

爬虫是用程序模拟浏览器的行为,发送请求给服务器,获取网页内容,解析网页数据,浏览器通过HTTP协议和服务器交流

HTTP三个要点

1. HTTP是无连接:每次连接只处理一个请求,每次请求完成后自动断开连接
2. HTTP是媒体独立的
3. HTTP是无状态的:指的是协议对于事务处理没有记忆能力,如果后续处理需要前面信息,必须重新传输,每次继续访问都会带着记忆id,这会减慢处理速度

HTTP的消息结构

请求:

1. 地址和动词:GET <http://www.baidu.com>
2. 请求头(header):用来描述请求和发送者的一些信息(接受语言,浏览器版本等等)
3. 请求参数:要搜索的关键词例如:http://baidu.com/helloworld\*\*参数

响应:

1,响应代码:200表示成功,404表示不存在等

1. 响应头:描述相应的一些信息
2. 响应内容:HTML,JSON,图片等



完整的网页请求过程:

1. 先请求网页文本
2. 根据HTML中的指定的地址,请求其他内容

浏览器的请求是一带多的过程,会先将网页信息(HTML)请求过来,再将关于网页的超链接(图片视频等)请求过来

如何解决HTTP请求给爬虫带来的困难:

1. 分析各个请求,找到关键的后续请求,用代码发送后续请求.通常,我们不需像浏览器一样发送所有后续请求,找到关键请求就够了
2. 使用selenium等浏览器驱动,模拟真实的浏览器操作,发送后续所有请求

get和post的区别

GET 请求一般不应产生副作用。就是说，它仅仅是获取资源信息，就像数据库查询一样，不会修改，增加数据，不会影响资源的状态

POST表示可能修改变服务器上的资源的请求



get请求参数会放在网页上









