Web Scraper爬虫工具

笔记本：	统计		
创建时间：	2020/5/27 15:54	**更新时间：**	2020/5/28 13:50
作者：	190072358@qq.com		

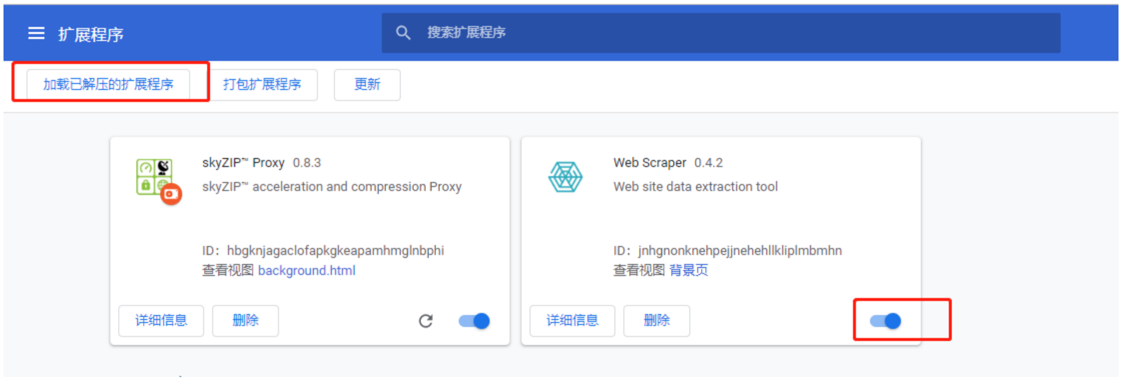
一、安装

百度网盘链接：https://pan.baidu.com/s/1vKtCFtVUA4q_bq3OFLJl2w

提取码：owf8

1、下载后，先解压。

2、goole浏览器打开扩展程序-加载已解压的扩展程序



skyZIP™ Proxy 0.8.3
skyZIP™ acceleration and compression Proxy
ID：hbgknjagaclofapkgkeapamhmglnbphi
查看视图 background.html

Web Scraper 0.4.2
Web site data extraction tool
ID：jnhgnonknehpejjnehehllkliplmbmhn
查看视图 背景页

二、使用

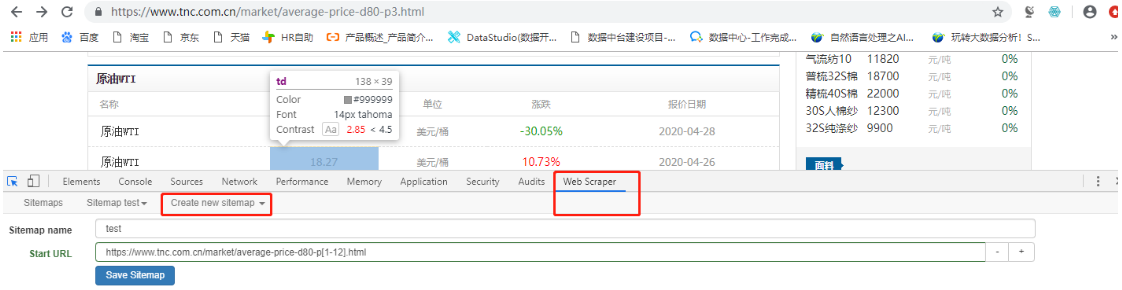
1、F12键，可以快速打开web scraper

2、create new sitemap

sitemap name 可以随便取个名字

start url 爬数据的网址，这里爬取的表格有多页，url写成：

https://www.tnc.com.cn/market/average-price-d80-p[1-12].html



https://www.tnc.com.cn/market/average-price-d80-p3.html

原油WTI

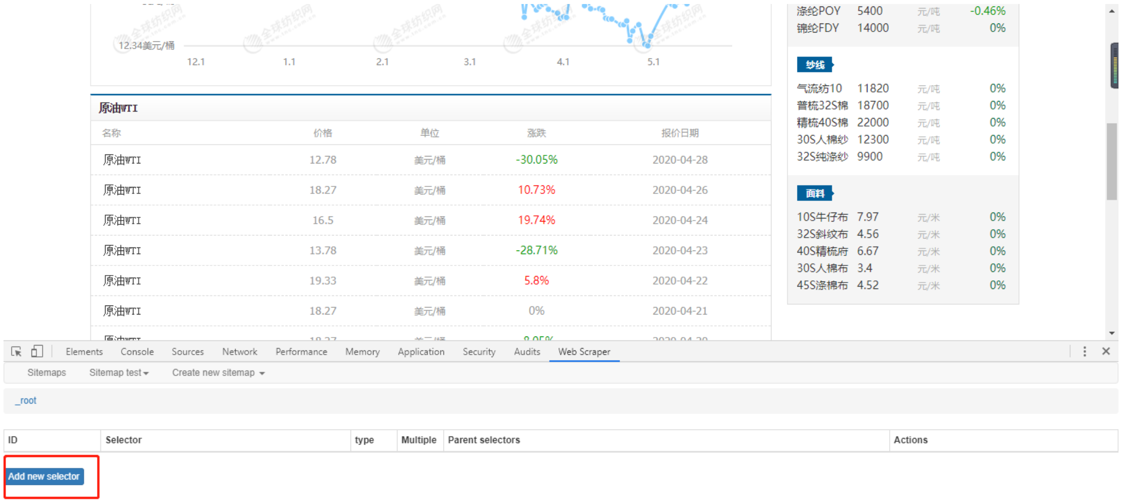
名称		单位	涨跌	报价日期
原油WTI		美元/桶	-30.05%	2020-04-28
原油WTI	18.27	美元/桶	10.73%	2020-04-26

气流纺10	11820	元/吨	0%
普梳32S棉	18700	元/吨	0%
精梳40S棉	22000	元/吨	0%
30S人棉纱	12300	元/吨	0%
32S纯涤纱	9900	元/吨	0%

Sitemap name: test

Start URL: https://www.tnc.com.cn/market/average-price-d80-p[1-12].html

3、add new selector



原油WTI

名称	价格	单位	涨跌	报价日期
原油WTI	12.78	美元/桶	-30.05%	2020-04-28
原油WTI	18.27	美元/桶	10.73%	2020-04-26
原油WTI	16.5	美元/桶	19.74%	2020-04-24
原油WTI	13.78	美元/桶	-28.71%	2020-04-23
原油WTI	19.33	美元/桶	5.8%	2020-04-22
原油WTI	18.27	美元/桶	0%	2020-04-21

涤纶POY	5400	元/吨	-0.46%
锦纶FDY	14000	元/吨	0%

纱线

气流纺10	11820	元/吨	0%
普梳32S棉	18700	元/吨	0%
精梳40S棉	22000	元/吨	0%
30S人棉纱	12300	元/吨	0%
32S纯涤纱	9900	元/吨	0%

面料

10S牛仔布	7.97	元/米	0%
32S斜纹布	4.56	元/米	0%
40S精梳府	6.67	元/米	0%
30S人棉布	3.4	元/米	0%
45S涤棉布	4.52	元/米	0%

ID	Selector	type	Multiple	Parent selectors	Actions

4、编辑抓取字段

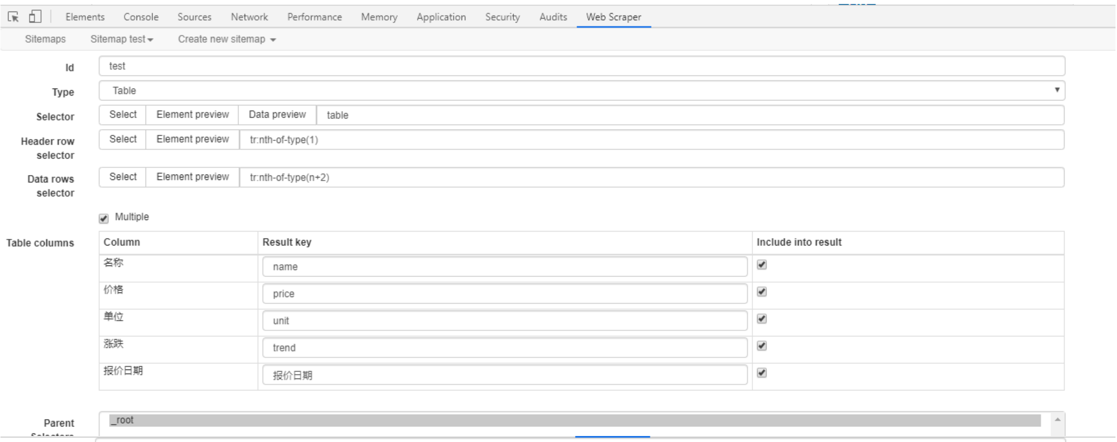
id可以随便取个

type这里选择table，因为抓取的是表格数据

select点击之后可以选择要抓取的表格，剩下的内容会自动填充

注意勾选Multiple，否则只会导出第一行

Result Key是中文保存时可能报错，改为英文即可。



Id: test

Type: Table

Selector: Select | Element preview | Data preview | table

Header row selector: Select | Element preview | tr:nth-of-type(1)

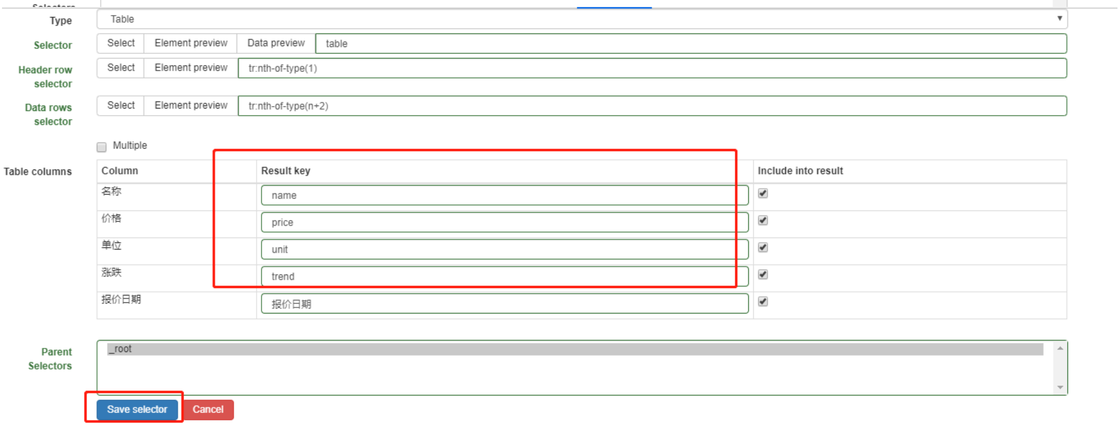
Data rows selector: Select | Element preview | tr:nth-of-type(n+2)

☑ Multiple

Table columns

Column	Result key	Include into result
名称	name	☑
价格	price	☑
单位	unit	☑
涨跌	trend	☑
报价日期	报价日期	☑

Parent Selectors: _root



Type: Table

Selector: Select | Element preview | Data preview | table

Header row selector: Select | Element preview | tr:nth-of-type(1)

Data rows selector: Select | Element preview | tr:nth-of-type(n+2)

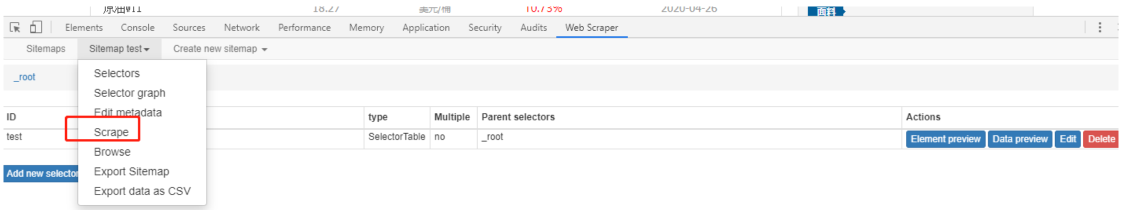
☐ Multiple

Table columns

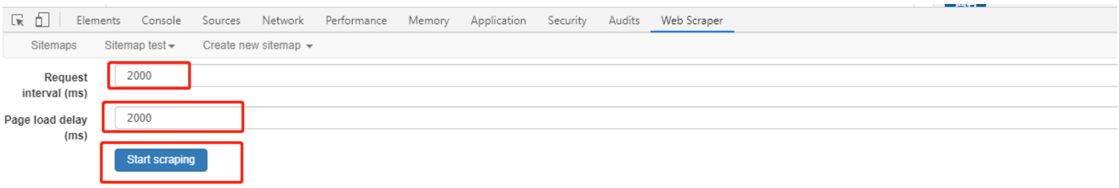
Column	Result key	Include into result
名称	name	☑
价格	price	☑
单位	unit	☑
涨跌	trend	☑
报价日期	报价日期	☑

Parent Selectors: _root

5、Scrape



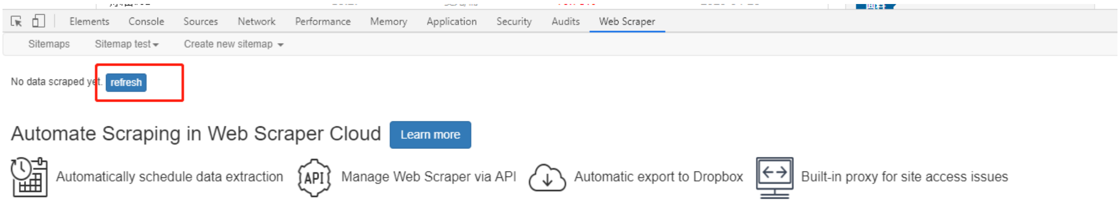
ID	type	Multiple	Parent selectors	Actions
test	SelectorTable	no	_root	Element preview Data preview Edit Delete



Request interval (ms) 2000

Page load delay (ms) 2000

6、refresh



No data scraped yet.

Automate Scraping in Web Scraper Cloud

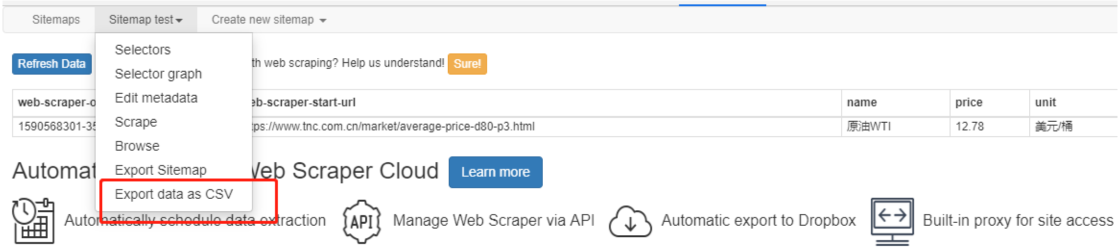
Automatically schedule data extraction

Manage Web Scraper via API

Automatic export to Dropbox

Built-in proxy for site access issues

7、导出



th web scraping? Help us understand!

web-scraper-o	eb-scraper-start-url	name	price	unit
1590568301-35	ps://www.tnc.com.cn/market/average-price-d80-p3.html	原油WTI	12.78	美元/桶

Automat Veb Scraper Cloud

Automatically schedule data extraction

Manage Web Scraper via API

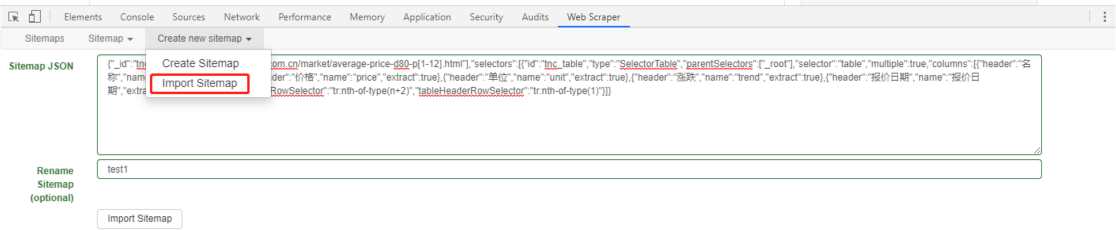
Automatic export to Dropbox

Built-in proxy for site access

三、直接导入别人写好的sitemap

1、单个导出，只导出ACN的数据。

{"_id":"tnc","startUrl":["https://www.tnc.com.cn/market/average-price-d90-p[1-12].html"],"selectors":[{"id":"tnc_table","type":"SelectorTable","parentSelectors":["_root"],"selector":"table","multiple":true,"columns":[{"header":"名称","name":"name","extract":true},{"header":"价格","name":"price","extract":true},{"header":"单位","name":"unit","extract":true},{"header":"涨跌","name":"trend","extract":true},{"header":"报价日期","name":"报价日期","extract":true}],"delay":0,"tableDataRowSelector":"tr:nth-of-type(n+2)","tableHeaderRowSelector":"tr:nth-of-type(1)"}]}

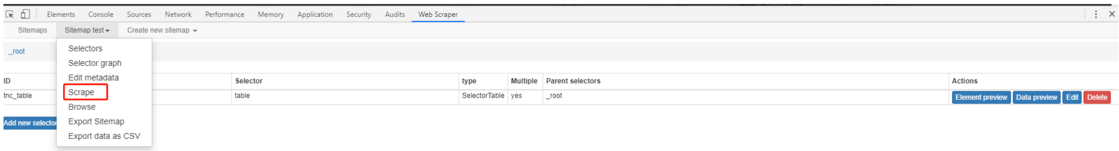


Sitemap JSON

{"_id":"tn om.cn/market/average-price-d80-p[1-12].html"],"selectors":[{"id":"tnc_table","type":"SelectorTable","parentSelectors":["_root"],"selector":"table","multiple":true,"columns":[{"header":"名
称","nam der":"价格","name":"price","extract":true},{"header":"单位","name":"unit","extract":true},{"header":"涨跌","name":"trend","extract":true},{"header":"报价日期","name":"报价日
期","extra RowSelector":"tr:nth-of-type(n+2)","tableHeaderRowSelector":"tr:nth-of-type(1)"}]}

Rename Sitemap (optional): test1

Import Sitemap



ID	Selector	type	Multiple	Parent selectors	Actions
tnc_table	table	SelectorTable	yes	_root	Element preview Data preview Edit Delete

2、批量导出

```
{"_id":"tnc-data","startUrl":["https://www.tnc.com.cn/market/average-price.html"],"selectors":[{"id":"links","type":"SelectorLink","parentSelectors":["_root"],"selector":"td div a","multiple":true,"delay":0},{"id":"page-link","type":"SelectorLink","parentSelectors":["links","page-link"],"selector":".page a","multiple":true,"delay":0},{"id":"data","type":"SelectorTable","parentSelectors":["page-link"],"selector":"table","multiple":true,"columns":[{"header":"名称","name":"name","extract":true},{"header":"价格","name":"price","extract":true},{"header":"单位","name":"unit","extract":true},{"header":"涨跌","name":"trend","extract":true},{"header":"报价日期","name":"date","extract":true}],"delay":0,"tableDataRowSelector":"tr:nth-of-type(n+2)","tableHeaderRowSelector":"tr:nth-of-type(1)"}]}
```

参考文档：

https://www.bilibili.com/video/BV1LJ411t7Gi/?spm_id_from=333.788.videocard.2