**如果内容页有分页，该如何爬取到完整数据？**

如果要爬取的某个内容页中有多个分页，该如何爬取这个内容页的完整数据呢？这里就无法使用on\_list\_page回调函数了，而需要使用field中的attached\_url来请求其他分页的数据。

举个栗子:  
爬取某网站文章时，发现有些文章有多个内容页面，处理过程如下：

$configs = array(

// configs 的其他成员

...

'fields' => array(

array(

'name' => "contents",

'selector' => "//div[@id='pages']//a//@href",

'repeated' => true,

'children' => array(

array(

// 抽取出其他分页的url待用

'name' => 'content\_page\_url',

'selector' => "//text()"

),

array(

// 抽取其他分页的内容

'name' => 'page\_content',

// 发送 attached\_url 请求获取其他的分页数据

// attached\_url 使用了上面抓取的 content\_page\_url

'source\_type' => 'attached\_url',

'attached\_url' => 'content\_page\_url',

'selector' => "//\*[@id='big-pic']"

),

),

),

),

);

在爬取到所有的分页数据之后，可以在on\_extract\_page回调函数中将这些数据组合成完整的数据

$spider->on\_extract\_field = function($fieldname, $data, $page)

{

if ($fieldname == 'contents')

{

if (!empty($data))

{

$contents = $data;

$data = "";

foreach ($contents as $content)

{

$data .= $content['page\_content'];

}

}

}

return $data;

};