微信公众号爬虫的基本原理

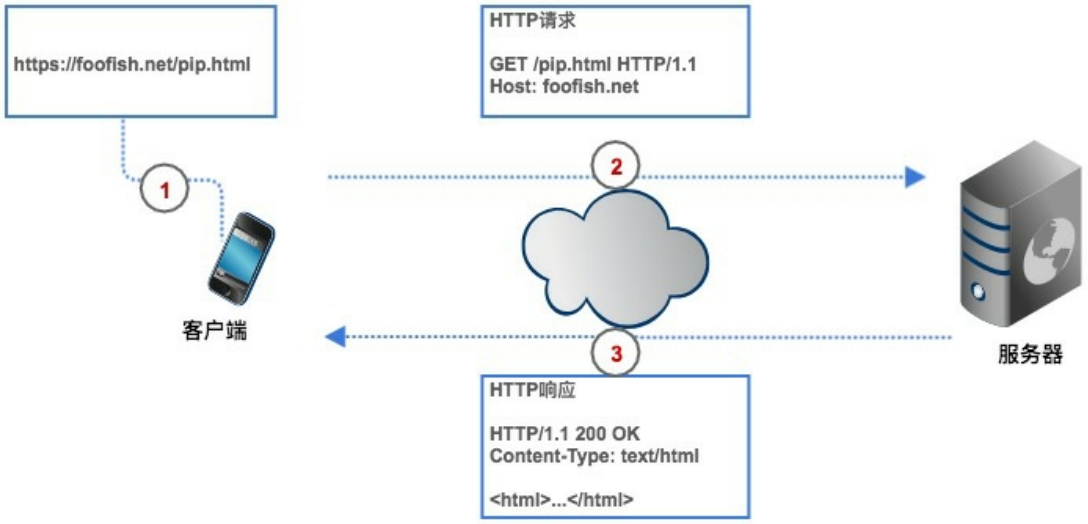
网上关于爬虫的教程多如牛毛，但很少有看到微信公众号爬虫教程，要有也是基于搜狗微信的，不过搜狗提供的数据有诸多弊端，比如文章链接是临时的，文章没有阅读量等指标，所以我想写一个比较系统的关于如何通过手机客户端利用 Python 爬微信公众号文章的教程，并对公众号文章做数据分析，为更好的运营公众号提供决策。

爬虫的基本原理

所谓爬虫就是一个自动化数据采集工具，你只要告诉它要采集哪些数据，丢给它一个 URL，就能自动地抓取数据了。其背后的基本原理就是爬虫程序向目标服务器发起 HTTP 请求，然后目标服务器返回响应结果，爬虫客户端收到响应并从中提取数据，再进行数据清洗、数据存储工作。

爬虫的基本流程

爬虫流程也是一个 HTTP 请求的过程，以浏览器访问一个网址为例，从用户输入 URL 开始，客户端通过 DNS 解析查询到目标服务器的 IP 地址，然后与之建立 TCP 连接，连接成功后，浏览器构造一个 HTTP 请求发送给服务器，服务器收到请求之后，从数据库查到相应的数据并封装成一个 HTTP 响应，然后将响应结果返回给浏览器，浏览器对响应内容进行数据解析、提取、渲染并最终展示在你面前。



HTTP 协议的请求和响应都必须遵循固定的格式，只有遵循统一的 HTTP 请求格式，服务器才能正确解析不同客户端发的请求，同样地，服务器遵循统一的响应格式，客户端才得以正确解析不同网站发过来的响应。

HTTP 请求格式

HTTP 请求由请求行、请求头、空行、请求体组成。

请求行由三部分组成：

1. 第一部分是请求方法，常见的请求方法有 GET、POST、

PUT、DELETE、HEAD
2. 第二部分是客户端要获取的资源路径
3. 第三部分是客户端使用的 HTTP 协议版本号

请求头是客户端向服务器发送请求的补充说明，比如 User-Agent 向服务器说明客户端的身份。

请求体是客户端向服务器提交的数据，比如用户登录时需要提高的账号密码信息。请求头与请求体之间用空行隔开。请求体并不是所有的请求都有的，比如一般的GET都不会带有请求体。

上图就是浏览器登录豆瓣时向服务器发送的HTTP POST 请求，请求体中指定了用户名和密码。

HTTP 响应格式

HTTP 响应格式与请求的格式很相似，也是由响应行、响应头、空行、响应体组成。

响应行也包含三部分，分别是服务端的 HTTP 版本号、响应状态码、状态说明，响应状态码常见有 200、400、404、500、502、304 等等，一般以 2 开头的表示服务器正常响应了客户端请求，4

开头表示客户端的请求有问题，5 开头表示服务器出错了，没法正确处理客户端请求。状态码说明就是对该状态码的一个简短描述。

第二部分就是响应头，响应头与请求头对应，是服务器对该响应的一些附加说明，比如响应内容的格式是什么，响应内容的长度有多少、什么时间返回给客户端的、甚至还有一些 Cookie 信息也会放在响应头里面。

第三部分是响应体，它才是真正的响应数据，这些数据其实就是网页的 HTML 源代码。

小结

这仅仅只是一个爬虫基本原理的介绍，涉及的 HTTP 协议的内容也非常有限，但不可能用一篇文章事无巨细的介绍完，因为 HTTP 协议是一个很大的话题，用一本书也写不完，深入了解推荐两本书《图解HTTP》、《HTTP权威指南》。