你好，我是你的爬虫老师崔庆才，欢迎来到我的专栏《52 讲轻松搞定网络爬虫》。最初接触和学习爬虫是兴趣使然，它看似简单却趣味无穷，入门容易但真的想做好也是需要下一番功夫的。

从 2015 年开始在博客上不断记录和分享自己的爬虫心得，受到读者挺多好评，至今博客阅读量已经过千万。2018 年我把爬虫知识结构化后出版了《Python3 网络爬虫开发实战》一书。这本书年销量 6w+ 册，是很多爬虫爱好者的启蒙教材，目前豆瓣评分是 9.0 分，京东上好评率也达到 99%。

但近期我看到了一些“负面评论”，主要集中在“由于反爬机制不断完善，好多案例已失效”。特地去看了下，我发现虽然网上有很多爬虫学习资料，但包括一些优质资料在内无一例外都存在这个问题。所以，我决定和拉勾教育联合推出一个“与时俱进”的爬虫实战专栏，希望可以帮到你。

为什么案例会失效?

随着大数据乃至人工智能的迅猛发展，数据变得越来越重要，甚至已成为很多企业赖以生存的根基。而想要获取数据，爬虫是必备工具之一。

前几年刮起的“全民学 Python”风，也促进了爬虫技术蓬勃发展，因为几乎所有 Python 课的实操案例都是“手把手教你写爬虫”。但发展的不止有爬虫技术，还有反爬技术和企业对数据保护的重视程度。你会发现之前学的爬虫案例过一段时间就失效了。

企业为了保护自己的数据不被轻易地爬取，采取了非常多的反爬虫措施，如 JavaScript 混淆和加密、App 加密、增强型验证码、封锁 IP、封锁账号等，甚至有不少企业有专门的更难破解的反爬措施。

为什么企业要求越来越高了?

数据爬取难度持续增大，也不完全是坏事，这让企业对爬虫工程师的需求量在逐步增多，薪资待遇也提升了不少。当然，技术要求也越来越高，例如 JavaScript、App 的逆向等几乎已经是爬虫工程师必备的技能，如果不懂，很多网站的数据是难以有效爬取的。另外，爬虫涉及的面很广，对计算机网络、编程基础、前端开发、后端开发、App 开发与逆向、网络安全、数据库、运维、机器学习、数据分析等方向也有一定的要求。

下面就几个爬虫目前所遇到的痛点来说一下。

- **比如，JavaScript 逆向。**很多网站为了保护数据不被轻易爬取到，会选择在前端进行一些保护，例如，将网站前端的代码进行加密或混淆，从而导致一些接口的请求难以直接用程序来模拟。如果要提高爬取效率，势必要对前端代码进行反混淆，进而进行数据爬取。
- **再比如，App 逆向。**移动互联网时代，许多公司会选择将数据放置于 App 端呈现，因此 App 也已经成了数据的重要载体。为了保护数据，企业会在数据接口中加入加密参数。这些加密参数的逻辑是写在 App 之中的，很多情况下，我们必须要对 App 进行逆向，才能分析出其中的逻辑，从而用爬虫进行模拟爬取。
- **还有，爬虫的运维和管理。**当爬虫数量较多的时候，如何方便地管理爬虫进程、如何进行定时任务的设置、如何进行扩容、如何进行监控、如何设置科学的报警机制变得非常重要。了解爬虫的运维和管理技巧，爬虫的管理才能不是难事。
- **识别验证码也常遇到。**现在很多网站都已经对接了各种各样的验证码，包括拖动、点选验证码等，如果不借助于人工方式识别，利用传统的算法是很难对此类验证码进行识别的。为了提高识别效率，有时候可能需要深度学习对此类验证码进行识别。如果掌握了深度学习的原理并将其应用于爬虫之上，会使你的爬虫技能如虎添翼。
- **最后说一下，网页的智能解析。**网页内容的解析在某些业务上是一件非常繁重的工作，现在很多人都会选择直接使用 XPath 等方式来解析，当网站类型变化多样的时候，单纯靠写 XPath 会耗费大量的精力，如果能有一种准确率较高的网页智能解析算法，那么网页的解析就会变得更加简单。

- **还有很多**，会在课程中详解，这里就不一一列举了。

这个课讲什么？怎么讲？

学习爬虫常有几大瓶颈：

- 不知道应该怎样从 0 到 1 完成数据爬取；
- 感觉爬虫体系太杂，无从下手；
- 难以应对各种反爬技术；
- 爬虫理论和项目难以很好结合

我会通过这个专栏带你一一突破，整个课程分为 7 大模块，从爬虫基础原理讲起：

- 体系化梳理爬虫整个技术栈涉及到的知识点，从易到难全面讲解；
- 讲解主流的爬虫和反爬应对措施，通过样例代码帮你了解爬虫基本用法和原理；
- 逐一详解多场景下数据爬取难题的解决方案；
- 一个知识点一个案例，带你实战演练，加深对爬虫技术的理解；
- 全面补充分享上述爬虫**新技术**，助你成为“爬虫高手”。

这个课适合你听吗？

在学习本门课程之前，最好是对 Python 有一定的基础了解，包括 Python 基本的语法和调用逻辑等。之所以没花篇幅去讲 Python，一个是很多人有 Python 基础，另一个是没有任何基础问题也不大，本课程会结合很多示例代码详细讲解，大家照着编写和学习，也能轻松理解，甚至成为爬虫大牛。

这门课是专门写给爬虫工程师的吗？当然不是，它适合所有有数据收集和获取需求的人。

- 比如，**学生和科研工作者**，如果在某些项目或研究上有数据需求，通过爬虫来获取是最省时省力行之有效的办法；
- 再比如，**企业从业者**，不论你是爬虫工程师还是产品运营或是别的什么，只要你的业务涉及数据需求，爬虫都会是你的好帮手；
- 还有，**学了 Python 想“变现”的人**，爬虫不仅可以帮你巩固Python知识，还可以让你实现学以致用，赚到钱；
- 总之，**对爬虫感兴趣的所有人**都可以订阅本专栏。

虽然爬虫涉及的知识点比较多，但经过我的系统梳理讲解和你的多加练习，相信你会对爬虫技术有全面透彻的理解，能应对绝大多数网站的爬取，加油！期待和你在爬虫的世界里一起进步，下节课我们开始学习爬虫基础原理，不见不散！