清华
科技大讲堂

Python 网络爬虫实战

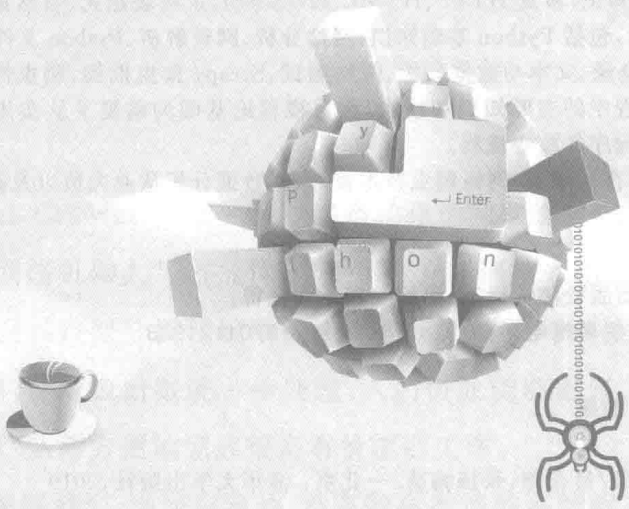
吕云翔 张扬 ◎ 编著

- 完整的项目案例
- 视频讲解
- 课件与源码

微课版
教学视频

清华大学出版社

清华
科技大讲堂
前沿·科技·分享



Python 网络爬虫实战

吕云翔 张扬 ◎ 编著

清华大学出版社
北京

内 容 简 介

本书介绍如何利用 Python 进行网络爬虫程序的开发，从 Python 语言的基本特性入手，详细介绍了 Python 爬虫开发的相关知识，涉及 HTTP、HTML、JavaScript、正则表达式、自然语言处理、数据科学等内容。全书共分为 14 章，包括 Python 基础知识、网站分析、网页解析、Python 文件的读写、Python 与数据库、AJAX 技术、模拟登录、文本与数据分析、网站测试、Scrapy 爬虫框架、爬虫性能等多个主题，内容覆盖网络抓取与爬虫编程中的主要知识和技术，在重视理论基础的前提下从实用性和丰富度出发，结合实例演示了编写爬虫程序的核心流程。

本书适合 Python 语言初学者、网络爬虫技术爱好者、数据分析从业人员以及高等院校计算机科学、软件工程等相关专业的师生阅读。

本书封面贴有清华大学出版社防伪标签，无标签者不得销售。
版权所有，侵权必究。侵权举报电话：010-62782989　13701121933

图书在版编目(CIP)数据

Python 网络爬虫实战/吕云翔，张扬编著. —北京：清华大学出版社，2019
(清华科技大讲堂)
ISBN 978-7-302-51592-0

Ⅰ. ①P…　Ⅱ. ①吕… ②张…　Ⅲ. ①软件工具—程序设计　Ⅳ. ①TP311.561

中国版本图书馆 CIP 数据核字(2018)第 257255 号

策划编辑：魏江江
责任编辑：王冰飞
封面设计：刘　键
责任校对：时翠兰
责任印制：丛怀宇

出版发行：清华大学出版社
网　　址：http://www.tup.com.cn，http://www.wqbook.com
地　　址：北京清华大学学研大厦 A 座　　**邮　　编**：100084
社 总 机：010-62770175　　**邮　　购**：010-62786544
投稿与读者服务：010-62776969，c-service@tup.tsinghua.edu.cn
质量反馈：010-62772015，zhiliang@tup.tsinghua.edu.cn
课件下载：http://www.tup.com.cn，010-62795954
印 装 者：清华大学印刷厂
经　　销：全国新华书店
开　　本：185mm×260mm　　**印　　张**：25.5　　**字　　数**：433 千字
版　　次：2019 年 5 月第 1 版　　**印　　次**：2019 年 5 月第1次印刷
印　　数：1～2000
定　　价：79.80 元

产品编号：075108-01

前言

网络爬虫(Web Crawler)是指一类能够自动化访问网络并抓取某些信息的程序，有时候也被称为“网络机器人”。它们被广泛用于互联网搜索引擎及各种网站的开发中，同时也是大数据和数据分析领域中的重要角色。爬虫可以按一定的逻辑大批量采集目标页面内容，并对数据做进一步处理，人们借此能够更好、更快地获得并使用他们感兴趣的信息，从而方便地完成很多有价值的工作。

Python是一种解释型、面向对象的、动态数据类型的高级程序设计语言，Python语法简洁、功能强大，在众多高级语言中拥有十分出色的编写效率，同时还拥有活跃的开源社区和海量程序库，十分适合进行网络内容的抓取和处理。本书将以Python语言为基础，由浅入深地探讨网络爬虫技术，同时通过具体的程序编写和实践来帮助读者了解和学习Python爬虫。

本书共分为14章，其中第1～3章为基础篇，第4～6章为进阶篇，第7～9章为高级篇，第10～14章为实践篇，最后为附录。第1章、第2章介绍了Python语言和编写爬虫程序的基础知识；第3章讨论了Python中对文件和数据的存储，涉及数据库的相关知识；第4章、第5章的内容针对相对复杂一些的爬虫抓取任务，主要着眼于动态内容和表单登录等方面；第6章涉及对抓取到的原始数据的深入处理和分析；第7～9章旨在从不同视角讨论爬虫程序，基于爬虫介绍了多个不同主题的内容；第10～14章通过一些实际的例子深入讨论爬虫编程的理论知识；最后在附录中介绍了Python语言和爬虫编程中常用的知识和工具。

本书的主要特点如下。

- 内容全面，结构清晰。本书详细介绍了网络爬虫技术的方方面面，讨论了数据抓取、数据处理和数据分析的整个流程。全书结构清晰，坚持理论知识与实践操作相结合。
- 循序渐进，生动简洁。本书从最简单的Python程序示例开始，在网络爬虫的核心主题之下一步步深入，兼顾内容的广度与深度，在内容编写上使用生动

简洁的阐述方式，力争详略得当。

- 示例丰富，实战性强。网络爬虫是实践性、操作性非常强的技术，本书将提供丰富的代码作为读者的参考，同时对必要的术语和代码进行解释。本书从生活实际出发，选取实用性、趣味性兼具的主题进行网络爬虫实践。
- 内容新颖，不落窠臼。本书中的程序代码均采用最新的 Python 3 版本，并使用了目前主流的各种 Python 框架和库来编写程序，注重内容的先进性。学习网络爬虫需要动手实践才能真正理解，本书最大限度地保证了代码与程序示例的易用性和易读性。

本书在第 10～14 章，针对 5 个爬虫实践，配有微课视频讲解，以方便读者更好地理解 Python 爬虫相关的理论和实践知识。

本书的编者为吕云翔、张扬，曾洪立参与了部分内容的编写及资料整理工作。

由于编者的水平有限，书中的不足在所难免，恳请广大读者批评指正。

编　者

2019 年 1 月

试读结束：需要全本请在线购买：　www.ertongbook.com

目 录

本书源码下载

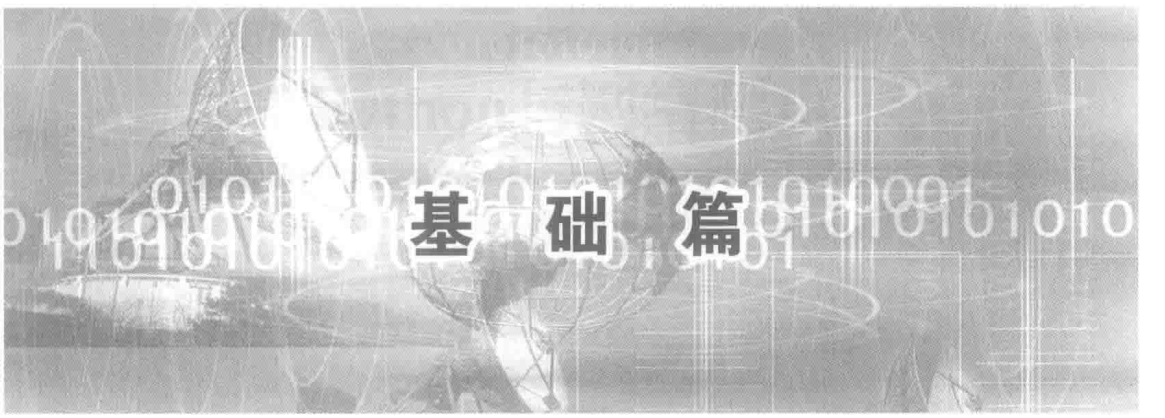
基 础 篇

进 阶 篇

高 级 篇

实 践 篇

试读结束：需要全本请在线购买： www.ertongbook.com


基 础 篇

第1章

Python与网络爬虫

网络爬虫(Web Crawler)有时候也叫网络蜘蛛(Web Spider),是指这样一类程序——它们可以自动连接到互联网站点,读取网页中的内容或者存放在网络上的各种信息,并按照某种策略对目标信息进行采集(例如对某个网站的全部页面进行读取)。实际上,世界上最大的搜索网站——Google搜索本身就建构在爬虫技术之上,像Google、百度这样的搜索引擎会通过爬虫程序来不断更新自身的网站内容和对其他网站的网络索引。从某种意义上说,用户每次通过搜索引擎查询一个关键词,就是在搜索引擎服务者的爬虫程序所"爬"到的信息中进行查询。当然,搜索引擎背后所使用的技术十分复杂,其爬虫技术通常也不是一般个人开发的小型程序所能比拟的。其实,爬虫程序本身并不复杂,用户只要懂一点编程知识,了解一点HTTP和HTML,就可以写出属于自己的爬虫,实现很多有意思的功能。

在众多编程语言中,本书选择Python来编写爬虫程序,因为Python不仅语法简洁、便于上手,而且拥有庞大的开发者社区和浩如烟海的模块库,对于普通的程序编写而言有极大的便利。虽然Python和C/C++等语言相比可能在性能上有所欠缺,但毕竟瑕不掩瑜,是目前最好的选择。

1.1 Python 语言

Python 是目前最流行的编程语言之一，本书对它的历史和发展作一些简单介绍，然后看看 Python 的基本语法，对于没有 Python 编程经验的读者而言，可以借此对 Python 有一个初步的了解。

1.1.1 什么是 Python

Guido van Rossum 在 1989 年发明了 Python，而 Python 的第一个公开发行版发行于 1991 年。因为 Guido 是电视剧 *Monty Python's Flying Circus* 的爱好者，所以将这种新的脚本语言命名为 Python。

从最根本的角度来说，Python 是一种解释型、面向对象的、动态数据类型的高级程序设计语言。值得注意的是，Python 是开源的，源代码遵循 GPL(GNU General Public License)协议，这就意味着它对所有个人开发者是完全开放的，这也使得 Python 在开发者中迅速流行开来，来自全球各地的 Python 使用者为这门语言的发展贡献了很多力量。Python 的哲学是优雅、明确和简单。著名的 *the Zen of Python*（Python 之禅）[①]这样说道：

“

优美胜于丑陋，

明了胜于晦涩，

简洁胜于复杂，

复杂胜于凌乱，

扁平胜于嵌套，

间隔胜于紧凑，

可读性很重要。

即便假借特例的实用性之名，也不可违背这些规则，

不要包容所有错误，除非你确定需要这样做，

① 作者为 Tim Peters，英文原文可见“https://www.python.org/dev/peps/pep-0020/”。

试读结束：需要全本请在线购买： www.ertongbook.com