01 导读：为什么要学爬虫？

更新时间：2019-07-02 19:05:43

> 人不可有傲气，但不可无傲骨。
>
> ——徐悲鸿

大家好，我是梁睿坤，是一家智能科技公司的联合创始人，从今天开始我们就将一起开始探讨 “网络爬虫” 这个话题了。

随着时代的发展，互联网成为了大量信息的载体，如何有效的获取这些信息成为了开发人员一个巨大的挑战。因为有了这样的需求所以网络爬虫就应运而生了：网络爬虫（又被称为网页蜘蛛，网络机器人，也被称为网页追逐者），是一种按照一定的规则，自动地抓取互联网信息的程序或者脚本。另外一些不常使用的名字还有蚂蚁、自动索引、模拟程序或者蠕虫。

简单来讲，爬虫就是一个探测机器，它的基本操作就是模拟人的行为去各个网站溜达，点点按钮，查查数据，把看到的信息给你捕获回来。就像一只虫子在一幢楼（爬取的网站）里不知疲倦地爬来爬去。可以说，有了爬虫，你就有了“分身术”。每一个爬虫都是你的分身，帮你在互联网上获取你所需要的数据。

我们日常生活中离不开的搜索引擎其实就是一个巨大的爬虫，当我们在百度的输入框中输入你想搜索的问题，并点击“百度一下”的时候，百度这个巨大的爬虫就会启动，并且会自动在互联网上根据你输入的关键词进行匹配，如果有匹配到的结果，爬虫就会把结果呈现在你的面前。

我们来举个例子：你是A公司的员工，B公司是你们的竞争对手。你的 Leader 让你获取竞争对手网站上商品的价格、简介还有购买人数等等信息来做竞品分析。这个时候你怎么办呢？手动去Ctrl+C和Ctrl+V？确实可以，但这是最笨的方法。数据量小的时候我们可以这样做，但是数据成千上万的时候你还要这样做吗？

你说可以找别人帮忙，但是你就不想自己动手写一个爬虫程序，在极短的时间内把你想要的数据抓取下来并且整理成数据表格，然后在你的 Leader 需要的时候甩到他的脸上吗？

我知道你想这么做，所以我为你准备了这个专栏，在这个专栏中我会详细的带着你一步一步的实现你自己的网络爬虫。当然了爬虫能做的不仅仅只有这些，它的实力很强大，在这里我也不一一举例了，在后面的小节里你会慢慢的发现它的魅力。

在学习爬虫之前你需要掌握以下的一些基础知识：

- 网络爬虫中常用的Python基础知识
- HTTP协议通信原理（我们在浏览网页的时候是怎样的一个过程，他是如何构成的？）
- HTML、CSS、JS入门基础（掌握网页结构以及从网页中定位具体的元素）

具备了这些基础，你就可以开始愉快的学习爬虫了。不过很多同学和朋友仍然有疑惑：学完爬虫之后有什么用呢？

在最新的编程语言排行榜上，Pyhton超越Java，成为了榜一，越来越多的程序员选择Python，甚至有人说，使用Python是“面向未来编程”。关于Python与“爬虫”的关系，我想不用多说你也能看出来爬虫的火热程度。

其次，掌握爬虫技术后，你会看到很多不同风景，在你使用爬虫爬取数据的过程中，你会感到非常好玩儿，相信我，这种趣味性和好奇心，会让你对Python有一种天生的喜爱感，为让你有深入学习Python的动力。

另外，在这个数据为王的时代，互联网上充斥着大量形形色色繁杂的数据，我们要从这个庞大的互联网上来获取到我们所需要的数据， 爬虫是不二之选。无论是过去的搜索引擎，还是时下火爆的数据分析，都离不开爬虫，除了好玩之外，爬虫是实实在在有非常多的用武之地的，事实上，很多公司在招聘时，对爬虫也是有要求的。

我们使用Python开发爬虫，Python最强大的地方不在于语言本身而是其庞大而活跃的开发者社区和上亿量级的第三方工具包。通过这些工具包我们可以快速的实现一个又一个的功能而不用我们自己去造轮子，掌握的工具包越多，我们在编写爬虫程序的时候也就越方便。另外，爬虫的工作目标是“互联网”，所以HTTP通信和HTML、CSS、JS这些技能在编写爬虫程序的时候都会用的到。不过不用担心，即使你对这些技术不太了解，在学习了本专栏之后也能够轻松的将这些知识运用到我们的爬虫程序中去。

作为开发人员，代码是最好的老师，在实践中学习，直接靠代码说话，是我们程序员的学习方式，所以，在设计这个专栏的时候我从众多素材中选出了几种具有代表性的课题，我们一起开发几种不同类型的爬虫，实际生产中，我们所需要的数据一般也逃不过这样的页面结构：

- 新闻供稿专用爬虫——爬取RSS订阅数据
- 网易新闻爬虫——泛爬网技术
- 网易爬虫优化——大规模数据处理技术
- 豆瓣读书爬虫——测试驱动设计与高级反爬技术实践
- 蘑菇街采集——处理深度继承javascript网站
- 慢速爬虫的应用举例——知乎爬虫

我会带着大家一一实现这些页面结构，实现技术各不相同的页面爬虫，让大家通过具体的代码实践了解在什么样的情况下可以采用什么样的技术来处理，遇到了反爬措施我们该如何去解决，通过具体应用建立起对爬虫的具体认知，再了解背后的技术理论。

那么讲到这可能有的同学要问了：编写完爬虫程序之后呢？不要着急，在编写完爬虫程序之后我还会带着大家将我们的爬虫程序部署，真正的让我们的爬虫“大展宏图”。关于网络爬虫的部署上线，很多其它的教程都没有讲解，但确实有很多同学对此有疑惑，不知道该如何部署。虽然看起来部署没有太多技术含量，但却是你不可不会的内容。

好了，讲到这里我就不再多言。随着本专栏内容的逐步展开，你会慢慢的进入一个新的领域，会发现网络爬虫不单单是一门简单的数据采集技术，而是一把能打开“全栈式开发”大门的钥匙。

还在等什么呢？现在就开始吧，我们一起学爬虫，一起玩爬虫，一起用爬虫吧。

02 网络爬虫基础知识准备及基本开发环境介绍