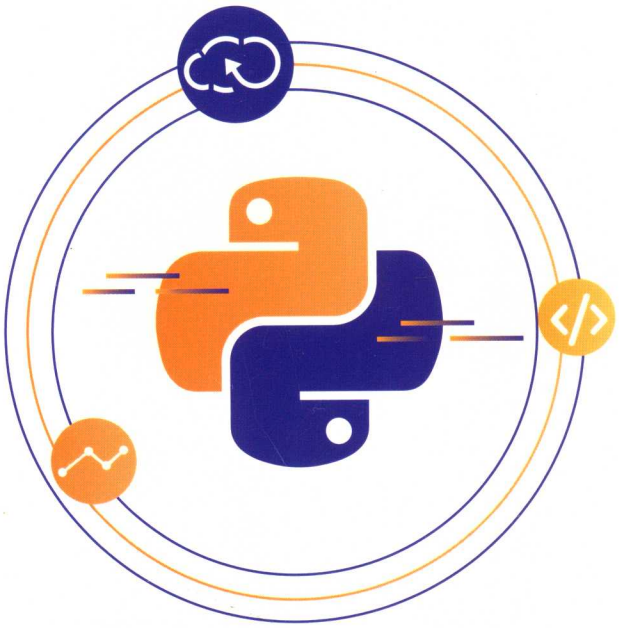
Python 开发从入门到精通系列



Python

web crawler from entry to proficiency

吕云翔　张扬　韩延刚　等／编著

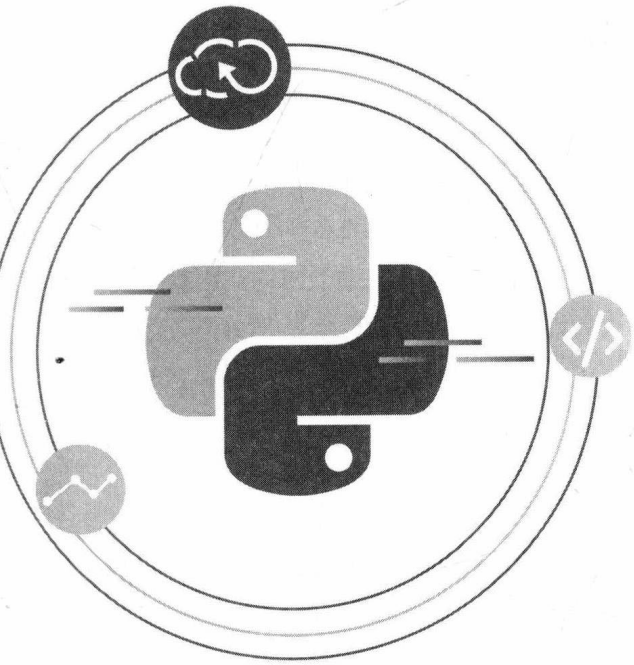
本书采用Python 3.6版本编写

Python 网络爬虫 从入门到精通

关注机械工业出版社计算机分社官方微信订阅号“IT有得聊”
即可获得本书配套案例程序的源代码、微课视频以及PPT电子教案

机械工业出版社
CHINA MACHINE PRESS

Python 开发从入门到精通系列



Python

web crawler from entry to proficiency

吕云翔　张扬　韩延刚　等／编著

本书采用Python 3.6版本编写

Python 网络爬虫 从入门到精通

机械工业出版社
CHINA MACHINE PRESS

本书的主旨是介绍如何结合 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 网络爬虫开发的各个方面，涉及 HTTP、HTML、JavaScript、正则表达式、自然语言处理、数据科学等不同领域的内容。全书共 15 章，包括 Python 基础知识、网站分析、网页解析、Python 文件读写、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。本书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，从实用性和丰富性出发，结合实例演示了爬虫编写的核心流程。

本书适合 Python 语言初学者、网络爬虫技术爱好者、数据分析从业人士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。

图书在版编目（CIP）数据

Python 网络爬虫从入门到精通 / 吕云翔等编著. —北京：机械工业出版社，2019.6
（Python 开发从入门到精通系列）
ISBN 978-7-111-62593-3

Ⅰ.①P… Ⅱ.①吕… 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（2019）第 079060 号

机械工业出版社（北京市百万庄大街 22 号 邮政编码 100037）
策划编辑：张淑谦 责任编辑：张淑谦 赵小花
责任校对：张艳霞 责任印制：孙 炜
保定市中画美凯印刷有限公司印刷
2019 年 5 月第 1 版 • 第 1 次印刷
184mm×260mm • 21.5 印张 • 505 千字
0001—3000 册
标准书号：ISBN 978-7-111-62593-3
定价：79.00 元

凡购本书，如有缺页、倒页、脱页，由本社发行部调换

电话服务
服务咨询热线：（010）88361066
读者购书热线：（010）68326294

网络服务
机 工 官 网：www.cmpbook.com
机 工 官 博：weibo.com/cmp1952
金 书 网：www.golden-book.com
教育服务网：www.cmpedu.com

前言

网络爬虫又叫网络蜘蛛，是指按照某种规则在网络上爬取所需内容的脚本程序。它们被广泛应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。众所周知，每个网页通常包含其他网页的入口，网络爬虫则通过一个网址依次进入其他网址获取所需内容。爬虫可以按一定逻辑大批量采集目标页面内容，并对数据进行进一步处理，人们借此能够更好、更快地获取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、动态数据类型的高级程序设计语言，其语法简洁、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，十分适合用于网络内容的抓取和处理。本书以 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 网络爬虫。

本书共 15 章，分 4 部分讲解，其中第 1～3 章为基础部分，第 4～6 章为进阶部分，第 7～9 章为高级部分，第 10～15 章为实践部分。第 1、2 章介绍了 Python 语言和爬虫编写的基础知识；第 3 章讨论了 Python 中文件和数据的存储，涉及数据库的相关知识；第 4、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 章探讨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 10～15 章通过一些实际的例子深入讨论了爬虫编程的理论知识。

本书的主要特点有：

- 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，围绕网络爬虫这个主题一步步深入讲解，兼顾内容的广度与深度。本书行文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，本书提供丰富的代码供读者参考，同时对必要的术语和代码进行解释。从实际应用出发，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。本书的程序代码均基于最新的 Python 3 版本编写，并使用了目

前主流的各种 Python 框架和库，注重内容的时效性。网络爬虫需要动手实践才能真正理解，本书最大程度地保证了代码与程序示例的易用性和易读性。

本书主要由吕云翔、张扬、韩延刚编写，另外，曾洪立参与了部分内容的编写及资料整理工作。

由于作者水平有限，不足之处在所难免，欢迎广大读者联系交流（邮箱：yunxianglu@hotmail.com）。

编　者

目录

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），它是指这样一类程序——它们可以自动连接到互联网站点，并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（如对某个网站的全部页面进行读取）。实际上，像Google、百度这样的搜索引擎就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。某种意义上说，用户每次通过搜索引擎查询一个关键词，就是在搜索引擎提供者的爬虫程序所“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所开发的小型程序所能比拟的。不过，爬虫程序本身其实并不复杂，只要懂一些编程知识，了解一些HTTP和HTML，就可以写出属于自己的爬虫程序，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。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非常便利。虽然Python与C/C++等语言相比可能在性能上有所欠缺，但毕竟瑕不掩瑜，开发人员普遍认为它是目前编写网络爬虫程序的最好选择。

1.1 Python语言

Python是目前最为流行的编程语言之一，本章首先对它的历史和发展做一些简单介绍，然后再介绍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年开发了Python语言，而Python的第一个公开发行版发行于1991年。因为Guido是一部电视剧《Monty Python's Flying Circus》的爱好者，因此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、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“Zen of Python”（Python 之禅）[⊖]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，
当存在多种可能，不要尝试去猜测，
而是尽量找一种，最好是唯一一种明显的解决方案，
虽然这并不容易，因为你不是 Python 之父。
做也许好过不做，但不假思索就动手还不如不做。
如果你无法向人描述你的方案，那肯定不是一个好方案；反之亦然。
命名空间是一种绝妙的理念，我们应当多加利用。

2000 年 Python 2.0 版本发布，Python 3.0 版本则于 2008 年发布，这一新版本不完全兼容之前的 Python 源代码。目前开发者主要接触到的是 Python 2.7 与 Python 3.5，以及更新一点的 Python 3.6。Python 3 在 Python 2 的基础上做出了不少很有价值的改进，3.5 和 3.6 也已逐步成为 Python 的主流版本，本书将完全使用 Python 3 作为开发语言。

1.1.2 Python 的应用现状

Python 的应用范围十分广泛，著名的应用案例有以下几个。

- Reddit：社交分享网站，美国最热门的网站之一。
- Dropbox：文件分享服务。
- Pylons：Web 应用框架。
- TurboGears：另一个 Web 应用快速开发框架。
- Fabric：用于管理Linux主机的程序库。
- Mailman：使用 Python 编写的邮件列表软件。
- Blender：以 C 与 Python 开发的开源 3D 绘图软件。

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

国内的例子也很多，著名的豆瓣网（国内一家很受欢迎的社区网站）和知乎（国内一家很受欢迎的网络问答社区）都大量使用了 Python 进行开发。可见，Python 在业界的应用可谓五花八门，总结起来，在系统编程、图形处理、科学计算、数据库、网络编程、Web 应用、多媒体应用等各个方面都有它的身影。在 2017 年的 IEEE Spectrum Ranking 中[㊀]，Python 力压群雄，成为最流行的编程语言。众所周知，学习一门程序语言最有效的方法就是边学边用，边用边学。通过对 Python 网络爬虫的逐步学习，相信能够很好地提高读者对整个 Python 语言的理解和应用。

【提示】 为什么要使用 Python 来编写爬虫程序？Python 的简明语法和各式各样的开源库使得 Python 在网络爬虫方向得天独厚，尤其对于个人开发爬虫程序而言，一般对于性能的要求不会太高，因此，虽然人们一般认为 Python 在性能上难以与 C/C++和 Java 相比，但总的来说，使用 Python 有助于更好、更快地实现开发者所需要的功能。另外，考虑到 Python 社区贡献了很多各有特色的库，很多都能直接用来编写爬虫程序，因此，Python 的确是目前更好的选择。

1.2　Python 的安装与开发环境配置

在开始探索 Python 之前，读者首先需要在自己的机器上安装 Python。值得高兴的是，Python 不仅免费、开源，而且坚持轻量级，安装过程并不复杂。如果使用 Linux 系统，其中可能已经内置了 Python（虽然版本有可能是较旧的）；使用苹果电脑（Mac OS 系统）的话，一般也已经安装了命令行版本的 Python 2.x。在 Linux 或 Mac OS X 系统上检测 Python 3 是否安装的最简单办法是使用终端命令，即在 terminal 应用中输入“Python3”命令并按〈Enter〉键执行，观察是否有对应的提示出现。至于 Windows 系统，在目前最新的 Windows 10 版本上并没有内置 Python，因此必须手动安装。

1.2.1　在 Windows 上安装

访问 python.org/download/并下载与计算机架构对应的 Python 3 安装程序，一般而言只要有新版本，就应该选择最新的版本。这里需要注意的是选择对应架构的版本前，读者需要首先搞清楚自己的系统是 32 位还是 64 位的，如图 1-1 所示。

根据安装程序的导引一步步执行，就能完成整个安装。如果最终看到类似图 1-2 所示的提示，就说明安装成功了。

这时检查“开始”菜单，就能看到 Python 3.x 的应用程序。如图 1-3 所示，其中有一个“IDLE”（Integrated Development Environment”）程序，单击此项目后就可以开始在交互式窗口中使用 Python Shell 了，如图 1-4 所示。

㊀ 可见 http://adtmag.com/articles/2017/07/24/ieee-spectrum-ranking.aspx。

Windows x86-64 embeddable zip file	Windows	for AMD64/EM64T/x64	04cc4f6f6a14ba74f6ae1a8b685ec471	7190516	SIG
Windows x86-64 executable installer	Windows	for AMD64/EM64T/x64	9e96c934f5d16399f860812b4ac7002b	31776112	SIG
Windows x86-64 web-based installer	Windows	for AMD64/EM64T/x64	640736a3894022d30f7babff77391d6b	1320112	SIG
Windows x86 embeddable zip file	Windows		b0b099a4fa479fb37880c15f2b2f4f34	6429369	SIG
Windows x86 executable installer	Windows		2bb6ad2ecca6088171ef923bca483f02	30735232	SIG
Windows x86 web-based installer	Windows		596667cb91a9fb20e6f4f153f3a213a5	1294096	SIG

图 1-1　Python.org/download 页面（部分）

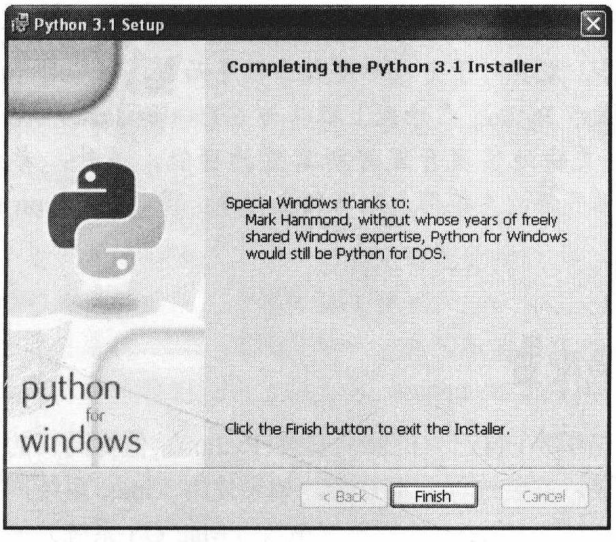


图 1-2　Python 安装成功的提示

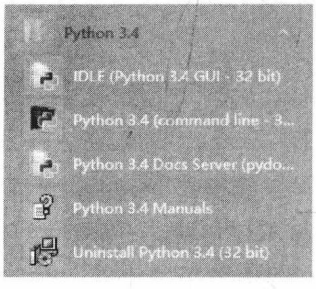


图 1-3　安装完成后的“开始”菜单

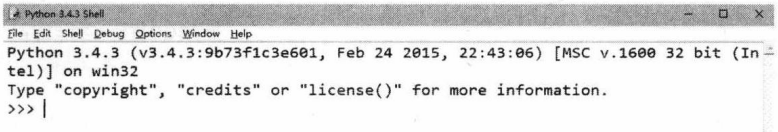


图 1-4　IDLE 的界面

1.2.2　在 Ubuntu 和 Mac OS 上安装

Ubuntu 是诸多 Linux 发行版中受众较多的一个系列。通过“Applicatons”（应用程序）中的添加应用程序进行安装，在其中搜索 Python 3，并在结果中找到对应的包进行下载。如果安装成功，大家将在“Applications”中找到 Python IDLE，单击后进入 Python Shell 中。

访问 python.org/download/并下载对应的 Mac OS 平台安装程序，根据安装包的指示进行操作，最后将看到类似图 1-5 所示的成功提示。

图 1-5　Mac OS 上的安装成功提示

关闭该窗口，并进入“Applications”（或者是从 LaunchPad 页面打开）中，就能找到 Python Shell，启动该程序，看到的结果应该和 Windows 平台上的结果类似。

1.2.3　PyCharm 的使用

虽然 Python 自带的 IDLE Shell 是绝大多数人对 Python 的第一印象，但如果通过 Python 语言编写程序、开发软件，它并不是唯一的工具，很多人更愿意使用一些特定的编辑器或者由第三方提供的集成开发环境（IDE）。借助 IDE 可以提高开发效率，但对开发者而言，只有最适合自己的，没有“最好的”，习惯一种工具后再接受另外一种总是不容易的。这里再简单介绍一下 PyCharm——一个由 JetBrain 公司出品的 Python 开发工具，并谈谈它的安装和配置。

在官网中可以下载到该软件：

```
https://www.jetbrains.com/pycharm/download/#section=windows
```

PyCharm 支持 Windows、Mac OS、Linux 三大平台，并提供 Professional 和 Community Edition 两种版本（见图 1-6）。其中前者需要购买正版（提供免费试用），后者可以直接下载使用。前者功能更为丰富，但后者也足以满足一些普通的开发需求。

选择对应的平台并下载后，安装程序（见图 1-7）将会引导用户完成安装。安装完成后，从“开始”菜单中（对于 Mac OS 和 Linux 系统是从“Applications”中）打开 PyCharm，就可以创建自己的第一个 Python 项目了（见图 1-8）。

创建项目后，还需要进行一些基本的配置。可以在菜单栏中使用“File”→“Settings”打开 PyCharm 设置窗口。

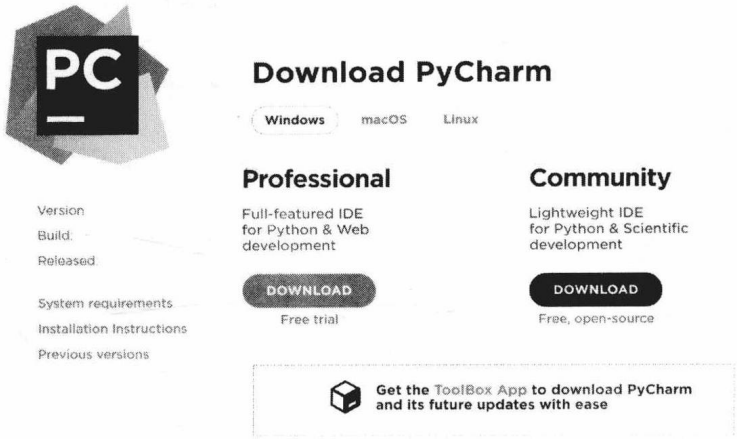


图 1-6　PyCharm 的下载页面

图 1-7　PyCharm 安装程序（Windows 平台）

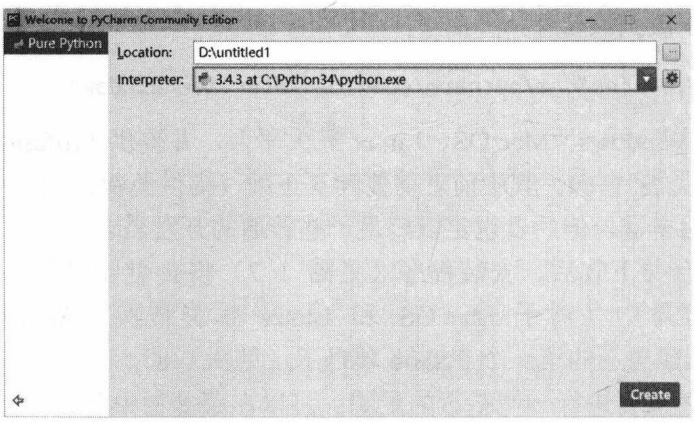


图 1-8　在 PyCharm 中创建新项目