HZ BOOKS
华章IT

零基础学习爬虫技术，从Python和Web前端基础开始讲起，由浅入深，包含大量案例，实用性强。

从静态网站到动态网站，从单机爬虫到分布式爬虫，涵盖Scrapy和PySpider框架的运用、去重方案的设计和分布式爬虫的搭建等。

實戰

范传辉 编著

The Fighting of Python Spider

Python爬虫
开发与项目实战

机械工业出版社
China Machine Press

實戰

随着大数据时代到来，网络信息量也变得更多更大，基于传统搜索引擎的局限性，网络爬虫应运而生。本书从基本的爬虫原理开始讲解，通过介绍Pthyon编程语言和Web前端基础知识引领读者入门，之后介绍动态爬虫原理以及Scrapy爬虫框架，最后介绍大规模数据下分布式爬虫的设计以及PySpider爬虫框架等。

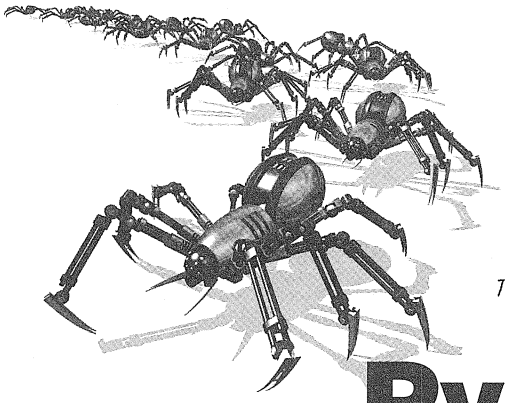
本书主要特点：

- 由浅入深，从Python和Web前端基础开始讲起，逐步加深难度，层层递进。
- 内容详实，从静态网站到动态网站，从单机爬虫到分布式爬虫，既包含基础知识点，又讲解了关键问题和难点分析，方便读者完成进阶。
- 实用性强，本书共有9个爬虫项目，以系统的实战项目为驱动，由浅及深地讲解爬虫开发中所需的知识和技能。
- 难点详析，对js加密的分析、反爬虫措施的突破、去重方案的设计、分布式爬虫的开发进行了细致的讲解。

上架指导：计算机\程序设计
ISBN 978-7-111-56387-7
9 787111 563877
定价：79.00元

投稿热线：(010) 88379604
客服热线：(010) 88379426 88361066
购书热线：(010) 68326294 88379649 68995259

华章网站：www.hzbook.com
网上购书：www.china-pub.com
数字阅读：www.hzmedia.com.cn



The Fighting of Python Spider

Python爬虫
开发与项目实战

范传辉 编著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Python 爬虫开发与项目实战 / 范传辉编著 . —北京：机械工业出版社，2017.3
（实战）

ISBN 978-7-111-56387-7

I. P… II. 范… III. 软件工具 – 程序设计 IV. TP311.561

中国版本图书馆 CIP 数据核字（2017）第 061009 号

Python 爬虫开发与项目实战

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）
责任编辑：吴 怡 责任校对：殷 虹
印 刷：北京文昌阁彩色印刷有限责任公司 版 次：2017 年 6 月第 1 版第 1 次印刷
开 本：186mm × 240mm 1/16 印 张：27.25
书 号：ISBN 978-7-111-56387-7 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
客服热线：（010）88379426 88361066 投稿热线：（010）88379604
购书热线：（010）68326294 88379649 68995259 读者信箱：hzit@hzbook.com

Preface 前　言

为什么写这本书

当你看前言的时候，不得不说你做出了一个聪明的选择，因为前言中有作者对整本书的概括和学习建议，这会对大家之后的阅读产生事半功倍的效果。在聊这本书之前，首先给大家一个本书所有配套源码和说明的链接：https://github.com/qiyeboy/SpiderBook。大家可以在Github中对不懂的内容进行提问，我会尽可能地帮助大家解决问题。其实在前言开头放这个链接是挺突兀的，不过确实是担心大家不会完整地看完前言。

接下来聊一聊这本书，写这本书的原因来自于我个人的微信公众号：七夜安全博客。我经常在博客园、知乎和微信平台上发布技术文章，分享一些知识和见解，有很多热心的朋友愿意和我进行交流讨论。记得2016年4月初的某一天，有一个朋友在微信后台留言，问我怎样将Python爬虫技术学好，有什么书籍可以推荐。我当时回答了好长一段建议，但是那个朋友依然希望能推荐一本书籍帮助入门和提高。其实我特别能理解初学者的心情，毕竟我也是从初学者走过来的，但是确实挺纠结，不知从何推荐。于是，我专门找了一下这方面的书籍，只找到一本外国人写的书，中文版刚出版没多久，名字为《Python 网络数据采集》。我花了半天看了一下里面的内容，整本书条理比较清晰，容易理解，但是很多知识点都谈得很浅，系统的实战项目基本上没有，更多的是一些代码片段，仅仅适合一些刚刚入门的朋友。自从这件事情以后，我就下定决心写一本Python爬虫方面的书籍，既然国内还没有人写这方面的书籍，我愿意做一个抛砖引玉的人，帮助大家更好地学习爬虫技术。

有了写书的想法后，开始列提纲，确定书的主题和内容。由于爬虫是一项实践性很强的技术，因此书的主题是以实战项目为驱动，由浅及深地讲解爬虫技术，希望你看这本书的时候是个菜鸟，认真学习完之后不再是个菜鸟，可以自主地开发Python爬虫项目了。从写书的那一刻开始，我就知道在书写完之前，我应该是没有周末了。这本书写了大半年的时间，由

于我平时有写笔记、做总结的习惯，因此写书的时间不是特别长，不过直到 2017 年年初我依然在更新内容，毕竟爬虫技术更新得比较快，我努力将比较新的知识贡献给大家。

在写书的过程中，我的内心变得越来越平静，越来越有耐心，不断地修改更新，对每个实战项目进行反复验证和敲定，尽可能地贴近初学者的需求，希望能帮助他们完成蜕变。

最后做一下自我介绍，本人是一位信息安全研究人员，比较擅长网络安全、软件逆向，同时对大数据、机器学习和深度学习有非常浓厚的兴趣，欢迎大家和我交流，共同进步。

前路多艰，学习的道路不可能一帆风顺，爬虫技术只是个开始，愿与诸君一道共克难关。

本书结构

本书总共分为三个部分：基础篇、中级篇和深入篇。

基础篇包括第 1 ~ 7 章，主要讲解了什么是网络爬虫、如何分析静态网站、如何开发一个完整的爬虫。

第 1 ~ 2 章帮助大家回顾了 Python 和 Web 方面的知识，主要是为之后的爬虫学习打下基础，毕竟之后要和 Python、Web 打交道。

第 3 ~ 5 章详细介绍了什么是网络爬虫、如何分析静态网站、如何从 HTML 页面中提取出有效的数据，以及对如何将数据合理地存储成各类文件以实现持久化。

第 6 ~ 7 章包含了两个实战项目。第一个项目是基础爬虫，也就是一个单机爬虫，功能是爬取百度百科的词条，并据此讲解了一个爬虫所应该具有的全部功能组件以及编码实现。第二个项目是分布式爬虫，功能和基础爬虫一致，在单机爬虫的基础上进行分布式改进，帮助大家从根本上了解分布式爬虫，消除分布式爬虫的神秘感。

中级篇包括第 8 ~ 14 章，主要讲解了三种数据库的存储方式、动态网站的抓取、协议分析和 Scrapy 爬虫框架。

第 8 章详细介绍了 SQLite、MySQL 和 MongoDB 三种数据库的操作方式，帮助大家实现爬取数据存储的多样化。

第 9 章主要讲解了动态网站分析和爬取的两种思路，并通过两个实战项目帮助大家理解。

第 10 章首先探讨了爬虫开发中遇到的两个问题——登录爬取问题和验证码问题，并提供了解决办法和分析实例。接着对 Web 端的爬取提供了另外的思路，当在 PC 网页端爬取遇到困难时，爬取方式可以向手机网页端转变。

第 11 章接着延伸第 10 章的问题，又提出了两种爬取思路。当在网页站点爬取遇到困难时，爬取思路可以向 PC 客户端和移动客户端转变，并通过两个实战项目帮助大家了解实施过程。

第 12 ~ 14 章由浅及深地讲解了著名爬虫框架 Scrapy 的运用，并通过知乎爬虫这个实战项目演示了 Scrapy 开发和部署爬虫的整个过程。

深入篇为第 15 ~ 18 章，详细介绍了大规模爬取中的去重问题以及如何通过 Scrapy 框架开发分布式爬虫，最后又介绍了一个较新的爬虫框架 PySpider。

第 15 章主要讲解了海量数据的去重方式以及各种去重方式的优劣比较。

第 16 ~ 17 章详细介绍了如何通过 Redis 和 Scrapy 的结合实现分布式爬虫，并通过云起书院实战项目帮助大家了解整个的实现过程以及注意事项。

第 18 章介绍了一个较为人性化的爬虫框架 PySpider，并通过爬取豆瓣读书信息来演示其基本功能。

以上就是本书的全部内容，看到以上介绍之后，是不是有赶快阅读的冲动呢？不要着急，接着往下看。

本书特点及建议

本书总体来说是一本实战型书籍，以大量系统的实战项目为驱动，由浅及深地讲解了爬虫开发中所需的知识和技能。本书是一本适合初学者的书籍，既有对基础知识点的讲解，也涉及关键问题和难点的分析和解决，本书的初衷是帮助初学者夯实基础，实现提高。还有一点要说明，这本书对编程能力是有一定要求的，希望读者尽量熟悉 Python 编程。

对于学习本书有两点建议，希望能引起读者的注意。第一点，读者可根据自己的实际情况选择性地学习本书的章节，假如之前学过 Python 或者 Web 前端的知识，前两章就可以蜻蜓点水地看一下。第二点，本书中的实战项目是根据当时网页的情况进行编写的，可能当书籍出版的时候，网页的解析规则发生改变而使项目代码失效，因此大家从实战项目中应该学习分析过程和编码的实现方式，而不是具体的代码，授人以渔永远比授人以鱼更加有价值，即使代码失效了，大家也可以根据实际情况进行修改。

致谢

写完这本书，才感觉到写书不是一件容易的事情，挺耗费心血的。不过除此之外，更多的是一种满足感，像一种别样的创业，既紧张又刺激，同时也实现了我分享知识的心愿，算是做了一件值得回忆的事情。这是我写的第一本书，希望是一次有益的尝试。

感谢父母的养育之恩，是他们的默默付出支持我走到今天。

感谢我的女朋友，在每个写书的周末都没有办法陪伴她，正是她的理解和支持才让我如此准时地完稿。

感谢长春理工大学电子学会实验室，如果没有当年实验室的培养，没有兄弟们的同甘共苦，就没有今天的我。

感谢西安电子科技大学，它所营造的氛围使我的视野更加开阔，使我的技术水平更上一层楼。

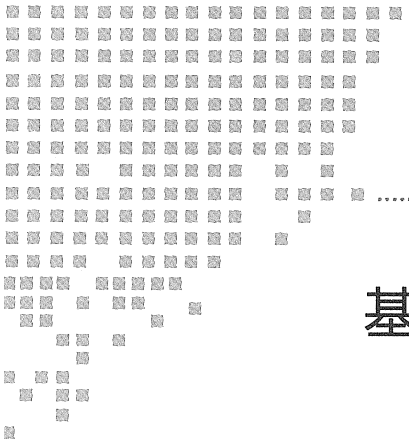
感谢机械工业出版社的吴怡编辑，没有她的信任和鼓励，就没有这本书的顺利出版。

感谢 Python 中文社区的大力支持。

感谢本书中所用开源项目的作者，正是他们无私的奉献才有了开发的便利。

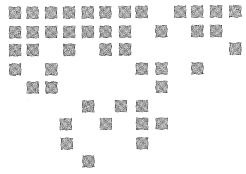
由于作者水平有限，书中难免有误，欢迎各位业界同仁斧正！

Contents 目　录



基础篇

- 第 1 章　回顾 Python 编程
- 第 2 章　Web 前端基础
- 第 3 章　初识网络爬虫
- 第 4 章　HTML 解析大法
- 第 5 章　数据存储（无数据库版）
- 第 6 章　实战项目：基础爬虫
- 第 7 章　实战项目：简单分布式爬虫



Chapter 1 第 1 章

回顾 Python 编程

本书所要讲解的爬虫技术是基于 Python 语言进行开发的，拥有 Python 编程能力对于本书的学习是至关重要的，因此本章的目标是帮助之前接触过 Python 语言的读者回顾一下 Python 编程中的内容，尤其是与爬虫技术相关的内容。

1.1 安装 Python

Python 是跨平台语言，它可以运行在 Windows、Mac 和各种 Linux/Unix 系统上。在 Windows 上编写的程序，可以在 Mac 和 Linux 上正常运行。Python 是一种面向对象、解释型计算机程序设计语言，需要 Python 解释器进行解释运行。目前，Python 有两个版本，一个是 2.x 版，一个是 3.x 版，这两个版本是不兼容的。现在 Python 的整体方向是朝着 3.x 发展的，但是在发展过程中，大量针对 2.x 版本的代码都需要修改才能运行，导致现在许多第三方库无法在 3.x 版本上直接使用，因此现在大部分的云服务器默认的 Python 版本依然是 2.x 版。考虑到上述原因，本书采用的 Python 版本为 2.x，确切地说是 2.7 版本。

1.1.1 Windows 上安装 Python

首先，从 Python 的官方网站 www.python.org 下载最新的 2.7.12 版本，地址是 https://www.python.org/ftp/python/2.7.12/python-2.7.12.msi。然后，运行下载的 MSI 安装包，在选择安装组件时，勾选上所有的组件，如图 1-1 所示。

特别要注意勾选 pip 和 Add python.exe to Path，然后一路点击 Next 即可完成安装。

pip 是 Python 安装扩展模块的工具，通常会用 pip 下载扩展模块的源代码并编译安装。

Add python.exe to Path 是将 Python 添加到 Windows 环境中。

安装完成后，打开命令提示窗口，输入 python 后出现如图 1-2 情况，说明 Python 安装成功。

当看到提示符“>>>”就表示我们已经在 Python 交互式环境中了，可以输入任何 Python 代码，回车后会立刻得到执行结果。现在，输入 exit()并回车，就可以退出 Python 交互式环境。

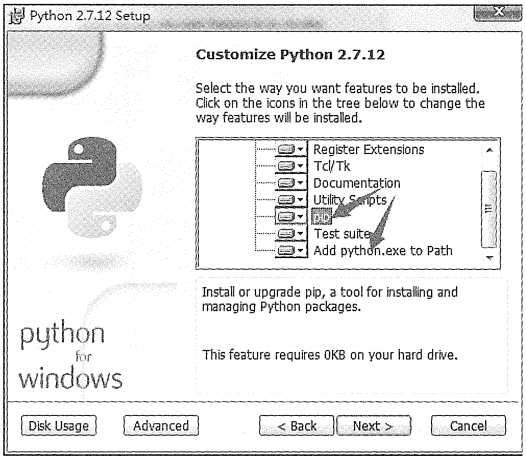


图 1-1　Python 安装界面

1.1.2　Ubuntu 上的 Python

本书采用 Ubuntu 16.04 版本，系统自带了 Python 2.7.11 的环境，如图 1-3 所示，所以不需要额外进行安装。

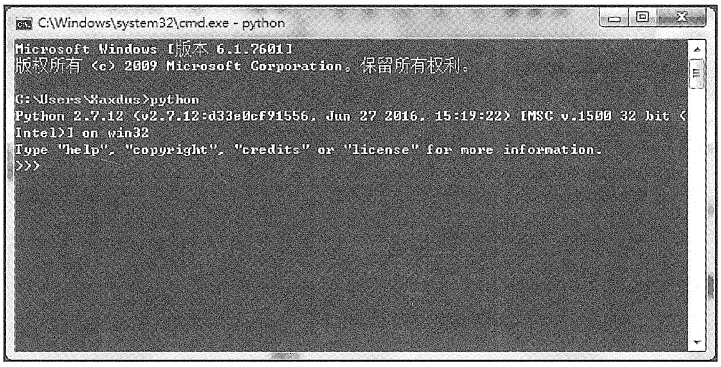


图 1-2　Python 命令行窗口

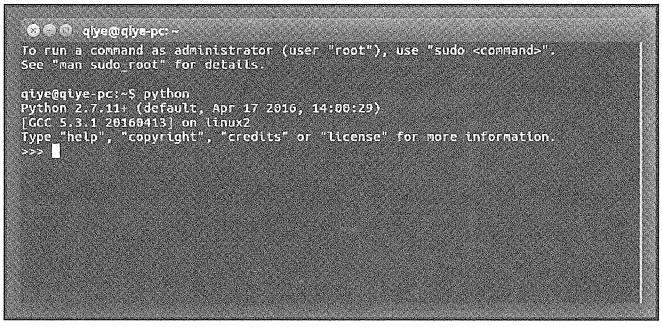


图 1-3　Python 环境

拥有了 Python 环境，但为了以后方便安装扩展模块，还需要安装 python-pip 和 python-dev，在 shell 中执行：sudo apt-get install python-pip python-dev 即可安装，如图 1-4 所示。