

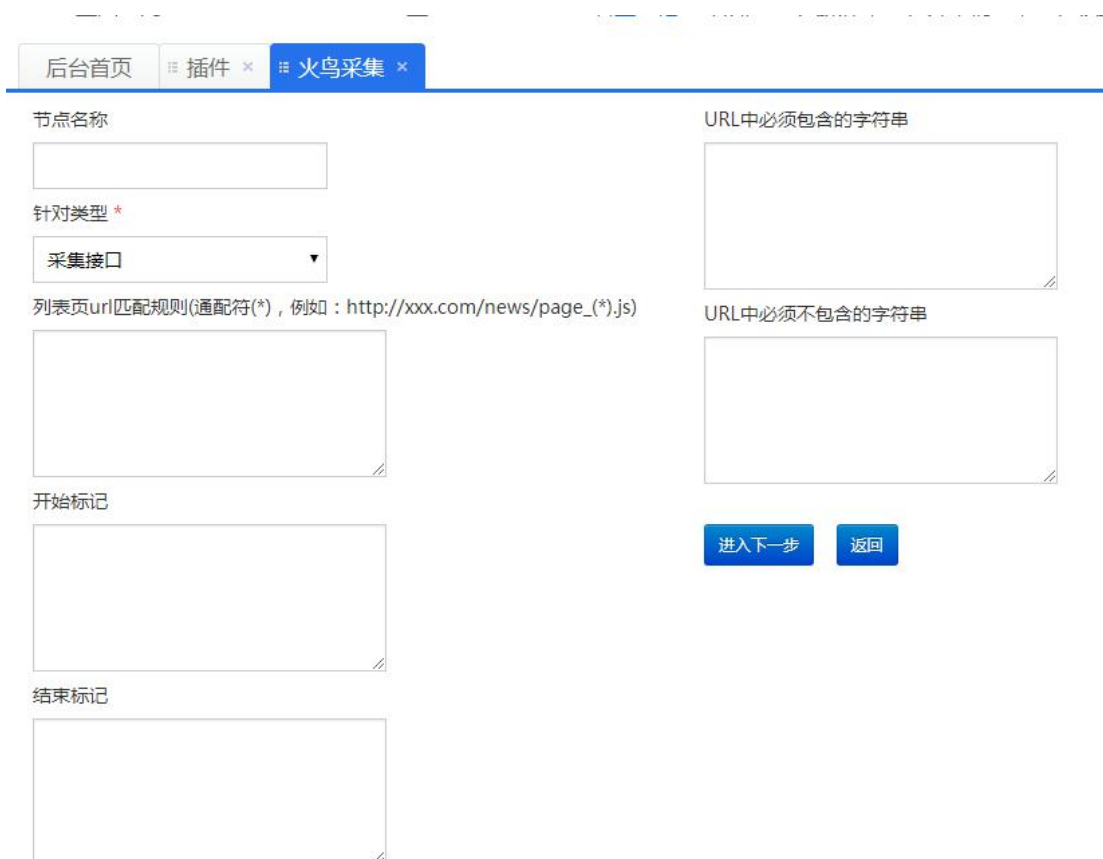
信息资讯通用采集插件 (使用手册)

[Huoniao.Plugins]

一、添加任务



点击添加任务之后出现以下界面：



1. 节点名称可以自己根据要采集的内容进行填写
2. 针对类型有三种
采集单个界面：就是采集单独的一个 HTML 页面
采集多个界面：
例如：
<https://ihuoniao.cn/sz/article/yule/?page=1>
<https://ihuoniao.cn/sz/article/yule/?page=2>
[https://ihuoniao.cn/sz/article/yule/?page=\(*\)](https://ihuoniao.cn/sz/article/yule/?page=(*))
所以类似这种可以匹配页码规则的都为此类型。

采集接口类型:

以下例子为网易新闻的列表页面，数据是通过

http://temp.163.com/special/00804KVA/cm_guonei_03.js?callback=data_callback

http://temp.163.com/special/00804KVA/cm_guonei_02.js?callback=data_callback

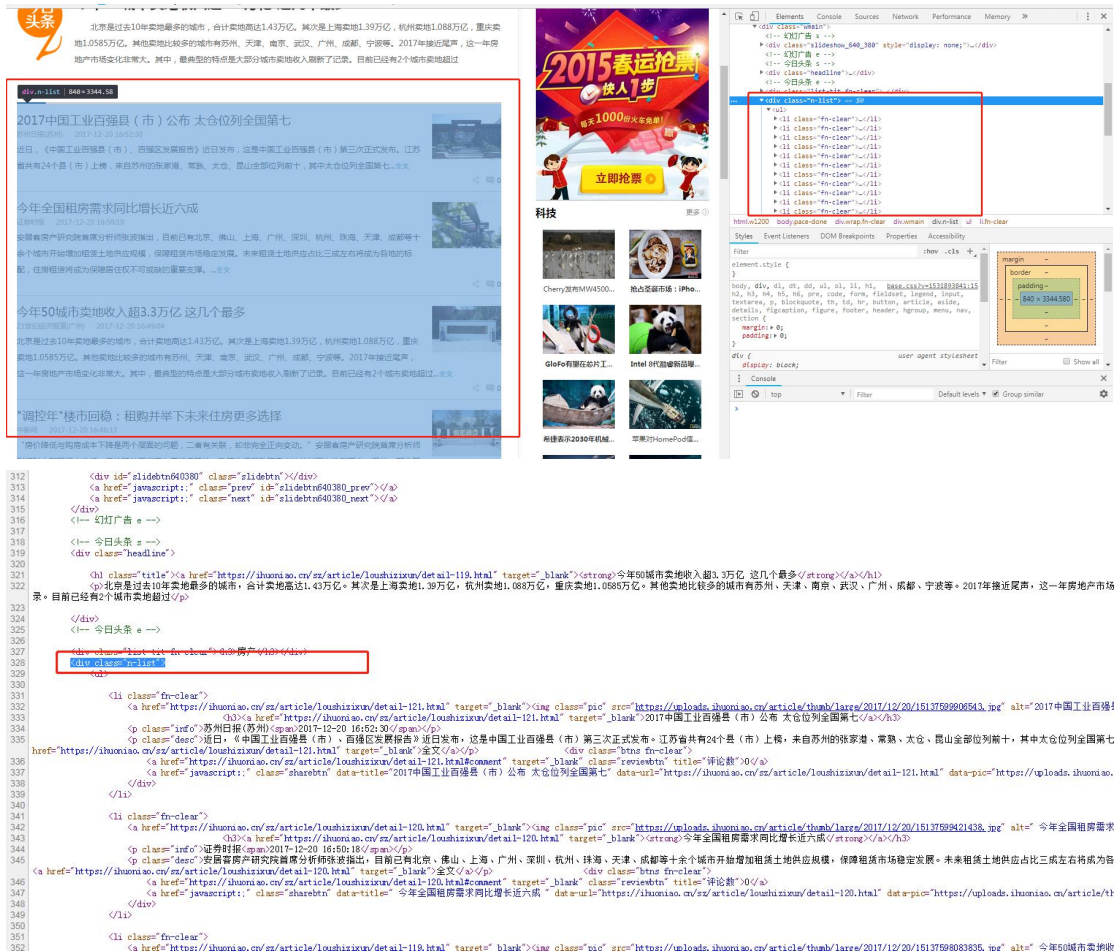
等接口获取的。若没有出现类似的接口，则说明数据与页面是同步的，则选择针对类型为网页 HTML



3. 第三个输入框根据选择的针对类型填写通配符统一填写 (*)

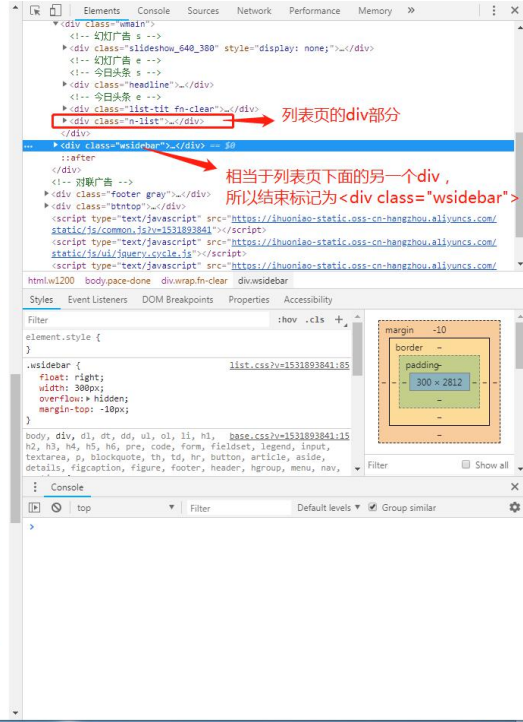
4. 开始标记为页面列表部分 div 的开始。

采集网页 HTML 类型:



需要右键查看源代码，复制开始标记: <div class="n-list">

5. 结束标记



采集接口类型:

```
data_callback([
```

开始标记尽量为左中括号之前的全部内容

```
{
  "title": "数据堂倒卖公民信息0.2元/条买病例 百度是大客户",
  "digest": "",
  "docurl": "http://news.163.com/18/0716/02/DMQ68MHS00018AOP.html",
  "commenturl": "http://comment.news.163.com/news2_bbs/DMQ68MHS00018AOP.html",
  "tium": 355,
  "tlastid": "",
  "tlink": "http://news.163.com/18/0716/02/DMQ68MHS00018AOP.html",
  "label": "其它",
  "keywords": [
```

```
["akey_link": "http://news.163.com/keywords/4",
["akey_link": "http://news.163.com/keywords/0
```

```
"time": "07/16/2018 00:00:00",
```

所以开始标记为 data_callback(结束标记为

```
["akey_link": "http://news.163.com/keywords/4/8/4e8b52a190e8/1.html", "keyna
["akey_link": "http://news.163.com/keywords/4/2/4e24592e519b59d4/1.html", "keyname": "中央军委"}
["akey_link": "http://news.163.com/keywords/5/5/5b597e0d9a8b/1.html", "keyname": "孙绍骋"}
],
"time": "07/16/2018 09:55:45",
"newstype": "article",
"pics3": [
],
"chaoneiname": "guomei",
"imgurl": "http://cms-bucket.nosdn.127.net/2018/07/14/b4c9bbf0df940b9a586a4bf73f6c42.png",
"add1": "",
"add2": "",
"add3": ""
])
```

结束标记尽量为右中括号右边的所有内容, 此处为)

6. 必须包含为此列表页包括的所有新闻内页的 url 中共有的部分 (只要是公共部分就行,

也可直接填写网页后缀)

```
{
  "title": "云南整治违建"豪华墓" "活人墓" 打击捆绑收费行为",
  "digest": "",
  "docurl": "http://news.163.com/18/0714/13/DMM9OSQNO001875N.html",
  "commenturl": "http://comment.news.163.com/news_guonei8_bbs/DMM9OSQNO001875N.html",
  "titem": 2,
  "tlastid": "<a href='http://news.163.com/'>新闻</a>",
  "tlink": "http://news.163.com/18/0714/13/DMM9OSQNO001875N.html",
  "label": "其它",
  "keywords": [

    {"akey_link": "http://news.163.com/keywords/"}
  ],
  "time": "07/14/2018 11:36:00",
  "newstype": "article",
  "pics3": [
  ],
  "channelname": "guonei",
  "imgurl": "http://cms-bucket.nosdn.127.net/2018/07/14/07da5758c8ae474882004c9641b37ba2.png",
  "add1": "",
  "add2": "",
  "add3": ""
}

{
  "title": "台湾水果滞销农民只能贱卖或作废 连战向汪洋请命",
  "digest": "",
  "docurl": "http://news.163.com/18/0714/11/DMMOUVWV10001875N.html",
  "commenturl": "http://comment.news.163.com/news_guomei8_bbs/DMMOUVWV10001875N.html",
  "titem": 10903,
  "tlastid": "",
  "tlink": "http://news.163.com/18/0714/11/DMMOUVWV10001875N.html",
  "label": "其它",
  "keywords": [

    {"akey_link": "http://news.163.com/keywords/"}
  ]
}
```

经过验证这两条url为新闻详细内容页面，所以必须包括
填写http://news.163.com/18 (此处可以直接写后缀也可以
比如：.html)

7. 必须不包含一般为列表页面中的一些图片的链接或者关键词评论的链接，填写之后采集时会过滤这些内容。

最终效果：

后台首页 插件 × 火鸟采集 ×

节点名称
火鸟

针对类型 *
采集多个页面

列表页url匹配规则(通配符*), 例如: http://xxx.com/news/page_(*).js
https://ihuoniao.cn/sz/article/yule/?page=(*)

URL中必须包含的字符串
article

URL中必须不包含的字符串
.jpg

开始标记
<div class="n-list">

结束标记
<div class="pagination">

进入下一步 返回

二、点击进入下一步添加新闻详细页面的匹配规则

正文开始: <div class="post_text" id="endText" style="border-top:1px solid #ddd;">

正文结束: <div class="post_btmsahre">

来源开始: <div class="ep-source cDGray">

来源结束标记:

以上为举例。所有的结束标记都是匹配到开始标记之后遇到的第一个结束标记为止。

最终效果:

采集节点	58
采集列表页	https://ihuoniao.cn/sz/article/fangchan/

下一步

点击下一步进入采集页面。

页面下方为匹配到的新闻链接。

输入每页采集条数和间隔时间

点击开始采集。

以下为最终效果。

节点名称	火鸟
每页采集	10 条
间隔时间	1 秒

开始采集 查看已采集

当前已完成: 32%

获取到的种子网址：

https://ihuoniao.cn/sz/article/loushizixun/detail-121.html
https://ihuoniao.cn/sz/article/loushizixun/detail-120.html
https://ihuoniao.cn/sz/article/loushizixun/detail-119.html
https://ihuoniao.cn/sz/article/loushizixun/detail-118.html
https://ihuoniao.cn/sz/article/loushizixun/detail-117.html
https://ihuoniao.cn/sz/article/loushizixun/detail-116.html

采集完成之后可以点击查看已采集内容：

火鸟门户

新闻采集 插件管理 新闻分类 导出新闻 管理新闻 新闻站 新闻管理 模板管理 / 管理 友情链接

后台首页 插件 × 火鸟采集 ×

已抓取新闻列表

新闻站 发布内容

节点	新闻正文	新闻标题	来源	作者	发布时间	操作
58	> <h1 class="h1">买车位的五大好处从这告别停车难	买车位的五大好处从这告别停车难	网易房产苏州	李文	1513758897	删除
58	> <h1 class="h1">19楼的电梯, 楼的整层电梯信息不明确海澜资讯	楼的整层电梯信息不明确海澜资讯	网易房产苏州	张瑾	1513758973	删除
58	> <h1 class="h1">2万方住宅带, 万方住宅带地所幼儿园翰园西片区	万方住宅带地所幼儿园翰园西片区	名城苏州	李文	1513759181	删除
58	> <h1 class="h1">"调控年"楼市, 调控年楼市调控措施将带来什么	调控年楼市调控措施将带来什么	中新网	李文	1513759693	删除
58	> <h1 class="h1">2017中国工, 中国工业百强县市公布大位位列佳	中国工业百强县市公布大位位列佳	苏州日报苏州	张瑾	1513759950	删除
58	> <h1 class="h1">今年50城市, 今年城市财政收入超万亿这几个	今年城市财政收入超万亿这几个	世纪经济报道广州	李文	1513759744	删除

发布内容按钮可以把采集到的内容发布到网站的新闻中心。

新闻采集 插件管理 新闻分类 导出新闻 管理新闻 新闻站 新闻管理 模板管理 / 管理 友情链接

后台首页 插件 × 火鸟采集 ×

发布新闻

当前可导出条数：	6
选择城市：	苏州
导出分类：	综艺
每批导出数量	20 条
每批导出间隔	1 秒

开始导出 返回主界面

同样输入每批导出数量和时间间隔。

导出完成之后，则可以在新闻管理界面看到采集的内容。