**爬虫原理：<https://www.cnblogs.com/value-code/p/8270037.html>**

**HTTP协议：<https://www.cnblogs.com/an-wen/p/11180076.html>**

**HTTPS：<https://zhuanlan.zhihu.com/p/72616216>**

**HTTP“三次握手”**

浏览器请求tcp连接（第一次握手）

服务器答应进行tcp连接（第二次握手）

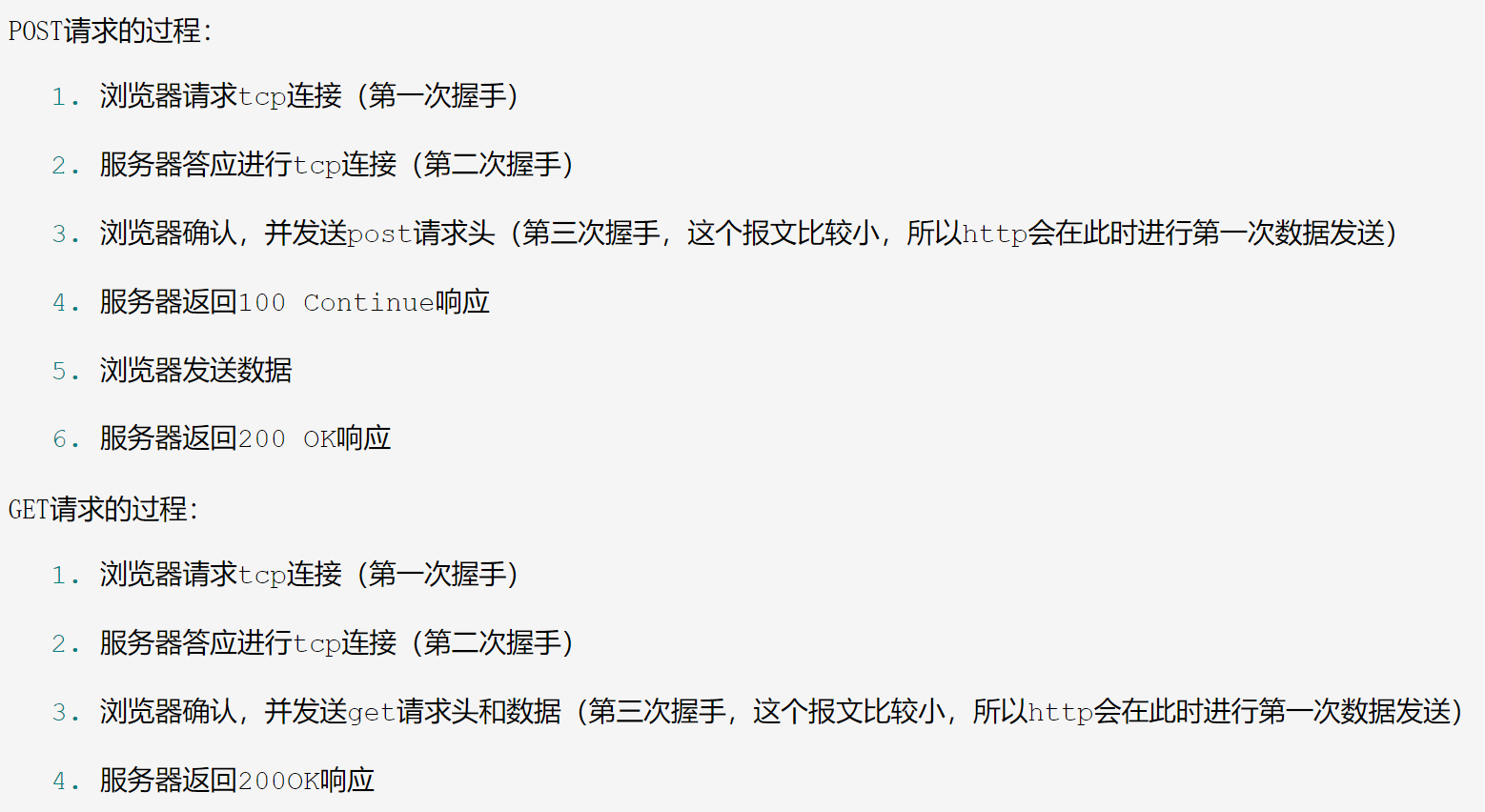
浏览器确认，并发送请求头，正式建立连接（第三次握手）

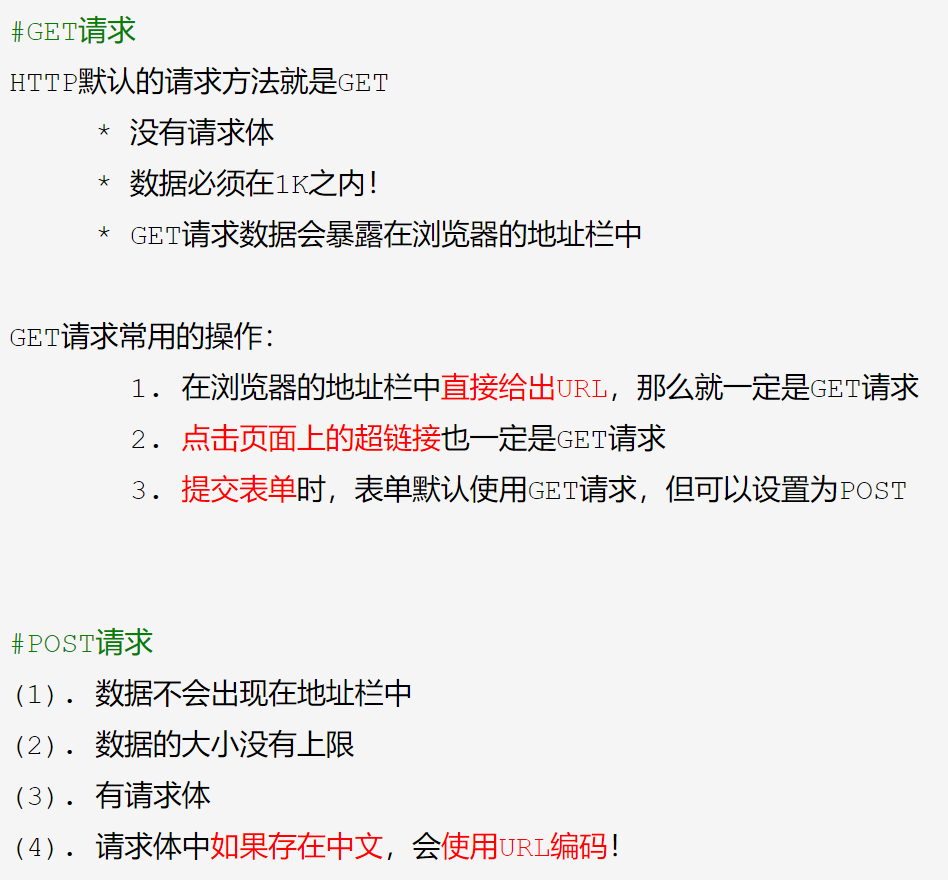
**建立连接的八种方式**

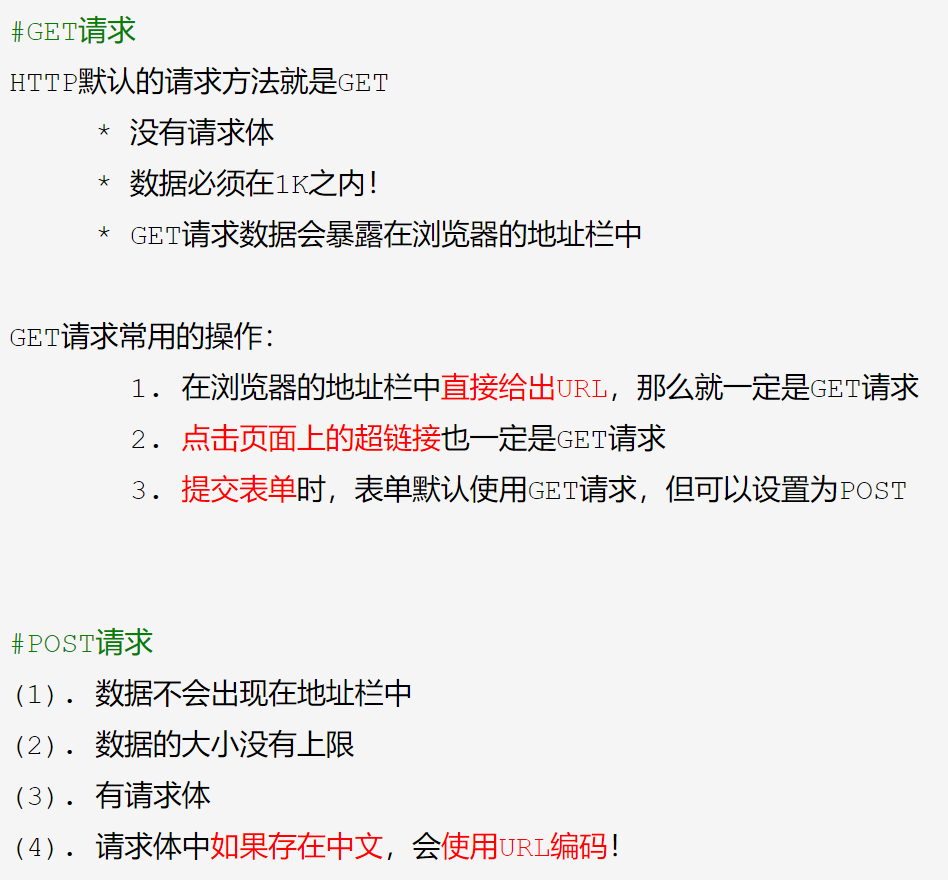


对于爬虫而言，只涉及GET和POST

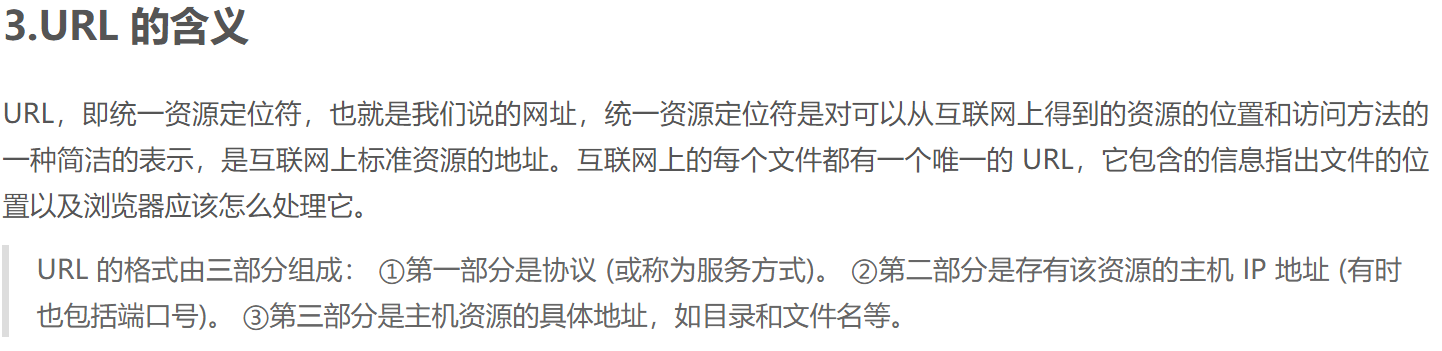
**GET&POST差别**







**URL的含义**



如果使用了GET方式上传数据，数据会在url里以键值对的形式跟在一个’?’后，并以’&’连接如：

http://httpbin.org/get?key2=value2&key1=value1

**爬虫工作的实质**

访问网页的过程其实就是用户输入URL之后，经过 DNS 服务器，找到服务器主机，向服务器发出一个请求，服务器经过解析之后，发送给用户的浏览器 HTML、JS、CSS 等文件，浏览器解析出来，用户便可以看到网页内容。 因此，用户看到的网页实质是经过渲染的HTML文件，爬虫爬来的便是HTML文件，通过对HTML文件代码的解析，实现对图片、文字等资源的获取。

**请求**

请求包含请求头和请求尾，对应的，发回的Response也有头和尾

**Cookie**

指某些网站为了辨别用户身份、进行 session 跟踪而储存在用户本地终端上的数据

**访问状态码及含义：**

**1XX- 信息型**，服务器收到请求，需要请求者继续操作。

100：继续客户端应当继续发送请求。客户端应当继续发送请求的剩余部分，或者如果请求已经完成，忽略这个响应。

101：转换协议在发送完这个响应最后的空行后，服务器将会切换到在Upgrade消息头中定义的那些协议。只有在切换新的协议更有好处的时候才应该采取类似措施。

102：继续处理由WebDAV（RFC2518）扩展的状态码，代表处理将被继续执行。

**2XX- 成功型**，请求成功收到，理解并处理。

200：请求成功——处理方式：获得响应的内容，进行处理

201：请求完成，结果是创建了新资源。新创建资源的URI可在响应的实体中得到——处理方式：爬虫中不会遇到

202：请求被接受，但处理尚未完成——处理方式：阻塞等待

204：服务器端已经实现了请求，但是没有返回新的信息。如果客户是用户代理，则无须为此更新自身的文档视图。——处理方式：丢弃

**3XX - 重定向**，需要进一步的操作以完成请求。

300：该状态码不被HTTP/1.0的应用程序直接使用，只是作为3XX类型回应的默认解释。存在多个可用的被请求资源。——处理方式：若程序中能够处理，则进行进一步处理，如果程序中不能处理，则丢弃

301：请求到的资源都会分配一个永久的URL，这样就可以在将来通过该URL来访问此资源——处理方式：重定向到分配的URL

302：请求到的资源在一个不同的URL处临时保存——处理方式：重定向到临时的URL

304：请求的资源

未更新——处理方式：丢弃

**4XX - 客户端错误**，请求包含语法错误或无法完成请求。

400：非法请求——处理方式：丢弃

401：未授权——处理方式：丢弃

403：禁止——处理方式：丢弃

404：没有找到——处理方式：丢弃

**5XX - 服务器错误**，服务器在处理请求的过程中发生了错误。

500：服务器内部错误服务器遇到了一个未曾预料的状况，导致了它无法完成对请求的处理。一般来说，这个问题都会在服务器端的源代码出现错误时出现。

501：服务器无法识别服务器不支持当前请求所需要的某个功能。当服务器无法识别请求的方法，并且无法支持其对任何资源的请求。

502：错误网关作为网关或者代理工作的服务器尝试执行请求时，从上游服务器接收到无效的响应。

503：服务出错由于临时的服务器维护或者过载，服务器当前无法处理请求。这个状况是临时的，并且将在一段时间以后恢复。