tab(在新的标签页中打开),打开后如下图



三、在新打开的标签页中,右键点击检查或者按 F12,接着点击 Network。开始前可以 先点击 clear 清除之前的一些内容。我们通过点击翻页按钮,来观察 Network 中请求的 变化,如图

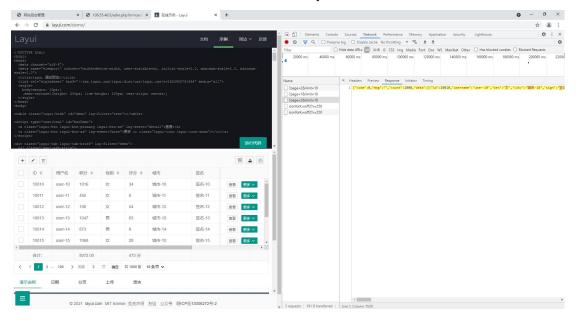


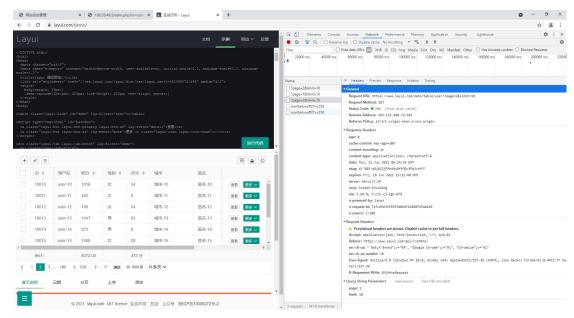


这里不断变化的是 page 后面的数字,打开阿乐数据采集器 V2.0.exe,将http://106.55.48.5/adm.php?m=carmodel&a=list&page=1 中&前面的部分填入 url 的框中,将 page 填入 variate(变量)中,在第三项 page(页数)中填入需要爬取的页数。如下图

		×
### ### ### ### ### ### ### ### ### ##		

如果是前端渲染类的网站,地址栏中 url 是不变的,在 Network 中可以看到翻页时,请求的 url 也是有规律的变化,服务器返回来的是 json 格式的数据,如下图

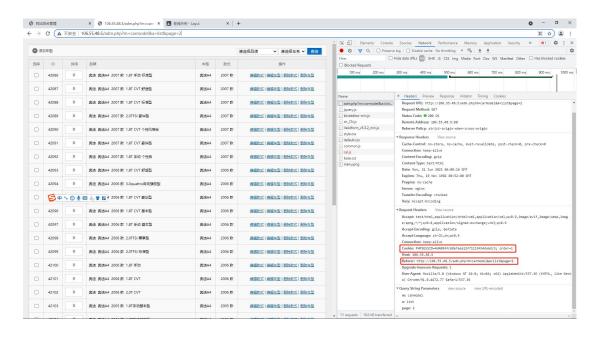




这类的网站,填写方式和前面是一样的,去掉中间变化的部分填入 url,参数填到 variate 中,如图

	om	=	×
	url(填写url不变化的部分*)		
ht	ttps://www.layui.com/demo/table/user/?limit=10		
	variate (填写url使用&连接时的参数名)		
To.	age (AE-201100110000000111038-80001)		
p.	page(填写爬取的页数 *)		
10			
110			
	cookie(填写分号分隔的键值对)		
	referer(填写来源页 留空默认等于url)		
	user-agent(浏览器标识 醫空默认随机ua)		
	线程数		
	4 連次延时		
	0		
	扩展名		
	元扩展名		
	显示浏览器		
	ē ■		
	截倒数量		
	不截图		
	The state of the s		
	start		
	log		

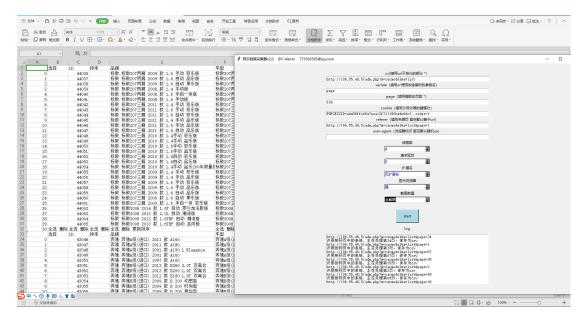
四、其他内容的填写。如果是网站后台,必须要填写 cookie。有些网站必须要填写 referer。这些内容也都在 Network 中可以找到,如下图填写



http://106.55.48.5/ada.php?n=carnodel&n=list	乐数据采集器v2.0 BY: Alenm 77359.	soop@dd.com	_	
http://l06.65.48.5/adm.php?m=carnodel&a=list		url/信写url 太恋// 的部分 *)		
page			-	
page (填写爬即的页数 *)  100				
page (填写配配的页数 *)   100				
Cookie (填写分写分隔的键值对)				
cookie(捷馬子等分隔的體值对) PHPSESSID=44408ffc68a7eaa12f71119fb4de6dc3: order=1 referer(捷馬来渡回 智致以等于中) http://106.55.48.5/adm.php?m=carmodel&n=list&page=1 user-agent (郑茂縣宗川 新空數以經析以由)				
PHPSESSID=44408ffc68a7eea12f71119fb4de6dc3; order=1 referer (馬馬来原贝爾亞默以達于中) http://106.55.48.5/adm.php?me-carnodelaw-listebages11 user-agent (海底器房贝留空默认随机由)				
http://106.55.48.5/adm.php?m=carmodel&m=list&page=  user-agent (知思語示以 篇至默认随例ua)    集程数			<del></del>	
http://106.55, 48.5/ada, plp?m=caraodelaw=lst&page=l    user-agent (知茂縣宗) 醫空數以經行與   接程数				
接程数			<del></del>	
接程数		user-agent (浏览器标识 留空默认随机ua)		
事來說时  D  F  F  F  F  F  E  E  S  S  S  S  S  S  S  S  S  S  S				
请求延时  0		线程数		
请求延时  0		4		
D				
扩展名  无扩展名  显示浏览器  否  截图数量  不截图				
医广测范器 否  数图数量  不新图  start				
量示測透器 否 執密数量 不新器				
图 數图数量 不數图 • start				
載函数量 不載型 start				
不載證 start				
start				
		不截图		
log		start		
log				
		log		

五、根据网络、网站性能调整线程数等设置。勾选截图会调用绿色版谷歌浏览器截图,会影响爬取速度。如果网站有.html 或者.jsp 等扩展名,则需要勾选相应的扩展名。

# 1.3 效果截图及性能



我自己测试的网站,PHP做的cms,里面有表格。服务器配置是腾讯云轻量级服务器,2核2g内存,带宽5m。网络是电信300m。

#### Speed:

单线程,爬取 800 页,处理表格,不截图,用时 130s 单线程,爬取 800 页,处理表格,全截图,用时 280s 四线程,爬取 800 页,处理表格,不截图,用时 40s 四线程,爬取 800 页,处理表格,全截图,用时 108s 八线程,爬取 800 页,处理表格,不截图,用时 36s 八线程,爬取 800 页,处理表格,全截图,用时 57s

# 2.1 其他正在开发的功能

- 一、浏览器输入网址,自动提取 url, cookie, referer 等。
- 二、爬取的日志,以及 html 文件的哈希值。
- 三、post 请求网站的爬取。
- 四、动态加载网站的爬取和保存。
- 五、多个参数变化的网站爬取。
- 六、服务器端,分布式爬取。
- 七、页数选择,从中间的页数开始提取。
- 八、代理,加入ip代理池。
- 九、文件锁,爬取网站时,打开 csv 文件用只读,不会使爬虫停止。
- 十、UI 太丑了,研究一下 qt5 重新做一下。

# 2.2 项目地址

GitHub: https://github.com/zmzmon/CJ Gitee: https://gitee.com/saa1028/cj