爬虫基本原理讲解

什么是爬虫?

什么是爬虫?

请求网站并提取数据的自动化程序

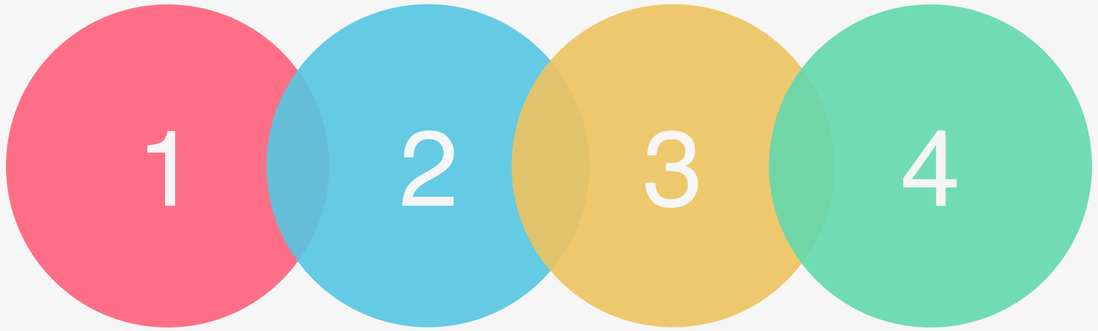
爬虫基本流程

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

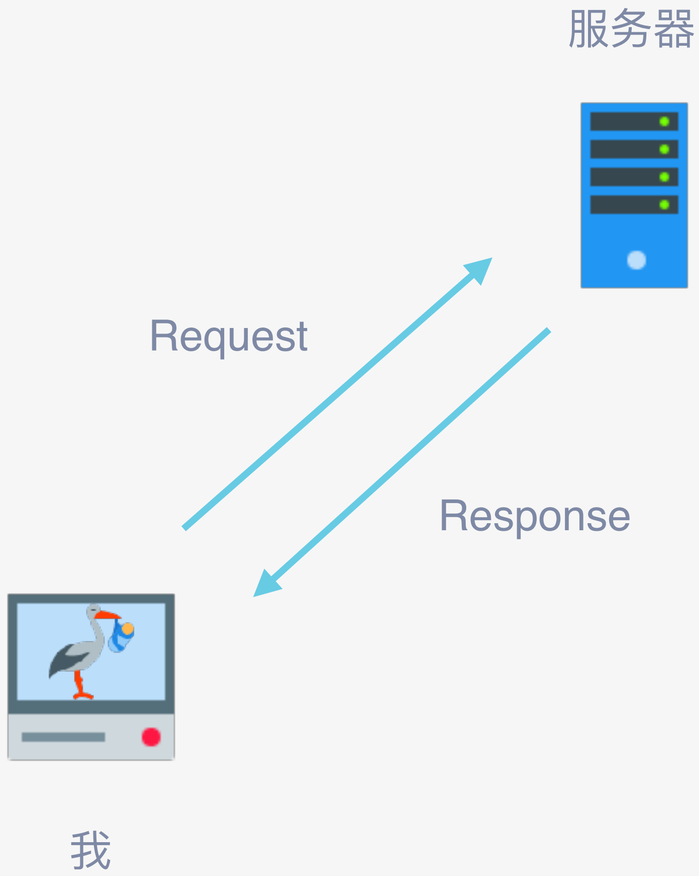
得到的内容可能是HTML，可以用正则表达式、网页解析库进行解析。可能是Json，可以直接转为Json对象解析，可能是二进制数据，可以做保存或者进一步的处理。

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。

什么是Request和Response?

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中包含什么?

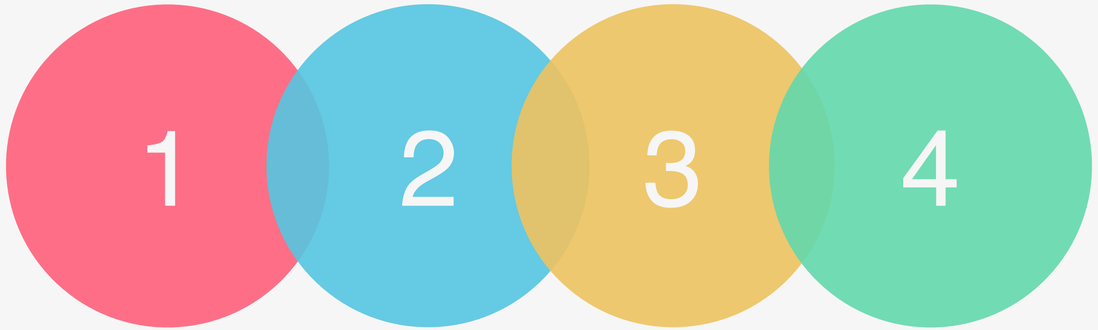
Request

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

Response中包含什么?

Response

1

响应状态

有多种响应状态，如200代表成功、301跳转、404找不到页面、502服务器错误

2

响应头

如内容类型、内容长度、服务器信息、设置Cookie等等。

3

响应体

最主要的部分，包含了请求资源的内容，如网页HTML、图片二进制数据等。

能抓怎样的数据?

能抓怎样的数据?

网页文本

如HTML文档、Json格式文本等。

1

2

图片

获取到的是二进制文件，保存为图片格式。

视频

同为二进制文件，保存为视频格式即可。

3

4

其他

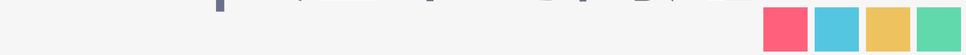
只要是能请求到的，都能获取。

怎样来解析?

解析方式

为什么我抓到的和浏览器看到的不一样?

怎样解决JavaScript渲染的问题?



怎样解决JavaScript渲染的问题?

- 分析Ajax请求
- Selenium/WebDriver
- Splash
- PyV8、Ghost.py

可以怎样保存数据?

怎样保存数据?

1

文本

纯文本、Json、Xml等。

2

关系型数据库

如MySQL、Oracle、SQL Server等具有结构化表结构形式存储。

3

非关系型数据库

如MongoDB、Redis等Key-Value形式存储。

4

二进制文件

如图片、视频、音频等等直接保存成特定格式即可。

谢谢