

- 爬取地址
  - <https://www.huxiu.com/>
  - 分析
    - 页面分析

Name	X	Headers	Preview	Response	Cookies	Timing
<input type="checkbox"/> pv?stm=1543992725092						
<input type="checkbox"/> recall?requestID=epvqm1q9rl...						
<input type="checkbox"/> send_log?data=%257B%252...						
<input type="checkbox"/> action?stm=1543992727822						
<input type="checkbox"/> action?stm=1543992733391						
<input checked="" type="checkbox"/> article_list						
<input type="checkbox"/> action?stm=1543992733938						
<input type="checkbox"/> action?stm=1543992737153						
<input type="checkbox"/> log?clientToken=48ab7b2e18...						
<input type="checkbox"/> log?clientToken=48ab7b2e18...						

  

Name	X	Headers	Preview	Response	Cookies	Timing
<input type="checkbox"/> pv?stm=1543892010489						
<input type="checkbox"/> recall?requestID=ggkyjg2x1d...						
<input type="checkbox"/> send_log?data=%257B%252...						
<input type="checkbox"/> action?stm=1543892014639						
<input type="checkbox"/> action?stm=1543892018731						
<input type="checkbox"/> action?stm=1543892023425						
<input checked="" type="checkbox"/> article_list						
<input type="checkbox"/> action?stm=1543892025050						
<input type="checkbox"/> log?clientToken=48ab7b2e18...						
<input type="checkbox"/> log?clientToken=48ab7b2e18...						
<input type="checkbox"/> action?stm=1543892028761						

  

▼ Form Data      view source      view URL encoded

huxiu\_hash\_code: 88038741b16b49cdf3eeba2f4da71de7

page: 2

last\_dateline: 1543826564

- 爬取目标数据

```

..
  <div class="mod-b mod-art clearfix" data-aid="274905"> == $0
  ::before
  <!--栏目链接-->
  <div class="mod-thumb pull-left">
    <a class="transition" href="/article/274905.html" target="_blank">
      <!--文章图片-->
      
    </a>
  </div>
  <div class="mob-ctt index-article-list-yh">
    <h2>
      <a href="/article/274905.html" class="transition msubstr-row2" target="_blank">百年前的世界和今日的世界</a>
    </h2>
    <div class="mob-author">
      <div class="author-face"></div>
      <a href="/member/2147218.html" target="blank">
        <span class="author-name">经济观察报观察家</span>
      </a>
      <a href="/vip" target="blank"></a>
      <span class="time">19小时前</span>
      <i class="icon icon-cmt"></i>
      <em>13</em>
      <i class="icon icon-fvr"></i>
      <em>84</em>
    </div>
    <!-- 外部文章/内部文章两者取其 -->
    <div class="mob-sub">
      推倒全球治理体系会让世界更好吗? </div>
  </div>
  <!--文章tag展示-->
  <div class="column-link-box"></div>
  ::after

```

## • 爬取工具

### • pyspider

#### • 架构

- Scheduler (调度器)
- Fetcher (抓取器)
- Processor (处理器)
- Monitor (监控器) : 监视整个爬取过程
- Result WOrder (结果处理器) : 处理最后抓取结果

#### • 大致流程

- 一个 pyppider 爬虫项目对应一个 Python 脚本，脚本里定义了一个 Handler 主类。爬取时首先调用 on\_start() 方法生成最初的抓取任务，然后发送给 Scheduler。
- Scheduler 将抓取任务分发给 Fetcher 进行抓取，Fetcher 执行然后得到 Response、随后将 Response 发送给 Processor。
- Processor 处理响应并提取出新的 URL 然后生成新的抓取任务，然后通过消息队列的方式通知 Scheduler 当前抓取任务执行情况，并将新生成的抓取任务发送给 Scheduler。如果生成了新的提取结果，则将其发送到结果队列等待 Result Worker 处理。
- Scheduler 接收到新的抓取任务，然后查询数据库，判断其如果是新的抓取任务或者是需要重试的任务就继续进行调度，然后将其发送回 Fetcher 进行抓取。
- 不断重复以上工作、直到所有的任务都执行完毕，抓取结束。
- 抓取结束后、程序会回调 on\_finished() 方法，这里可以定义后处理过程。

#### • 目录结构分析

- data目录 (文件路径就放在你第一次运行pyspider命令后的位置，里面有几个SQLite文件)
  - project.db : 俺村了用户的爬虫项目相关信息，包括项目的python代码
  - result.db : 项目运行的结果数据
  - task.db : 项目相关的任务信息

- 关于项目控制说明

Recent Active Tasks				
group	project name	status	rate/burst	avg time
<a href="#">[group]</a>	<a href="#">huxiu</a>	TODO	<a href="#">1/3</a>	
<a href="#">[group]</a>	<a href="#">tripadvisor</a>	STOP	<a href="#">1/3</a>	

- Projects are independent, but you can import another project as a module with `from projects import other_project`
- A project has 5 status: TODO, STOP, CHECKING, DEBUG and RUNNING
  - TODO - a script is just created to be written
  - STOP - you can mark a project as STOP if you want it to STOP (= =).
  - CHECKING - when a running project is modified, to prevent incomplete modification, project status will be set as CHECKING automatically.
  - DEBUG/RUNNING - these two status have no difference to spider. But it's good to mark it as DEBUG when it's running the first time then change it to RUNNING after being checked.

- The crawl rate is controlled by rate and burst with token-bucket algorithm.
  - rate - how many requests in one second
  - burst - consider this situation,  $\text{rate/burst} = 0.1/3$ , it means that the spider scrawls 1 page every 10 seconds. All tasks are finished, project is checking last updated items every minute. Assume that 3 new items are found, pypider will "burst" and crawl 3 tasks without waiting  $3 \times 10$  seconds. However, the fourth task needs wait 10 seconds.

- To delete a project, set group to delete and status to STOP, wait 24 hours.

#### • API讲解

- `@every(minutes=0, seconds=0)`
  - 设置多少分钟或秒运行一次此方法,常用来定时执行任务.
- `@config(age=10 * 24 * 60 * 60)`
  - 本参数用来指定任务的有效期, 在有效期内不会重复抓取. 默认值是-1 (永远不过期, 意思是只抓一次)
- `@config(priority=2)`
  - 这个参数用来指定任务的优先级, 数值越大越先被执行. 默认值为0

#### • 参考资料

- pypider交接文档 ( 主要项目转移 )
  - <https://www.zybuluo.com/twein89/note/782441>