

# 第 0 关、初识爬虫

## 1、初始爬虫

爬虫，从本质上来说，就是利用程序在网上拿到对我们有价值的信息。

## 2、明晰路径

### 2-1、浏览器工作原理



- (1) 解析数据：当服务器把数据响应给浏览器之后，浏览器并不会直接把数据丢给我们。因为这些数据是用计算机的语言写的，浏览器还要把这些数据翻译成我们能看懂的内容；
- (2) 提取数据：我们就可以在拿到的数据中，挑选出对我们有用的数据；
- (3) 存储数据：将挑选出来的有用数据保存在某一文件/数据库中。

### 2-2、爬虫工作原理

## 爬虫的四个步骤

0 获取数据

1 解析数据

2 提取数据

3 储存数据

by 风变编程

- (1) 获取数据：爬虫程序会根据我们提供的网址，向服务器发起请求，然后返回数据；
- (2) 解析数据：爬虫程序会把服务器返回的数据解析成我们能读懂的格式；
- (3) 提取数据：爬虫程序再从中提取出我们需要的数据；
- (4) 储存数据：爬虫程序把这些有用的数据保存起来，便于你日后的使用和分析。

## 3、体验爬虫

### 3-1、requests.get()

#### ①、安装 requests 库

- Mac电脑里打开终端软件（terminal），输入`pip3 install requests`，然后点击enter；
- Windows电脑里叫命令提示符（cmd），输入`pip install requests`。

提示：往后安装其他库时与上方类似，`pip install 模块名`

#### ②、requests 库作用

requests 库可以帮我们下载网页源代码、文本、图片，甚至是音频。其实，“下载”本质上是向服务器发送请求并得到响应。

#### ③、requests 库使用

```
1 res = requests.get('URL')
```

`requests.get` 是在调用requests库中的`get()`方法，它向服务器发送了一个请求，括号里的参数是你需要的数据所在的网址，然后服务器对请求作出了响应。我们把这个响应返回的结果赋值在变量`res`上。



### 3-2、Response对象的常用属性

Response对象的常用属性	
属性	作用
response.status_code	检查请求是否成功
response.content	把reponse对象转换为二进制数据
response.text	把reponse对象转换为字符串数据
response.encoding	定义response对象的编码

by 风变编程

#### ①、response.status\_code

打印 response 的响应状态码，以检查请求是否成功。

常见响应状态码解释			
响应状态码	说明	举例	说明
1xx	请求收到	100	继续提出请求
2xx	请求成功	200	成功
3xx	重定向	305	应使用代理访问
4xx	客户端错误	403	禁止访问
5xx	服务器端错误	503	服务不可用

by 风变编程

#### ②、response.content

把 Response 对象的内容以二进制数据的形式返回，适用于图片、音频、视频的下载。

### ③、response.text

把 Response 对象的内容以字符串的形式返回，适用于文字、网页源代码的下载。

### ④、response.encoding

能帮我们定义Response对象的编码。（遇上文本的乱码问题，才考虑用res.encoding）

## 3-3、汇总图解



## 4、爬虫伦理

### 4-1、Robots 协议

Robots 协议是互联网爬虫的一项公认的道德规范，它的全称是“网络爬虫排除标准”（Robots exclusion protocol），这个协议用来告诉爬虫，哪些页面是可以抓取的，哪些不可以。

### 4-2、协议查看

（1）在网站的域名后加上/robots.txt就可以了。如淘宝的robots协议（<http://www.taobao.com/robots.txt>）；

（2）协议里最常出现的英文是Allow和Disallow，Allow代表可以被访问，Disallow代表禁止被访问。