**如何提前生成列表页URL再提取内容？**

通常情况下，爬虫会从起始页(scan\_urls)开始通过列表页规则(list\_url\_regexes)寻找列表页，内容页同理，但是很多时候，第三方网站为了防止采集，会采用ajax的方式，不把列表页直接显式放在页面内容，而是通过js生成，又或者是直接显示前10页，因为正常的用户也只需要浏览前10页的数据就够了，现在我们针对这两种方式来看看抓取方法

只显示前10页的网页我们可以先生成列表页URL入队列

$configs = array(

// configs的其他成员

...

'scan\_urls' => array(

'https://www.itjuzi.com/investfirm?user\_id=305129'

),

'list\_url\_regexes' => array(

"https://www.itjuzi.com/investfirm\?user\_id=305129&page=\d+"

),

...

);

$spider->on\_start = function ($spider)

{

// 生成列表页URL入队列

for ($i = 0; $i <= 652; $i++)

{

$url = "https://www.itjuzi.com/investfirm?user\_id=305129&page={$i}";

$spider->add\_url($url);

}

};