

# 教你在 Chrome 浏览器轻松抓包

---

我们知道了什么是爬虫

也知道了爬虫的具体流程

那么在我们要对某个网站进行爬取的时候

要对其数据进行分析

就要知道应该怎么请求

就要知道获取的数据是什么样的

所以我们要学会怎么数据抓包

ok，打开 Chrome 浏览器之后呢

我们随便输入一个网址吧

输入一个人人都能上的网站

[www.baidu.com](http://www.baidu.com)

用力回车

一个熟悉的页面显示在你的面前



这个时候，你按下 F12

你可以看到弹出一个有点装逼的窗口



这个玩意

正是我们想要的

可以看到

Element 标签下对应的 HTML 代码

其实就是这个网页的代码

我们可以在这里除了看看它的代码之外

我们还可以修改一些东西

比如我把这个按钮改成Python小课



按下回车



是不是瞬间逼格满满

咳咳，回归正题

接下来我们点击 Network 这个标签

然后刷新一下

GET, POST, PUT, DELETE, HEAD, **OPTIONS**, TRACE

咱们就一一说道说道

## 学习 python 的正确姿势

然后我们就可以发现

**Baidu** 周杰伦

百度一下


百度网盘 设置 登录

---

网页 资讯 视频 图片 知道 文库 贴吧 采购 地图 更多»

百度为您找到相关结果的58,000,000个





**周杰伦** [歌曲在线试听\\_网易云音乐](#)



地区：港台  
生日：1979-01-18  
星座：摩羯座  
单曲：[15首](#)

[▶ 播放JAY热门歌曲](#)

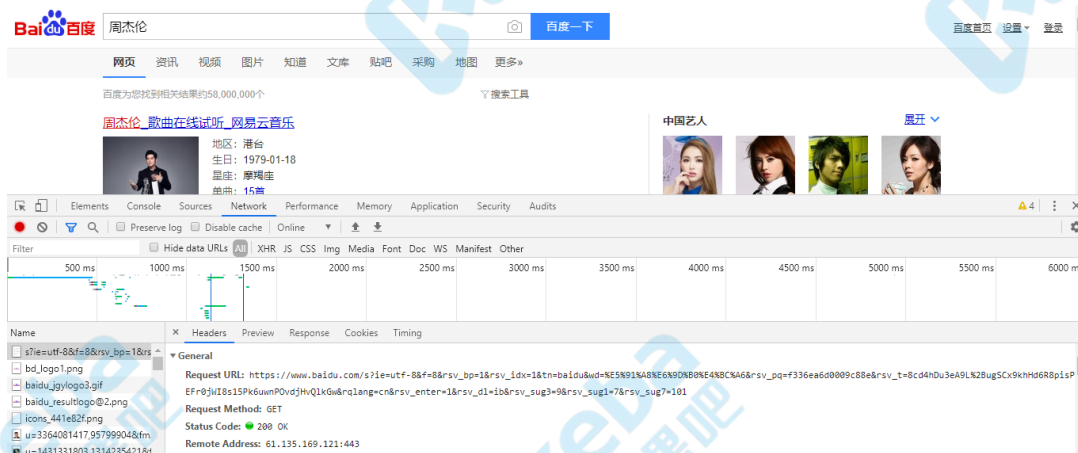
**中国艺人**

中流砥柱的 四大教主之 俊美狠辣的 温柔亲和的

我们随便点击一个请求进去

我们随便点击一个请求进去



可以看到我们的请求URL

[https://www.baidu.com/s?ie=utf-8&f=8&rsv\\_bp=1&rsv\\_idx=1&t=baidu&wd=周杰伦&rsv\\_pq=f336ea6d0009c88e&rsv\\_t=8cd4hDu3eA9L%28ugSCx9khHd6R8pisPEFr0jWI8s15Pk6uwnPOvdjHvQlkGw&rqlang=cn&rsv\\_enter=1&rsv\\_dl=ib&rsv\\_sug3=9&rsv\\_sug1=7&rsv\\_sug7=101](https://www.baidu.com/s?ie=utf-8&f=8&rsv_bp=1&rsv_idx=1&t=baidu&wd=周杰伦&rsv_pq=f336ea6d0009c88e&rsv_t=8cd4hDu3eA9L%28ugSCx9khHd6R8pisPEFr0jWI8s15Pk6uwnPOvdjHvQlkGw&rqlang=cn&rsv_enter=1&rsv_dl=ib&rsv_sug3=9&rsv_sug1=7&rsv_sug7=101)

在 ? 后面的这些玩意儿

就是 GET 请求的参数

这些参数以「键值对」的形式实现

比如这里的

`wd=周杰%20伦`

就是告诉百度

我们要查询的是周杰伦相关的内容

这种方式的请求方式是最简单的

所以以后我们在 Python 写 GET 请求的时候

直接在 URL 后面加个 ? 然后添加参数值就好了

比如

我要百度搜索五月天

那么就是

<https://www.baidu.com/s?wd=五月天>

不信你直接在浏览器这样搜

是一毛一样的

那么，啥是 POST 请求呢？

我们在做一些**信息提交**的时候

比如注册，登录

这时候我们做的就是 POST 请求

POST 的参数不会直接放在 URL 上

会以 Form 表单的形式将数据提交给服务器

我们来登录一下百度



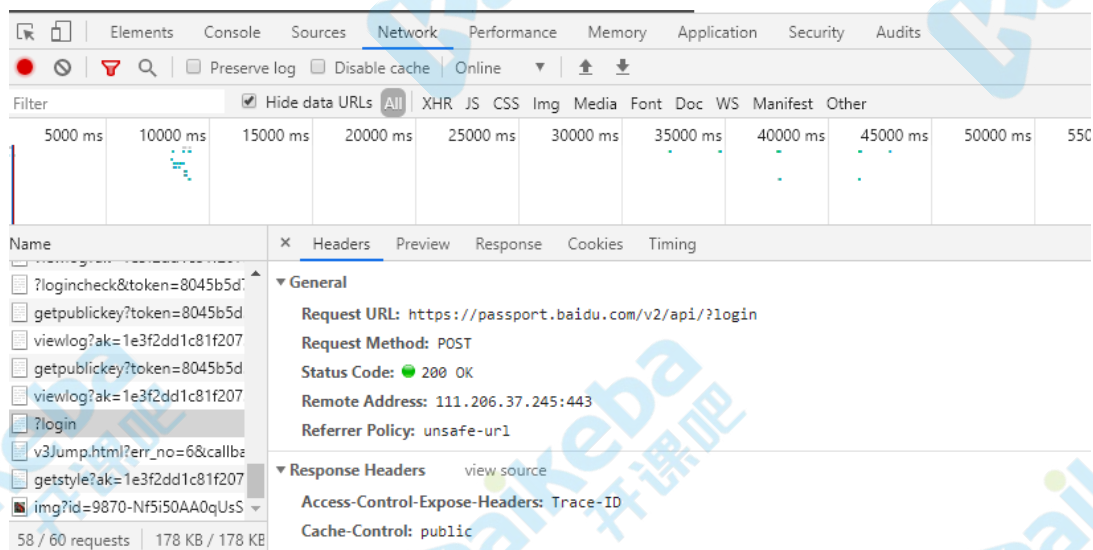
当我们点击登录的时候

就开始将我们的账号密码请求给百度服务器

可以看到我们请求了 login 这个接口

请求方法就是 POST





而我们的请求参数是以 Form 表单的方式提交的



拉到下面就可以看到

username 就是 可以显示出来

而密码，就是被加密了的

这些都是 POST 参数

可以发现

GET请求把请求参数都暴露在URL上

而POST请求的参数放在request body 里面

POST请求方式还对密码参数加了密

这样就相对安全一些

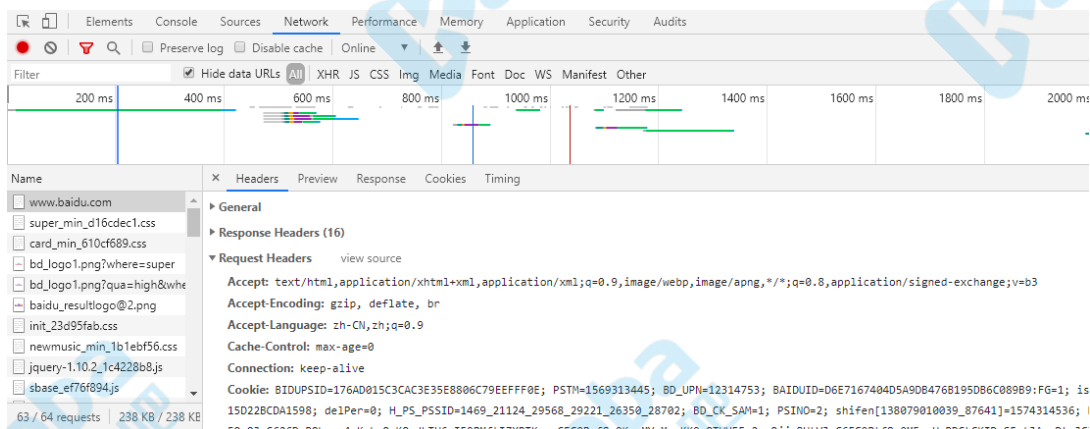
ok

你已经了解请求方式了

接下来说说请求头

我们刚刚在访问百度的时候

可以看到这个玩意



这个就是请求头

## Request Header

我们在做 HTTP 请求的时候

除了提交一些参数之外

我们还有定义一些 HTTP 请求的头部信息

比如 Accept、Host、cookie、User-Agent等等

这些参数也是我们在做爬虫要用到

通过这些信息，欺骗服务器，告诉它是正规请求

比如

我们可以在代码里面设置 cookie 告诉服务器我们就是在这个浏览器请求的会话

User-Agent 告诉服务器我们是浏览器请求的

说完我们这边的请求了

接着我们再说说服务器的**响应**

你一定遇到过 404 页面吧

或者服务器错误返回个 502 吧



这些 404 啊，200啊，301啊，502啊

都是服务器的响应码

一般服务器给我们返回 200

那就说明

我们成功请求了



再来说说响应头

当我们请求成功之后

服务器会给我们返回响应码之外

还有响应头

这个头主要是告诉我们数据以什么样的形式展现

告诉我们cookie的设置

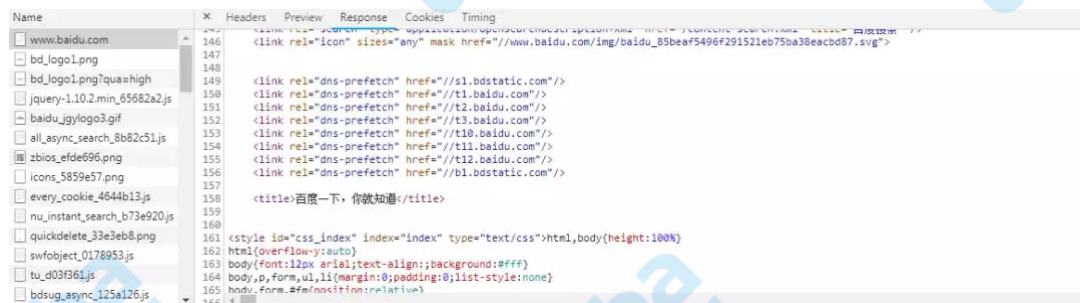
还有一个

就是响应体了

说白了，就是服务器返回给我们的数据



我们点击 Response 就可以看到相关的数据了



看，这些就是服务器返回给我们的 HTML 源代码

对于不同的请求

我们获取到的数据是不一样的

除了 HTML 的，也有 JSON 的

图片二进制数据等等

可以针对不同的情况

用不同的手段来解析这些数据