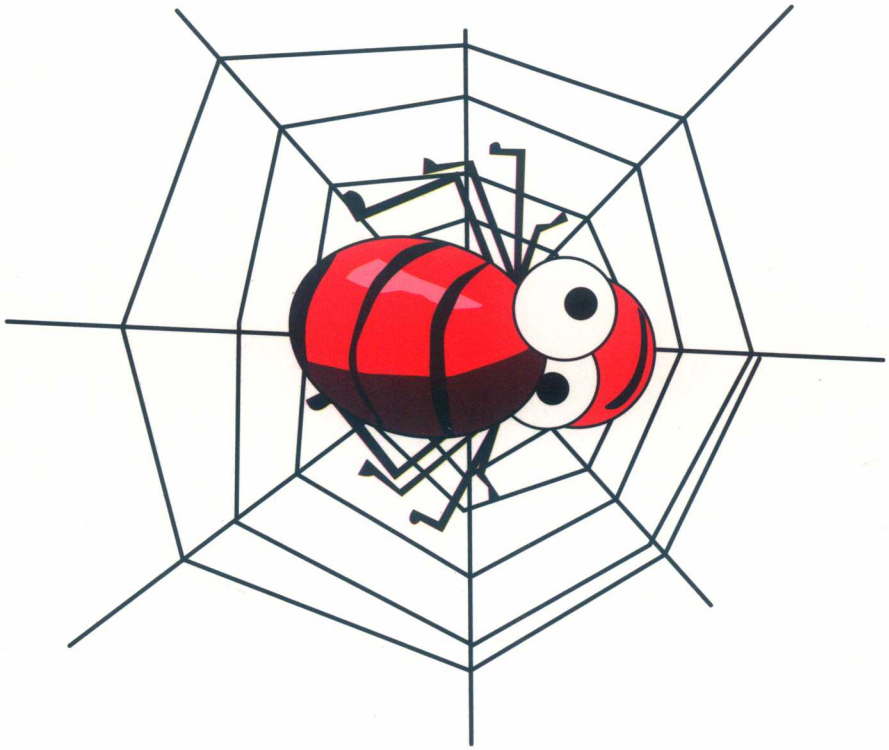
学会Scrapy爬虫框架，掌握网络数据采集技术



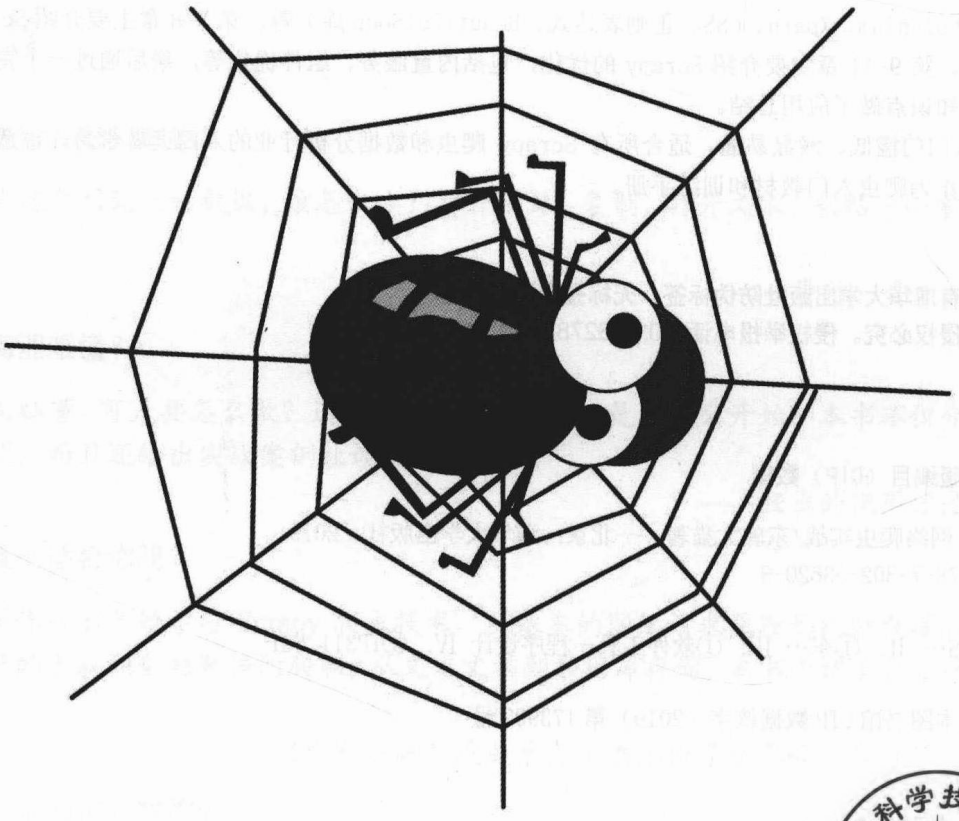
An open source and collaborative framework
for extracting the data you need from websites.

Scrapy网络爬虫实战

东郭大猫 著

示例源代码

清華大學出版社



Scrapy网络爬虫实战

东郭大猫 著

清華大學出版社
北京

内 容 简 介

随着大数据分析、大数据计算火热兴起，越来越多的企业发布了数据分析岗位，而数据分析的基础则是海量的数据。Python 中的 Scrapy 框架就是为了抓取数据而设计的。本书是一本 Scrapy 爬虫框架零基础起步的实战图书。

本书共分 11 章，第 1~2 章介绍 Python 环境的搭建、编辑器的使用、爬虫的一些基础知识（urllib、requests、Selenium、Xpath、CSS、正则表达式、BeautifulSoup 库）等。第 3~8 章主要介绍 Scrapy 框架的原理与使用。第 9~11 章主要介绍 Scrapy 的优化，包括内置服务、组件优化等，最后通过一个完整的大型示例对全书的知识点做了应用总结。

本书入门门槛低、浅显易懂，适合所有 Scrapy 爬虫和数据分析行业的入门读者学习，也适合高等院校和培训学校作为爬虫入门教材和训练手册。

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

Scrapy 网络爬虫实战/东郭大猫著. —北京：清华大学出版社，2019
ISBN 978-7-302-53620-8

Ⅰ. ①S… Ⅱ. ①东… 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 173907 号

责任编辑： 夏毓彦
封面设计： 王　翔
责任校对： 闫秀华
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

印 装 者： 北京嘉实印刷有限公司
经　　销： 全国新华书店
开　　本： 190mm×260mm　　**印　　张：** 15.75　　**字　　数：** 403 千字
版　　次： 2019 年 10 月第 1 版　　**印　　次：** 2019 年 10 月第 1 次印刷
定　　价： 59.00 元

产品编号：081510-01

前　言

读懂本书

还在复制粘贴找数据？

我想要这个网站上的数据，该怎么办？打开网站，复制，打开文本，粘贴……重复、重复、重复。

——费时、费力、错误多！

讲解晦涩难懂？

道理我都懂，可是要怎么做？这些数据我都想要，可是要怎么开始？本书不仅介绍 Scrapy 爬虫的原理，而且还给出实战案例让读者应用它们。

——爬虫的使用才是硬道理。

本书真的适合你吗？

本书帮你从零开始学习 Scrapy 爬虫技术，从基本的网络请求原理到抓取数据的保存，从单页面数据的下载到全站数据的爬取，从文本文档到数据库存储，本书介绍了实际使用中的各种基础知识。

——爬虫零基础？没关系，本书给出了从零开始学习的新手方案。

本书涉及的技术或框架

Python	HTTP	MySQL
Requests	JSON	MongoDB
BeautifulSoup	XPATH	Visual Studio
Selenium	CSS	Chrome 调试

本书涉及的示例和案例

抓取知乎热榜
名言网站抓取
博客园 Python 类文章抓取
深圳市社会保障局下载中心文件下载
链家数据保存至 MongoDB
豆瓣使用 Cookies 登录
抓取 cnBeta 科技类文章
伯乐在线订阅源数据抓取
伯乐在线最新文章抓取保存
起点小说网站小说封面抓取
豆瓣模拟提交表单登录
使用代理与统计链家小区信息
名言网站数据统计
IT 之家新闻抓取

猫眼电影 top100 抓取并发送邮件　　SegmentFault 全网用户信息抓取

本书特点

（1）本书不论是爬虫基础知识的介绍还是实例的开发，都是从实际应用的角度出发，精心选择典型的例子，讲解细致，分析透彻。

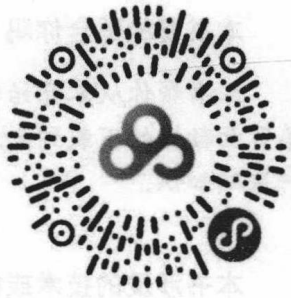
（2）深入浅出、轻松易学，以实例为主线，激发读者的学习兴趣，让读者能够快速学会 Scrapy 爬虫的实用技术。

（3）技术新颖、与时俱进，结合时下实用的技术，如 Requests、BeautifulSoup、Scrapy，使读者能够真正运用到实际工作中。

（4）贴近读者、贴近实际，大量成熟的第三方库和框架的使用和说明，帮助读者快速找到问题的最优解决方案，书中很多实例来自作者常用的数据源。

示例代码下载

本书示例代码请扫描二维码获得。如果下载有问题，请联系 booksaga@163.com，邮件主题为“Scrapy 网络爬虫实战”。



本书适用读者

Scrapy 网络爬虫初学者
从事 Web 网络数据分析的人员
从事数据存储的工作人员
高校与培训学校的师生

作　者
2019 年 5 月

试读结束：需要全本请在线购买：　www.ertongbook.com

目　录

试读结束：需要全本请在线购买： www.ertongbook.com

第1章

Python 开发环境的搭建

Scrapy 是使用 Python 编写的爬虫框架，在使用 Scrapy 之前，需要搭建开发环境。本章将为大家介绍 Python 的安装以及一些实用的编辑器的安装，以方便我们后续的开发工作。熟悉 Windows 系统的读者可以选择在 Windows 上搭建本书开发环境。

本章的主要知识点有：

- Python 安装
- PyCharm 编辑器安装
- Visual Studio 编辑器安装

1.1 Python SDK 安装

Python 是跨平台语言，可以运行在 Windows、Mac 及 Linux/UNIX 系统上，因此编写的代码在平台上没有运行的限制。目前，Python 有 Python 2 和 Python 3 两个版本，不幸的是两个版本很多地方不兼容。由于 Python 2 即将停止支持，而越来越多的库已经支持 Python 3，况且 Python 3 也提供了很多 Python 2 没有的新功能，因此本书使用 Python 3 搭建环境。

1.1.1 在 Windows 上安装 Python

首先从 Python 官网（https://www.python.org/downloads/）下载安装包，本书使用的是 3.6.3 版本，读者也可以下载更新的版本。下载后文件名为 python-3.6.3-amd64.exe，双击进行安装。

（1）在安装时先勾选 Add Python 3.6 to PATH 复选框，再选择 Customize installation 选项，如图 1.1 所示。

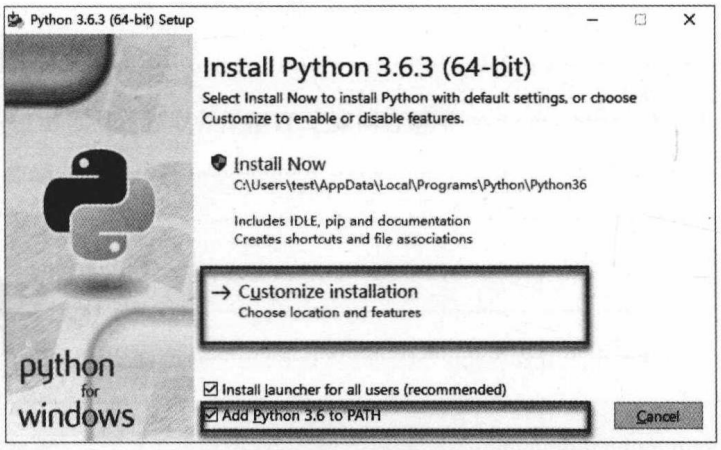


图 1.1　Python 3.6 安装首页

（2）务必选中 pip 复选框，单击 Next 按钮进行安装，如图 1.2 所示。

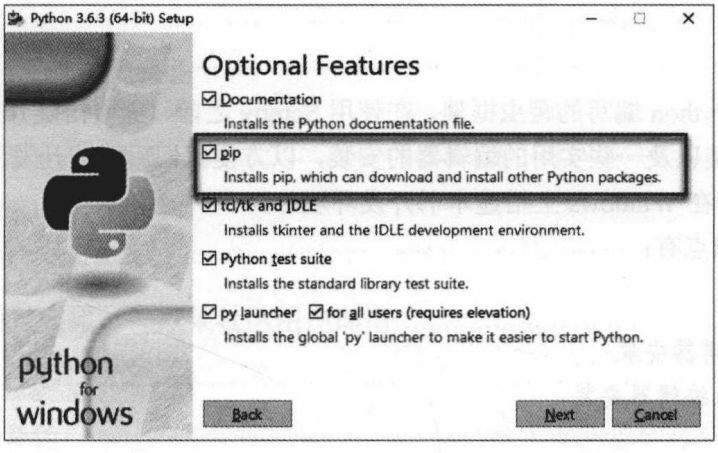


图 1.2　pip 选项

（3）选择安装路径，默认安装即可。打开命令提示窗口，输入 python，若出现 Python 版本号，则进入 Python 交互页面“>>>”，说明安装成功，如图 1.3 所示。

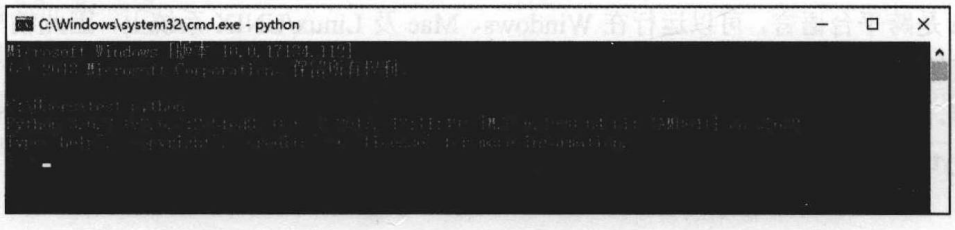


图 1.3　Python 3.6 安装检测

1.1.2　在 Ubuntu 上安装 Python

本书 Linux 发行版使用的是 Ubuntu 18.04，读者也可选择其他 Ubuntu 版本或者其他 Linux 发行版。Ubuntu 18.04 已经安装好了 Python 3，版本为 3.6.5。打开终端，输入 python3 命令，如图 1.4 所示。

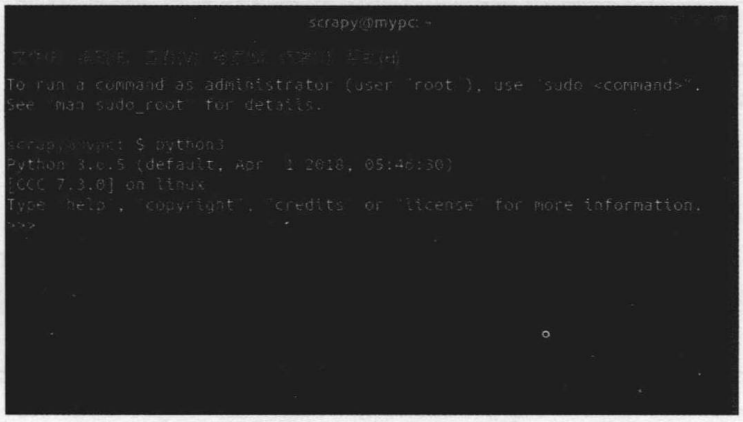


图 1.4　Python 3.6 安装检测

这里我们需要手动安装 pip，使用命令 sudo apt install python3-pip 进行安装，安装完成之后，运行 pip3，结果如图 1.5 所示。

图 1.5　pip 安装检测

1.2　安装开发工具 PyCharm 社区版

安装好 Python SDK 之后，我们需要一个方便的 IDE 来编写脚本。一个好的 IDE 能极大地提高工作效率，编者使用的是 PyCharm 这款编辑器。PyCharm 分为社区版和专业版，社区版为免费版本；专业版需付费，并且提供了更多的功能。针对爬虫开发来说，社区版已经足够使用了。

（1）进入 PyCharm 下载页面，以安装 Windows 版为例，下载社区版安装包，如图 1.6 所示。

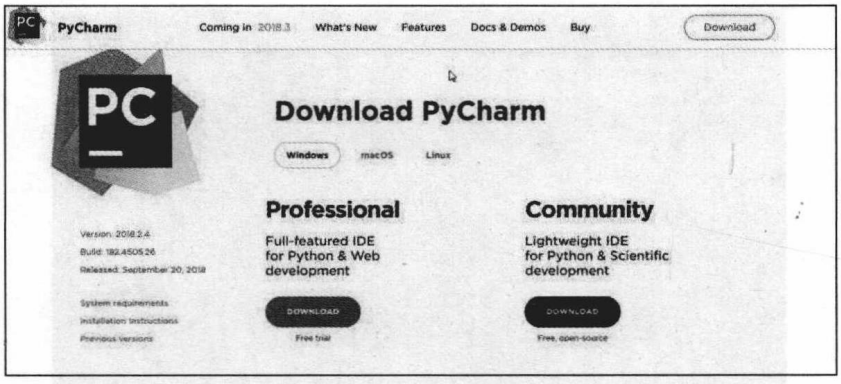


图 1.6　PyCharm 社区版下载

（2）下载安装包之后，双击进行安装，选择安装路径，关联 .py 文件，单击 Next 按钮安装，如图 1.7 所示。

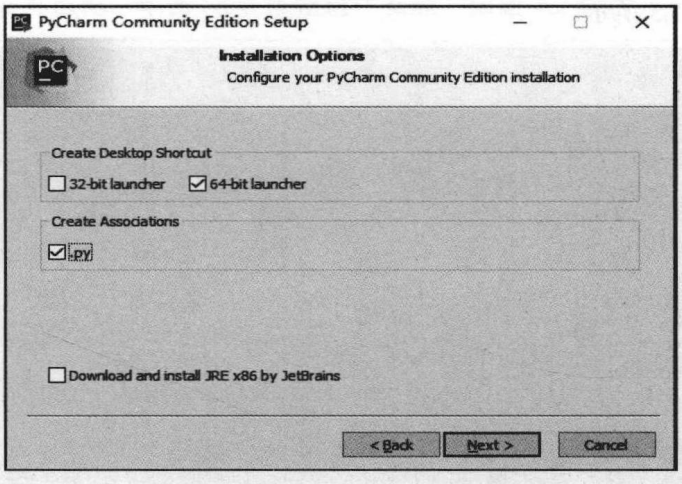


图 1.7　PyCharm 安装选项

（3）一直单击 Next 按钮，即可安装完成，做一些个性化的设置后即可开始创建项目，如图 1.8 所示。

图 1.8　PyCharm 创建项目

1.3　安装开发工具 Visual Studio 社区版

另一个编者使用起来也很方便的编辑器是 Visual Studio 社区版，同样免费，下载地址为 https://visualstudio.microsoft.com/zh-hans/，选择 Community 2017 进行下载。双击下载文件进行安装，由于是通过网络安装，因此需要下载一些安装文件。在安装时选择“Python 开发”复选框，之后开始安装直至完成，如图 1.9 所示。

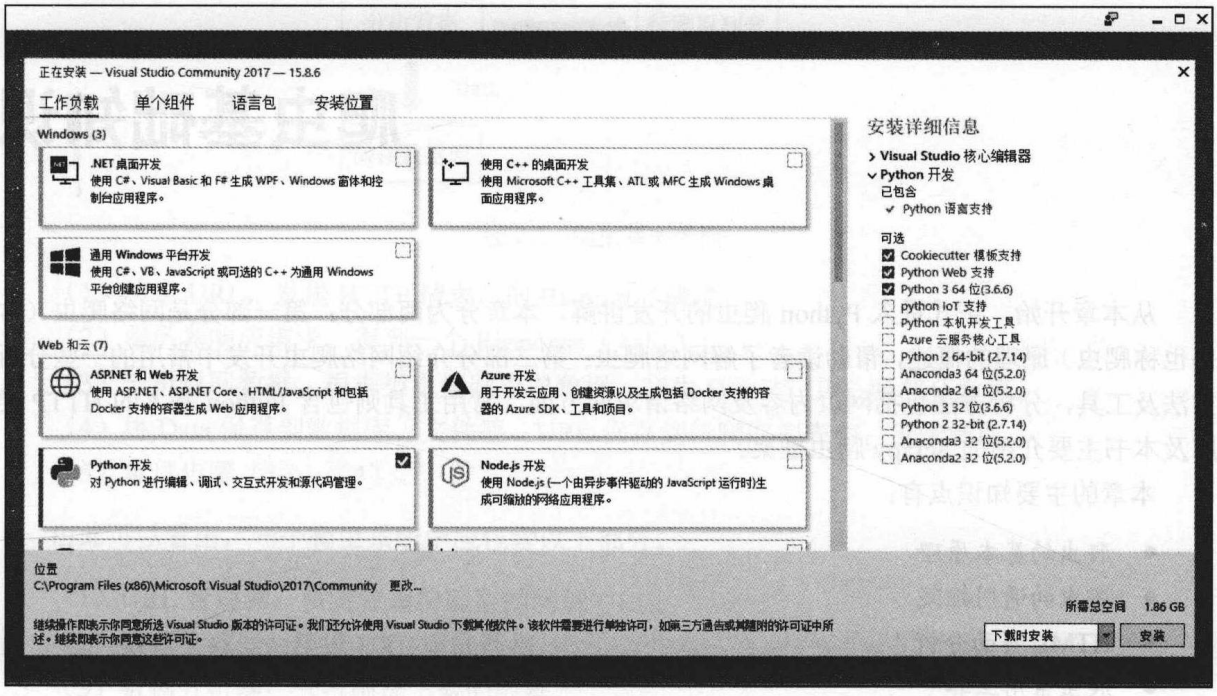


图 1.9　Visual Studio 社区版安装选项

第2章

爬虫基础知识

从本章开始，正式进入 Python 爬虫的开发讲解。本章分为两部分：第一部分是网络爬虫（本书也称爬虫）原理的概述，帮助读者了解网络爬虫；第二部分介绍网络爬虫开发中常用的一些分析方法及工具，分析方法包括网页内容及网络请求两方面，常用工具则包含 Python 基本的 HTTP 类库及本书主要介绍的 Scrapy 爬虫框架。

本章的主要知识点有：

- 爬虫的基本原理
- 爬虫的通用框架
- HTML 页面分析
- 爬虫常用工具

2.1 爬虫原理

网络爬虫在本质上就是模拟用户在浏览器上操作，发送请求，接收响应，然后分析并保存数据，只不过这个过程通过代码实现了大量的自动化操作。

2.1.1 爬虫运行基本流程

一般来说，一个爬虫的执行过程如图 2.1 所示。

试读结束：需要全本请在线购买： www.ertongbook.com