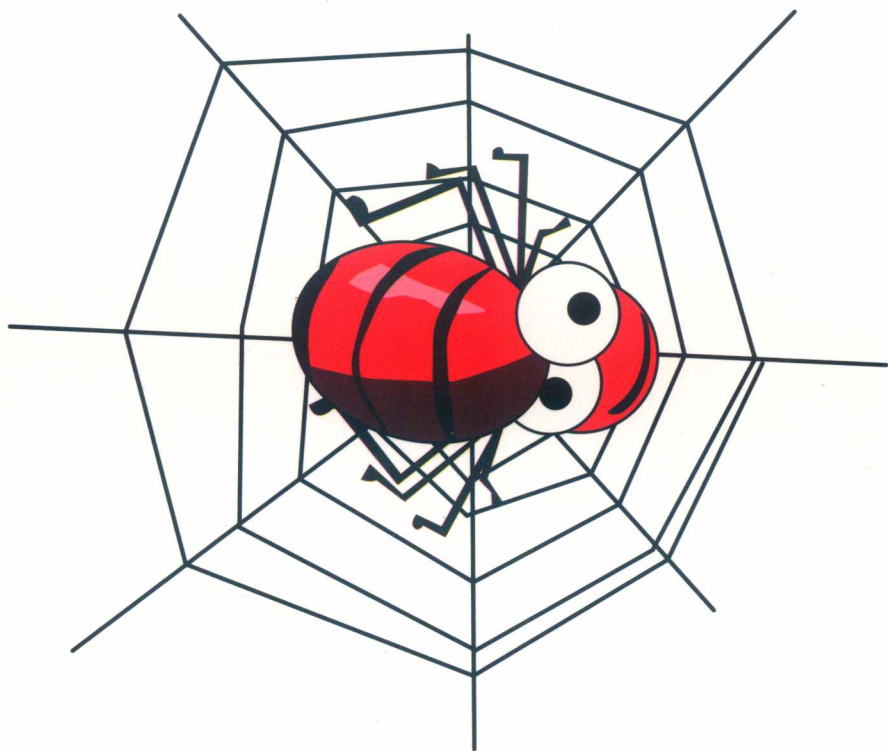


学会Scrapy爬虫框架，掌握网络数据采集技术



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

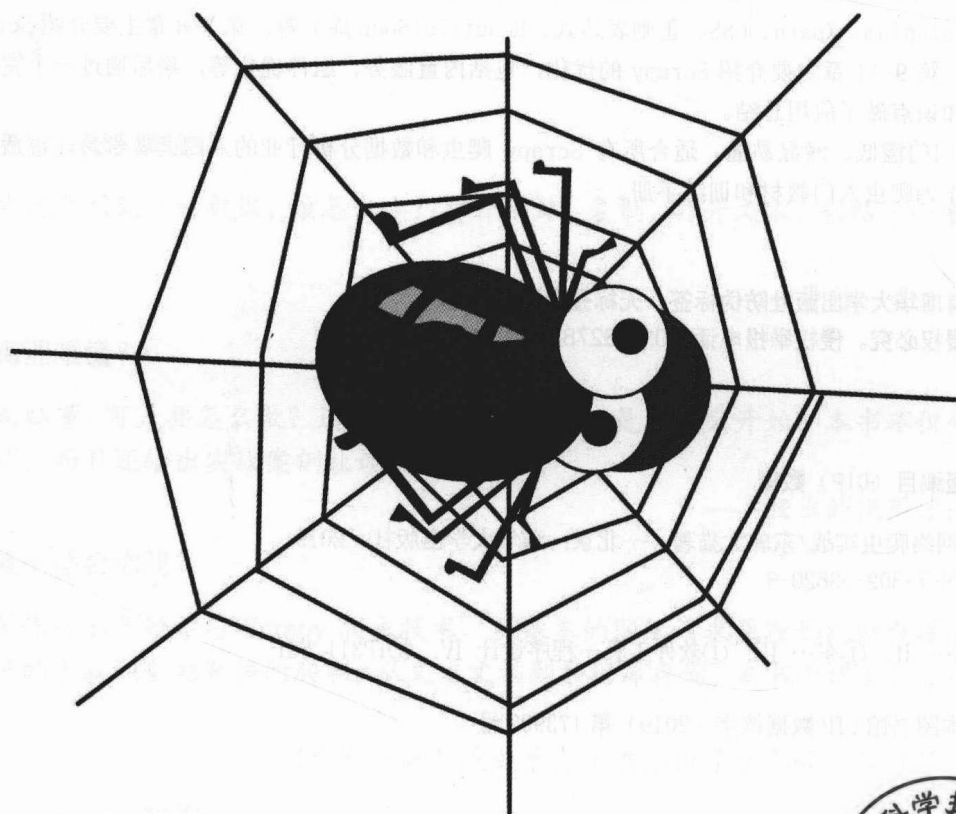
Scrapy网络爬虫实战

东郭大猫 著

 示例源代码



清华大学出版社



Scrapy网络爬虫实战



东郭大猫 著

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

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。

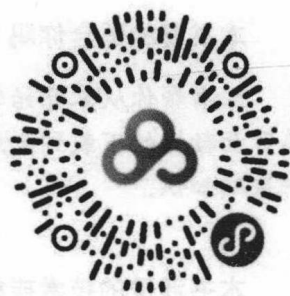
(2) 深入浅出、轻松易学，以实例为主线，激发读者的学习兴趣，让读者能够快速学会 Scrapy 爬虫的实用技术。

(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

目 录

第 1 章 Python 开发环境的搭建	1
1.1 Python SDK 安装	1
1.1.1 在 Windows 上安装 Python	1
1.1.2 在 Ubuntu 上安装 Python	2
1.2 安装开发工具 PyCharm 社区版	3
1.3 安装开发工具 Visual Studio 社区版	5
第 2 章 爬虫基础知识	6
2.1 爬虫原理	6
2.1.1 爬虫运行基本流程	6
2.1.2 HTTP 请求过程	8
2.2 网页分析方法 1: 浏览器开发人员工具	9
2.2.1 Elements 面板	10
2.2.2 Network 面板	11
2.3 网页分析方法 2: XPath 语法	14
2.3.1 XPath 节点	14
2.3.2 XPath 语法	15
2.3.3 XPath 轴	17
2.3.4 XPath 运算符	19
2.4 网页分析方法 3: CSS 选择语法	19
2.4.1 元素选择器	20
2.4.2 类选择器	21
2.4.3 ID 选择器	21
2.4.4 属性选择器	21
2.4.5 后代选择器	21
2.4.6 子元素选择器	22
2.4.7 相邻兄弟选择器	22
2.5 网页分析方法 4: 正则表达式	22
2.5.1 提取指定字符	23
2.5.2 预定义字符集	23
2.5.3 数量限定	23

2.5.4	分支匹配	24
2.5.5	分组	24
2.5.6	零宽断言	24
2.5.7	贪婪模式与非贪婪模式	25
2.5.8	Python 中的正则表达式	25
2.6	爬虫常用类库 1: Python 中的 HTTP 基本库 urllib	30
2.6.1	发送请求	30
2.6.2	使用 Cookie	31
2.7	爬虫常用类库 2: 更人性化的第三方库 requests	33
2.7.1	发送请求	34
2.7.2	请求头	35
2.7.3	响应内容	35
2.7.4	响应状态码	36
2.7.5	cookies 参数	37
2.7.6	重定向与请求历史	37
2.7.7	超时	38
2.7.8	设置代理	38
2.7.9	会话对象	38
2.8	爬虫常用类库 3: 元素提取利器 BeautifulSoup	39
2.8.1	安装 BeautifulSoup	39
2.8.2	安装解析器	40
2.8.3	BeautifulSoup 使用方法	41
2.8.4	BeautifulSoup 对象	43
2.8.5	遍历文档树	47
2.8.6	搜索文档树	52
2.8.7	BeautifulSoup 中的 CSS 选择器	57
2.9	爬虫常用类库 4: Selenium 操纵浏览器	58
2.9.1	安装 Selenium	59
2.9.2	Selenium 的基本使用方法	59
2.9.3	Selenium Webdriver 的原理	61
2.9.4	Selenium 中的元素定位方法	61
2.9.5	Selenium Webdriver 基本操作	63
2.9.6	Selenium 实战: 抓取拉钩网招聘信息	64
2.10	爬虫常用类库 5: Scrapy 爬虫框架	67
2.10.1	安装 Scrapy	67
2.10.2	Scrapy 简介	68

2.11 基本爬虫实战：抓取 cnBeta 网站科技类文章	69
2.11.1 URL 管理器	70
2.11.2 数据下载器	71
2.11.3 数据分析器	72
2.11.4 数据保存器	74
2.11.5 调度器	75
第 3 章 Scrapy 命令行与 Shell	78
3.1 Scrapy 命令行介绍	78
3.1.1 使用 startproject 创建项目	80
3.1.2 使用 genspider 创建爬虫	81
3.1.3 使用 crawl 启动爬虫	82
3.1.4 使用 list 查看爬虫	82
3.1.5 使用 fetch 获取数据	83
3.1.6 使用 runspider 运行爬虫	84
3.1.7 通过 view 使用浏览器打开 URL	85
3.1.8 使用 parse 测试爬虫	85
3.2 Scrapy Shell 命令行	85
3.2.1 Scrapy Shell 的用法	85
3.2.2 实战：解析名人名言网站	86
第 4 章 Scrapy 爬虫	89
4.1 编写爬虫	89
4.1.1 scrapy.Spider 爬虫基本类	89
4.1.2 start_requests()方法	90
4.1.3 parse(response)方法	91
4.1.4 Selector 选择器	91
4.2 通用爬虫	94
4.2.1 CrawlSpider	94
4.2.2 XMLFeedSpider	95
4.2.3 CSVFeedSpider	96
4.2.4 SitemapSpider	97
4.3 爬虫实战	98
4.3.1 实战 1：CrawlSpider 爬取名人名言	98
4.3.2 实战 2：XMLFeedSpider 爬取伯乐在线的 RSS	102
4.3.3 实战 3：CSVFeedSpider 提取 csv 文件数据	104
4.3.4 实战 4：SitemapSpider 爬取博客园文章	106

第 5 章 Scrapy 管道	109
5.1 管道简介	109
5.2 编写自定义管道	110
5.3 下载文件和图片	113
5.3.1 文件管道	114
5.3.2 图片管道	117
5.4 数据库存储 MySQL	121
5.4.1 在 Ubuntu 上安装 MySQL	121
5.4.2 在 Windows 上安装 MySQL	122
5.4.3 MySQL 基础	125
5.4.4 MySQL 基本操作	127
5.4.5 Python 操作 MySQL	129
5.5 数据库存储 MongoDB	131
5.5.1 在 Ubuntu 上安装 MongoDB	132
5.5.2 在 Windows 上安装 MongoDB	132
5.5.3 MongoDB 基础	135
5.5.4 MongoDB 基本操作	137
5.5.5 Python 操作 MongoDB	143
5.6 实战：爬取链家二手房信息并保存到数据库	144
第 6 章 Request 与 Response	157
6.1 Request 对象	157
6.1.1 Request 类详解	158
6.1.2 Request 回调函数与错误处理	160
6.2 Response	162
6.2.1 Response 类详解	162
6.2.2 Response 子类	163
第 7 章 Scrapy 中间件	165
7.1 编写自定义 Spider 中间件	165
7.1.1 激活中间件	165
7.1.2 编写 Spider 中间件	166
7.2 Spider 内置中间件	168
7.2.1 DepthMiddleware 爬取深度中间件	168
7.2.2 HttpErrorMiddleware 失败请求处理中间件	168
7.2.3 OffsiteMiddleware 过滤请求中间件	169
7.2.4 RefererMiddleware 参考位置中间件	169

7.2.5	UrlLengthMiddleware 网址长度限制中间件	170
7.3	编写自定义下载器中间件	170
7.3.1	激活中间件	170
7.3.2	编写下载器中间件	171
7.4	下载器内置中间件	173
7.4.1	CookiesMiddleware	173
7.4.2	HttpProxyMiddleware	174
7.5	实战：为爬虫添加中间件	174
第 8 章	Scrapy 配置与内置服务	178
8.1	Scrapy 配置简介	178
8.1.1	命令行选项（优先级最高）	178
8.1.2	每个爬虫内配置	179
8.1.3	项目设置模块	179
8.1.4	默认的命令配置	181
8.1.5	默认全局配置（优先级最低）	182
8.2	日志	182
8.3	数据收集	184
8.4	发送邮件	187
8.4.1	简单例子	187
8.4.2	MailSender 类	187
8.4.3	在 settings.py 中对 Mail 进行设置	188
8.5	实战：抓取猫眼电影 TOP100 榜单数据	188
8.5.1	分析页面元素	189
8.5.2	创建项目	189
8.5.3	编写 items.py	190
8.5.4	编写管道 pipelines.py	190
8.5.5	编写爬虫文件 top100.py	191
第 9 章	模拟登录	194
9.1	模拟提交表单	194
9.2	用 Cookie 模拟登录状态	197
9.3	项目实战	198
9.3.1	实战 1：使用 FormRequest 模拟登录豆瓣	198
9.3.2	实战 2：使用 Cookie 登录	202

第 10 章 Scrapy 爬虫优化	205
10.1 Scrapy+MongoDB 实战：抓取并保存 IT 之家博客新闻	205
10.1.1 确定目标	205
10.1.2 创建项目	206
10.1.3 编写 items.py 文件	207
10.1.4 编写爬虫文件 news.py	207
10.1.5 编写管道 pipelines.py	209
10.1.6 编写 settings.py	210
10.1.7 运行爬虫	211
10.2 用 Benchmark 进行本地环境评估	212
10.3 扩展爬虫	214
10.3.1 增大并发	214
10.3.2 关闭 Cookie	214
10.3.3 关闭重试	214
10.3.4 减少下载超时时间	215
10.3.5 关闭重定向	215
10.3.6 AutoThrottle 扩展	215
第 11 章 Scrapy 项目实战：爬取某社区用户详情	217
11.1 项目分析	217
11.1.1 页面分析	217
11.1.2 抓取流程	221
11.2 创建爬虫	221
11.2.1 cookies 收集器	222
11.2.2 Items 类	225
11.2.3 Pipeline 管道编写	226
11.2.4 Spider 爬虫文件	227
11.2.5 Middlewares 中间件编写	235

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。

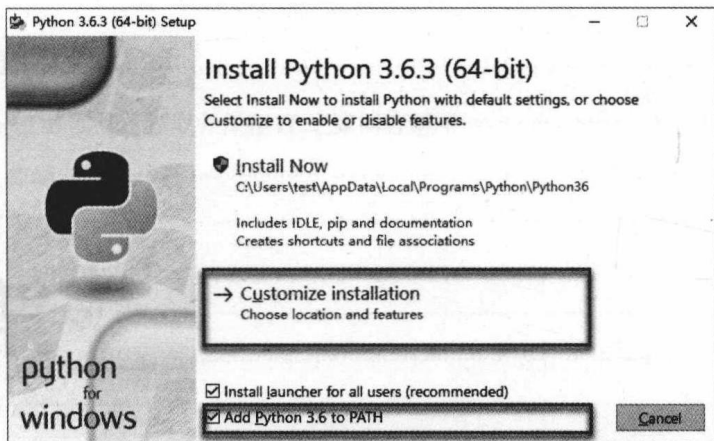


图 1.1 Python 3.6 安装首页

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

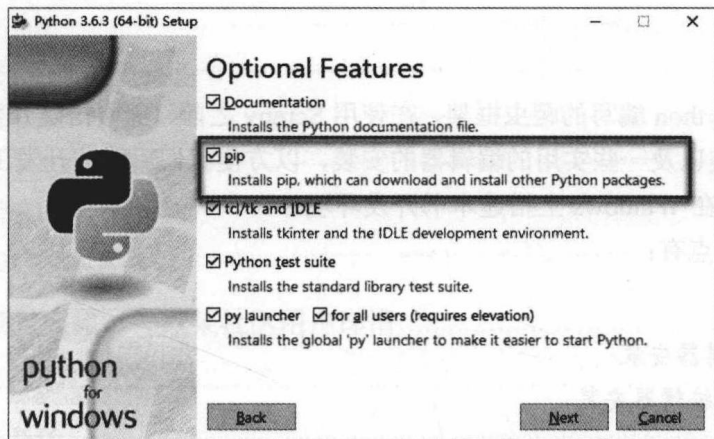


图 1.2 pip 选项

(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。

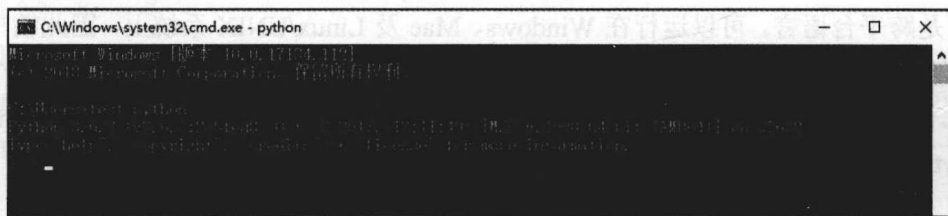


图 1.3 Python 3.6 安装检测

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。

```

scrapy@mypc: ~
Python 3.6.5 (default, Apr 1 2018, 05:46:30)
[GCC 7.3.0] on linux
Type 'help', 'copyright', 'credits' or 'license()' for more information.
>>>

```

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。

```

scrapy@mypc: ~
Python 3.6.5 (default, Apr 1 2018, 05:46:30)
[GCC 7.3.0] on linux
Type 'help', 'copyright', 'credits' or 'license()' for more information.
>>>

```

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。

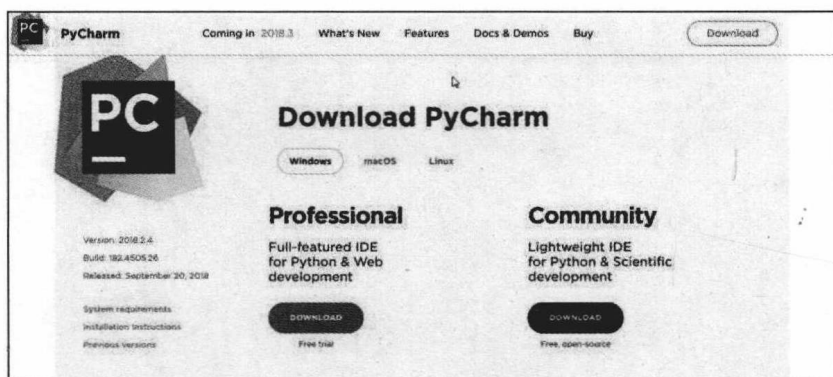


图 1.6 PyCharm 社区版下载

(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。

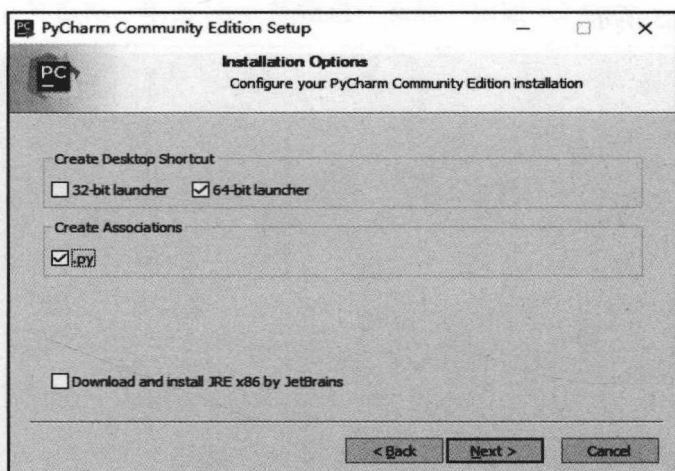


图 1.7 PyCharm 安装选项

(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。

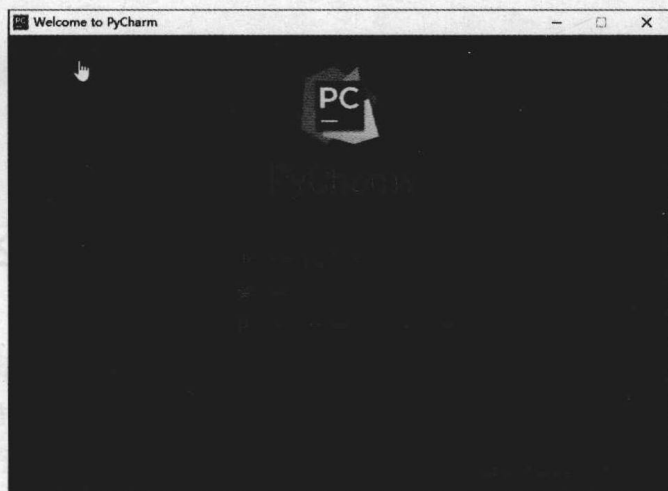


图 1.8 PyCharm 创建项目

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。

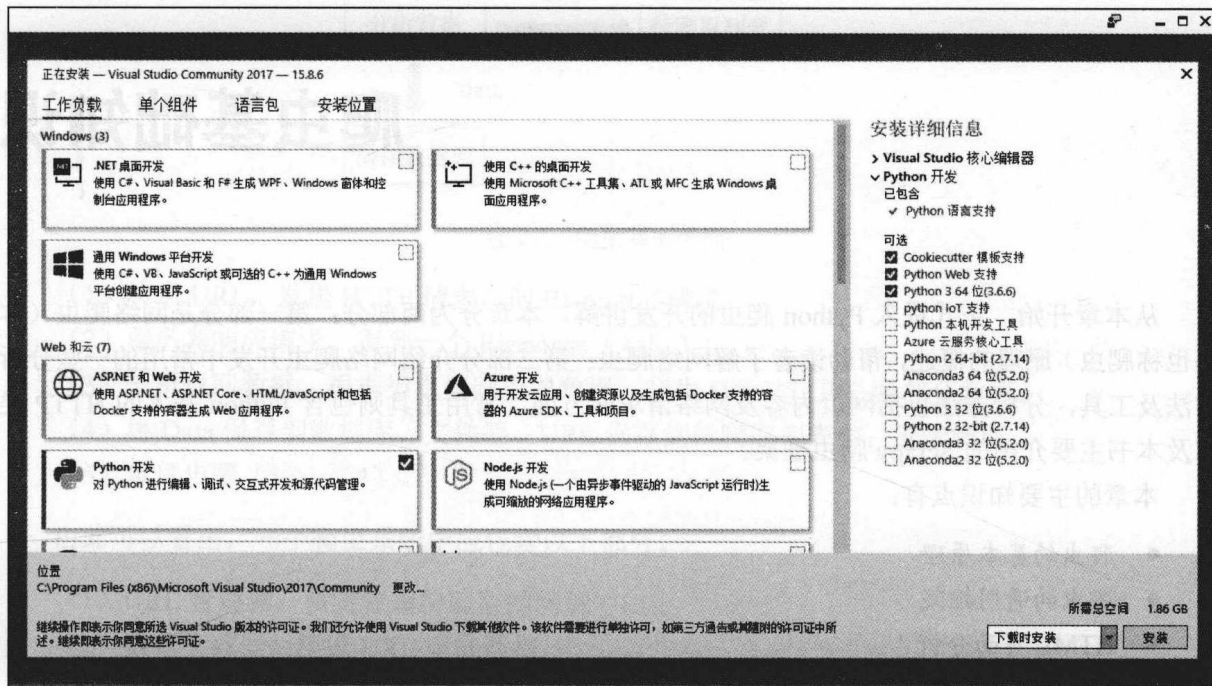


图 1.9 Visual Studio 社区版安装选项

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。