**爬虫进阶开发——之内置方法**

**本节介绍爬虫的内置方法**

**add\_url($url, $options = array())**

一般在on\_scan\_page和on\_list\_page回调函数（在[爬虫进阶开发——之回调函数](https://doc.phpspider.org/callback.html)中会详细描述）中调用, 用来往待爬队列中添加url

@param $url 待添加的url  
@param $options 成员包括method, headers, params, context\_data, reserve和 proxy, 如下所示:

@param $options['method'] 默认为"get"请求, 也支持"post"请求  
@param $options['headers'] 此url的Headers, 可以为空  
@param $options['params'] 发送请求时需添加的参数, 可以为空  
@param $options['context\_data'] 此url附加的数据, 比如内容页需要列表页一些数据, 可以为空  
@param $options['proxy'] 访问此url时使用的代理服务器，不使用请留空

栗子1:

$spider->on\_scan\_page = function($page, $content, $phpspider)

{

$regex = "#http://pic.qiushibaike.com/system/pictures/\d+#";

$urls = array();

preg\_match\_all($regex, $content, $out);

$urls = empty($out[0]) ? array() : $out[0];

if (!empty($urls)) {

foreach ($urls as $url)

{

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

}

}

...

return false;

};

**add\_scan\_url($url, $options = array())**

一般在on\_start回调函数（在[爬虫进阶开发——之回调函数](https://doc.phpspider.org/callback.html)中会详细描述）中调用, 用来往待爬队列中添加scan\_url

@param $url 待添加的scan\_url  
@param $options 成员包括method, headers, params, context\_data, reserve和 proxy, 如下所示:

@param $options['method'] 默认为"get"请求, 也支持"post"请求  
@param $options['headers'] 此url的Headers, 可以为空  
@param $options['params'] 发送请求时需添加的参数, 可以为空  
@param $options['context\_data'] 此url附加的数据, 比如内容页需要列表页一些数据, 可以为空  
@param $options['proxy'] 访问此url时使用的代理服务器，不使用请留空

举个栗子:

$spider->on\_start = function($page, $content, $phpspider)

{

for ($i = 0; $i < 10; $i++)

{

$phpspider->add\_scan\_url("http://www.qiushibaike.com/page.php?page=".$i);

}

};

栗子2:

$spider->on\_list\_page = function($page, $content, $phpspider)

{

...

$next\_url = str\_replace($page['url'], "page=".$current\_page\_num, "page=".$page\_num);

$options = array(

'method' => 'post',

'params' => array(

'page' => $page\_num,

'size' => 18

)

);

$phpspider->add\_url($next\_url, $options);

return false;

};

**request\_url($url, $options = array())**

一般在on\_start,on\_download\_page,on\_scan\_page和on\_list\_page回调函数（在[爬虫进阶开发——之回调函数](https://doc.phpspider.org/callback.html)中会详细描述）中调用, 下载网页, 得到网页内容

@param $url 待添加的url  
@param $options 成员包括method, headers, params, context\_data, reserve 和 proxy, 如下所示:

@param $options['method'] 默认为"get"请求, 也支持"post"请求  
@param $options['headers'] 此url的Headers, 可以为空  
@param $options['params'] 发送请求时需添加的参数, 可以为空  
@param $options['context\_data'] 此url附加的数据, 比如内容页需要列表页一些数据, 可以为空  
@param $options['proxy'] 访问此url时使用的代理服务器，不使用请留空

举个栗子:

$spider->on\_download\_page = function($page, $phpspider)

{

$url = "https://checkcoverage.apple.com/cn/zh/?sn=FK1QPNCEGRYD";

$options = array(

'method' => 'post',

'params' => array(

'sno' => 'FK1QPNCEGRYD',

'CSRFToken' => $result[1]

)

);

// 通过发送带参数的post请求，下载网页，并将网页内容赋值给result

$request = $phpspider->request\_url($url, $options);

...

return $page;

};