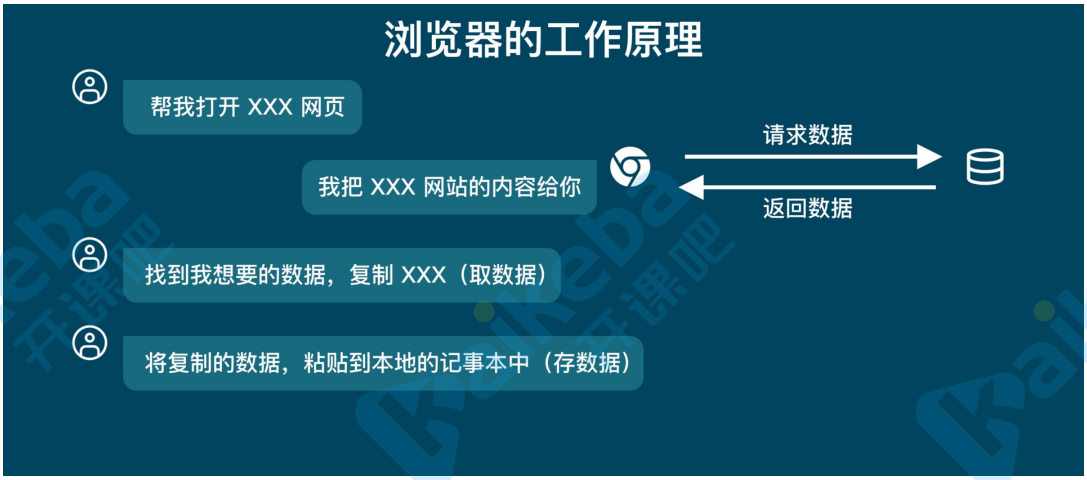
第1课 了解爬虫和浏览器的原理

1、初始爬虫

爬虫就是一段程序，它会在网络上“自由”穿梭，拿到编写这段程序的人需要的数据。

2、什么是爬虫

2-1、浏览器工作原理



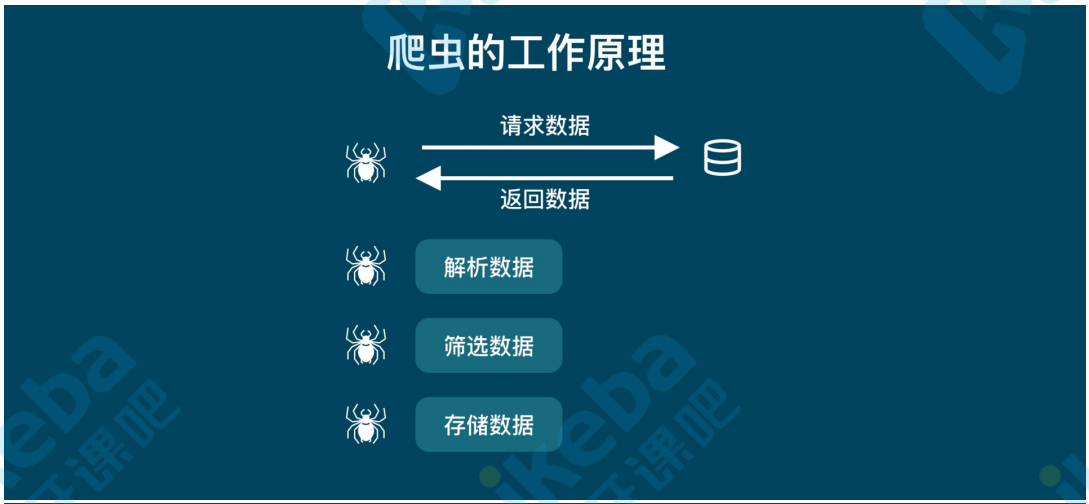
当浏览器收到服务器返回的数据时，它会先「解析数据」，把数据变成人能看得懂的网页页面。

当我们浏览这个网页的时候，我们会「筛选数据」，找出我们需要的数据，比如说一篇文章，一份论文等。

然后我们把这一篇文章，或者是一篇论文保存到本地，这就叫做「存储数据」

（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得懂的内容；
（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



（1）获取数据。爬虫会拿到我们要它去爬的网址，像服务器发出请求，获得服务器返回的数据；

（2）解析数据。爬虫会将服务器返回的数据转换成人能看懂的样式；

（3）筛选数据。爬虫会从返回的数据中筛选出我们需要的特定数据；

（4）存储数据。爬虫会根据我们设定的存储方式，将数据保存下来，方便我们进行后一步的操作。

3、编写爬虫

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

```
import requests
#在使用前需要先通过 import 来引入 requests 库
res = requests.get('URL')
#我们通过调用requests库中的get()方法来获取数据，这个方法需要一个参数，这个参数就是你需要请求的网址。当请求得到「响应」时，服务器返回的数据就被赋值到 res 这个变量上面
```

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对象常用属性

属性	知识点
response.status_code	检查请求是否成功
response.content	response对象的二进制数据
response.text	response对象的字符串数据
response.encoding	response对象的编码

①、response.status_code

打印 response 的响应状态码，以检查请求是否成功。

常见相应状态码解释

响应状态码	说明	例子	例子说明
1xx	请求接收	100	继续提出请求
2xx	请求成功	200	请求成功
3xx	重定向	305	应使用代理访问
4xx	客户端错误	403	禁止访问
5xx	服务器错误	503	服务不可用

tips：状态码这么多记不住怎么办？没关系，常用的就是除了200之外，其他都是请求遇到问题的情况。遇到了之后再查阅就行了。

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。

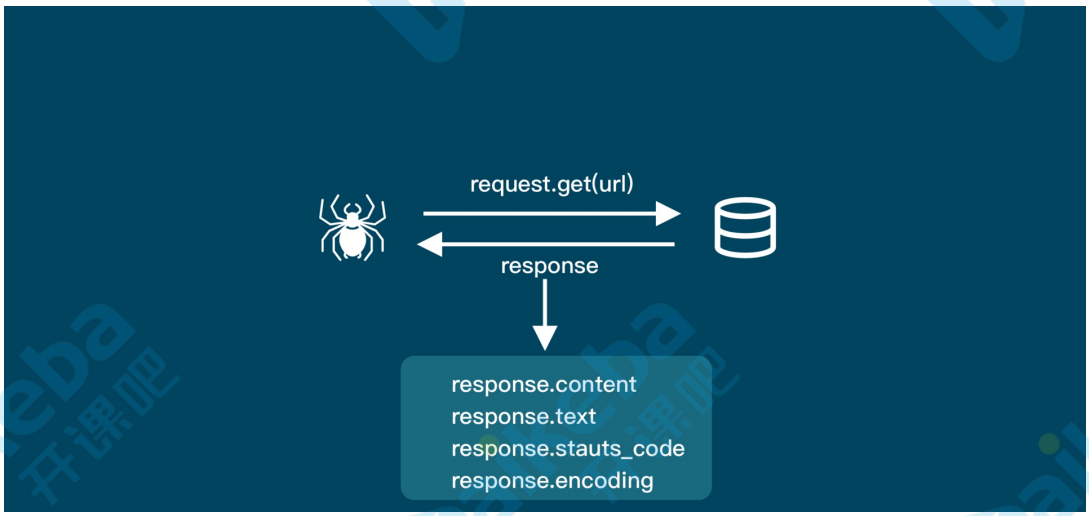
③、response.text

将 response.content 的二进制数据转换为字符串，适用于文字或者是网页源代码的下载。

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、网络世界中的爬虫规范

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。