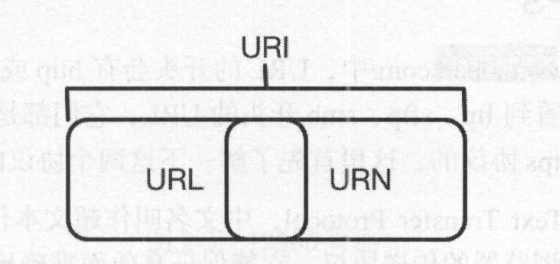
# 一、基础知识

## 1.1 URL和URI

举例来说，https://github.com/favicon.ico是GitHub的网站图标链接，它是一个URL，也是一个URI。即有这样的一个图标资源，我们用URL/URI来唯一指定了它的访问方式，这其中包括了访问协议https、访问路径(／即根目录)和资源名称favicon.ico。通过这样一个链接，我们便可以从互联网上找到这个资源，这就是URL/URI.

URL是URI的子集，也就是说每个URL都是URI，但不是每个URI都是URL。



但是在目前的互联网中，URN用得非常少，所以几乎所有的URI都是URL，一般的网页链接我们既可以称为URL，也可以称为URI。

## 1.2 超文本

接下来，我们再了解一个概念——超文本，其英文名称叫作hypertext，我们在浏览棉里看到的网页就是超文本解析而成的，其网页源代码是一系列HTML代码，里面包含了一系列标签，比如img显示图片，p指定显示段落等。浏览器解析这些标签后，便形成了我们平常看到的网页，而网页的源代码HTML就可以称作超文本（F12-Elements选项卡）。

## 1.3 HTTP和HTTPS

在淘宝的首页https://www.taobao.com/中，URL的开头会有http或https，这就是访问资源需要的协议类型。有时，我们还会看到ftp、sftp、smb开头的URL，它们都是协议类型。在爬虫中，我们抓取的页面通常就是http或https协议的，这里首先了解一下这两个协议的含义。

HTTP的全称是HyperTextTransferProtocol，中文名叫作超文本传输协议。HTTP协议是用于从网络传输超文本数据到本地浏览器的传送协议，它能保证高效而准确地传送超文本文档。HTTP由万维网协会（WorldWideWebConsortium）和Internet工作小组IETF(InternetEngineeringTaskForce)共同合作制定的规范，目前广泛使用的是HTTP1.1版本。

HTTPS的全称是HyperTextTransferProtocoloverSecureSocketLayer，是以安全为目标的HTTP通道，简单讲是HTTP的安全版，即HTTP下加入SSL层，简称为HTTPS。

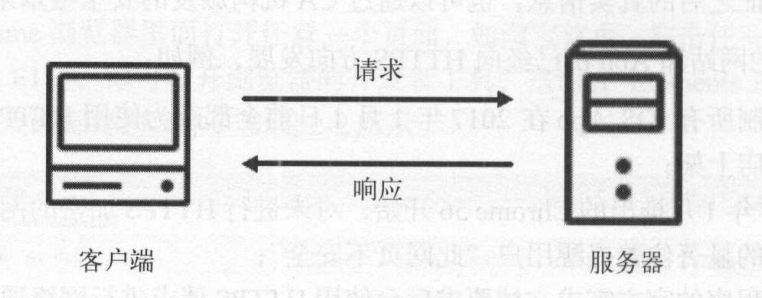
HTTPS的安全基础是SSL，因此通过它传输的内容都是经过SSL加密的它的主要作用可以分为两种。

（1）建立一个信息安全通道来保证数据传输的安全。

（2）确认网站的真实性，凡是使用了HTTPS的网站，都可以通过点击浏览器地址栏的锁头标志来查看网站认证之后的真实信息，也可以通过CA机构颁发的安全签章来查询。

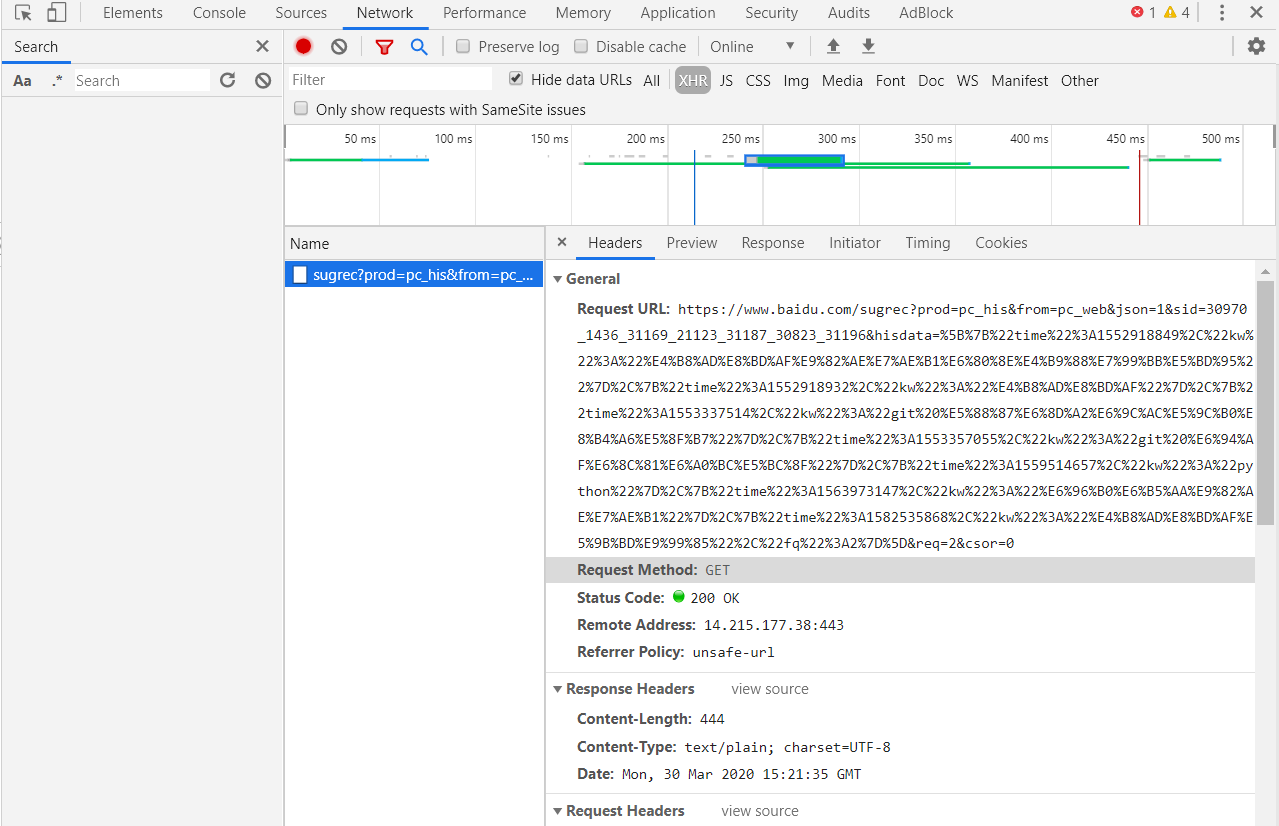
# 二、HTTP请求过程

我们在浏览器中输入一个URL，回车之后便会在浏览器中观察到页面内容。实际上，这个过程是浏览器向网站所在的服务器发送了一个请求，网站服务器接收到这个请求后进行处理和解析，然后返回对应的响应，接着传回给浏览器。响应里包含了页面的源代码等内容，浏览器再对其进行解析，便将网页呈现了出来，模型如图所示。



为了更直观地说明这个过程，这里用Chrome浏览器的开发者模式下的Network监听组件来做下演示，它可以显示访问当前请求网页时发生的所有网络请求和响应。

打开Chrome浏览器，右击并选择“检查”项，即可打开浏览器的开发者工具。这里访问百度http://www.baidu.com/，输入该URL后回车，观察这个过程中发生了怎样的网络请求。可以看到，在Network页面下方出现了一个个的条目，其中一个条目就代表一次发送请求和接收响应的过程，如图所示。



其中各列的含义如下。

首先是General部分，RequestURL为请求的URL，RequestMethod为请求的方法，StatusCode为响应状态码，RemoteAddress为远程服务器的地址和端口，ReferrerPolicy为Referrer判别策略。

再继续往下，可以看到，有ResponseHeaders和RequestHeaders，这分别代表响应头和请求头。请求头里带有许多请求信息，例如浏览器标识、Cookies、Host等信息，这是请求的一部分，服务器会根据请求头内的信息判断请求是否合法，进而作出对应的响应。图中看到的ResponseHeaders就是响应的一部分，例如其中包含了服务器的类型、文档类型、日期等信息，浏览器接受到响应后，会解析响应内容，进而呈现网页内容。

下面我们分别来介绍一下请求和响应都包含哪些内容。

## 2.1 请求

### 2.1.1 请求方法

常见的请求方法有两种：GET和POST。

GET和POST请求方法有如下区别。

(a) GET请求中的参数包含在URL里面，数据可以在URL中看到，而POST请求的URL不会包含这些数据，数据都是通过表单形式传输的，会包含在请求体中。

(b) GET请求提交的数据最多只有1024字节，而POST方式没有限制。

### 2.1.2 请求网址

请求的网址，即统一资惊定位符URL，它可以唯一确定我们想请求的资源。

### 2.1.3 请求头

请求头，用来说明服务器要使用的附加信息，比较重要的信息有Cookie、Referer、User-Agent等。

下面简要说明一些常用的头信息。

(a) Host：用于指定请求资源的主机IP和端口号，其内容为请求URL的原始服务器或网关的位置。从HTTP1.l版本开始，请求必须包含此内容。

(b) Referer：此内容用来标识这个请求是从哪个页面发过来的，服务器可以拿到这一信息并做相应的处理，如做来源统计、防盗链处理等。

(c) User-Agent：简称UA，它是一个特殊的字符串头，可以使服务器识别客户使用的操作系统及版本、浏览器及版本等信息。在做爬虫时加上此信息，可以伪装为浏览器；如果不加，很可能会被识别州为爬虫。

### 2.1.4 请求体

请求体－般承载的内容是POST请求中的表单数据，而对于GET请求，请求体则为空。

## 2.2 响应

响应，由服务端返回给客户端，可以分为三部分：响应状态码（ResponseStatusCode）、响应头(ResponseHeaders）和响应体（ResponseBody）。

### 2.1 响应状态码

响应状态码表示服务器的响应状态，如200代表服务器正常响应，404代表页面未找到，500代表服务器内部发生错误。在爬虫中，我们可以根据状态码来判断服务器响应状态，如状态码为200，则证明成功返回数据，再进行进一步的处理，否则直接忽略。

### 2.2 响应头

响应头包含了服务器对请求的应答信息，如Content-Type、Server、Set-Cookie等。

### 2.3 响应体

最重要的当属响应体的内容了。响应的正文数据都在响应体中，比如请求网页时，它的响应体就是网页的HTML代码；请求一张图片时，它的响应体就是图片的二进制数据。我们做爬虫请求网页后，要解析的内容就是响应体

# 三、代码演示操作