

爬虫基本流程

发起请求

通过HTTP库向目标站点发起请求，即发送一个Request，请求可以包含额外的headers等信息，等待服务器响应。

获取响应内容

如果服务器能正常响应，会得到一个Response，Response的内容便是所要获取的页面内容，类型可能有HTML，Json字符串，二进制数据（如图片视频）等类型。



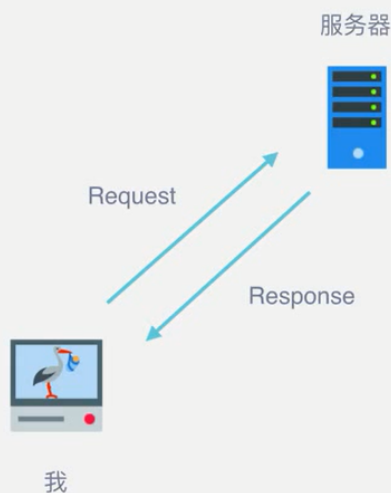
解析内容

得到的内容可能是HTML，可以用正则表达式、网页解析库进行解析。可能是Json，可以直接转为Json对象解析，可能是二进制数据，可以做保存或者进一步的处理。

保存数据

保存形式多样，可以存为文本，也可以保存至数据库，或者保存特定格式的文件。

Request与Response



(1) 浏览器就发送消息给该网址所在的服务器，这个过程叫做HTTP Request。

(2) 服务器收到浏览器发送的消息后，能够根据浏览器发送消息的内容，做相应处理，然后把消息回传给浏览器。这个过程叫做HTTP Response。

(3) 浏览器收到服务器的Response信息后，会对信息进行相应处理，然后展示。

Request

请求方式

主要有GET、POST两种类型，另外还有HEAD、PUT、DELETE、OPTIONS等。

请求URL

URL全称统一资源定位符，如一个网页文档、一张图片、一个视频等都可以用URL唯一来确定。



请求头

包含请求时的头部信息，如User-Agent、Host、Cookies等信息。

请求体

请求时额外携带的数据如表单提交时的表单数据

怎样解决JavaScript渲染的问题？

分析Ajax请求



Splash



Selenium/WebDriver



PyV8、Ghost.py

