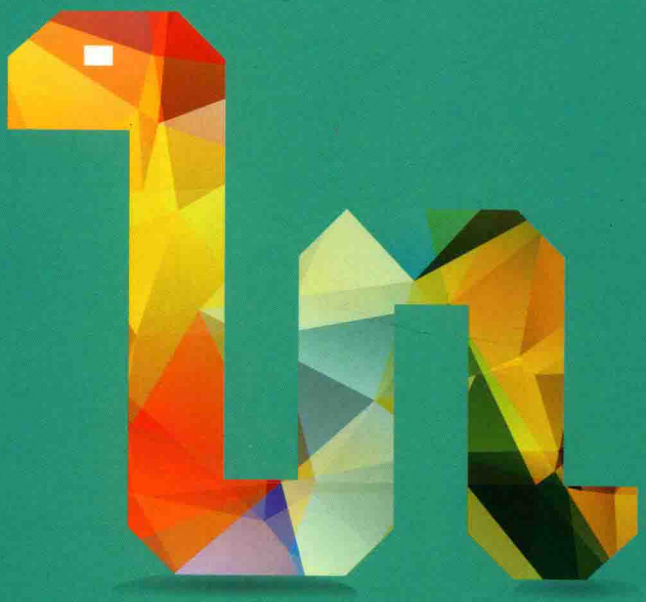
HZ BOOKS
华章科技

作者多年从事网络爬虫领域的教学及研究工作，有着丰富的实践经验。

面向初学者全面介绍 Python 网络爬虫的实战宝典，涵盖网络爬虫的核心概念、算法和技术实现，内容系统，案例丰富。



Python 网络爬虫技术与实战

赵国生 王健 ◎ 编著

WEB CRAWLER TECHNIQUE AND ACTUAL APPLICATION BASED ON PYTHON

机械工业出版社
China Machine Press

Python网络爬虫技术与实战

赵国生 王健 ◎ 编著

WEB CRAWLER TECHNIQUE
AND ACTUAL APPLICATION
BASED ON PYTHON

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

Python 网络爬虫技术与实战 / 赵国生，王健编著 . —北京：机械工业出版社，2021.1

ISBN 978-7-111-67411-5

I. P… II. ①赵… ②王… III. 软件工具 – 程序设计 IV. TP311.56

中国版本图书馆 CIP 数据核字（2021）第 009282 号

Python 网络爬虫技术与实战

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：栾传龙　　责任校对：殷　虹

印　　刷：北京市荣盛彩色印刷有限公司　　版　　次：2021 年 1 月第 1 版第 1 次印刷

开　　本：186mm × 240mm　1/16　　印　　张：29.5

书　　号：ISBN 978-7-111-67411-5　　定　　价：89.00 元

客服电话：（010）88361066　88379833　68326294　　投稿热线：（010）88379604

华章网站：www.hzbook.com　　读者信箱：hzit@hzbook.com

Preface 前　言

为什么写作本书

大数据时代已经到来，网络爬虫技术已成为这个时代不可或缺的一项技术，企业需要数据来分析用户行为、产品的不足之处以及竞争对手的信息等，而这一切的首要条件就是数据的采集。在互联网社会中，数据是无价之宝，一切皆为数据，谁拥有了大量有用的数据，谁就拥有了决策的主动权。如何有效地采集并利用这些信息成了一个巨大的挑战，而网络爬虫是自动采集数据的有效手段。网络爬虫是一种按照一定的规则，自动抓取互联网海量信息的程序或脚本。网络爬虫的应用领域很广泛，如搜索引擎、数据采集、广告过滤、大数据分析等。

笔者多年来一直从事网络爬虫相关课程的讲授及科学研究工作，有着丰富的教学和实践经验。在内容编排上，本书采用梯度层次化结构，由浅入深地介绍爬虫的知识点、原理及应用，并结合大量实例讲解操作步骤，使读者能够快速地理解网络爬虫的核心技术。

内容介绍

全书共 14 章，具体内容如下：

第 1 章主要介绍 Python 的安装、配置和基础语法，以及 Python 的字符串、数据结构、控制语句和函数等；

第 2 章主要介绍爬虫的类型、爬虫的抓取策略以及深入学习爬虫所需的网络基础等相关知识；

第 3 章主要对爬虫技术中经常使用到的 urllib、request、lxml 和 Beautiful Soup 库等进行详细介绍，最后展示了 4 个利用 Python 爬取数据的实例；

第 4 章主要对 Python 中正则表达式的语法、匹配规则和 re 模块常用函数进行详细阐述，并给出了实例；

第 5 章主要对 3 种主流库（PIL 库、Tesseract 库和 TensorFlow 库）的语法、类型、识别方法和案例进行介绍；

第 6 章详细介绍 Fiddler 的安装与配置、捕获会话、QuickExec 命令行的使用和 Fiddler 的

断点功能等；

第 7 章主要介绍数据存储在文件中和存储在数据库中这两种存储方式；

第 8 章重点介绍 Scrapy 框架的 Selector 用法，以及 Beautiful Soup 库和 CrawlSpider 的使用，然后介绍了 Scrapy Shell 和 Scrapyrt 的使用；

第 9 章主要介绍多线程和 Threading 模块的基本概念；

第 10 章主要介绍如何对动态网页进行信息爬取，首先介绍了浏览器开发工具的使用，然后介绍了异步加载技术、AJAX 技术和 Selenium 模拟浏览器；

第 11 章主要介绍分布式爬虫的原理及实现过程，然后介绍了 Scrapy-redis 分布式组件的工作机制和安装配置；

第 12 章主要介绍如何利用 Selenium 抓取并用 pyquery 解析电商网站的商品信息，然后将其保存到 MongoDB；

第 13 章主要介绍静态网页和动态网页的爬取方法，并对请求 - 响应关系进行了介绍，然后介绍了请求头和请求体；

第 14 章主要讲解如何通过 urllib 模块和 Scrapy 框架实现图片爬虫项目，以及利用 TensorFlow、KNN 和 CNN 等机器学习框架进行训练的方法与过程。

主要特点

本书针对网络爬虫学习的特点，结合作者多年使用网络爬虫的教学和实践经验，由浅入深、从简到繁、图文并茂地介绍了 Python 基础语法、爬虫原理、爬虫常用库模块、正则表达式、验证码识别、抓包工具 Fiddler、数据存储、Scrapy 爬虫框架、多线程爬虫、动态网页爬虫和分布式爬虫等方面的内容。本书内容条理清晰、针对性强，语言通俗易懂，在讲解的过程中配合大量的实例操作，符合读者的学习习惯。每章都是从基础知识开始介绍，然后是实例分析，最后附以练习题巩固学习效果，将理论与实践紧密结合。

具体来讲，本书具有以下鲜明的特点：

- 内容系统，由浅入深；
- 案例讲解，通俗易懂；
- 综合实战，注重实践。

读者对象

本书适合网络爬虫初学者，以及具有一定网络爬虫基础，但希望更深入了解、掌握爬虫原理与应用的中级读者阅读。

本书可以作为本科或者大专院校网络安全、电子信息、数据科学、网络工程等相关专业的教材，也可作为从事网络爬虫相关工作的科研或者工程技术人员的参考书。

致谢

本书由哈尔滨师范大学的赵国生和哈尔滨理工大学的王健编写。其中，赵国生主要负责第1～11章的编写，王健负责第12～14章的编写。参与本书大量辅助性工作的研究生有邹伊凡、刘冬梅、张婧婷、廖玉婷、晁绵星、谢宝文等，在此表示感谢。

特别感谢以下项目对本书的支持：国家自然科学基金项目“可生存系统的自主认知模式研究”（61202458）、国家自然科学基金项目“基于认知循环的任务关键系统可生存性自主增长模型与方法”（61403109）、高等学校博士点基金项目（20112303120007）、哈尔滨市科技创新人才研究专项（2016RAQXJ036）和黑龙江省自然科学基金（F2017021）。

感谢您选择本书，虽然笔者在编写过程中力求叙述准确、完善，但由于水平有限，书中仍可能存在欠妥之处，希望您可以把对本书的意见和建议告诉我们。

最后，再次希望本书能够对您的工作和学习有所帮助！

目　录 *Contents*

第 1 章 Chapter 1

Python 环境搭建及基础学习

Python 是一种跨平台的计算机语言，也是一种解释型的、面向对象和动态数据类型的高级程序设计语言。Python 由吉多·范罗苏姆在 1989 年发明，第一个公开发行的版本出现于 1991 年。近年来，由于 Python 在人工智能和云计算领域的应用，其热度越来越高。自从 Facebook 开源了 PyTorch，Python 在 AI 时代登上“第一语言”的位置已成定局。Google 的 TensorFlow 大部分代码都是由 Python 语言编写的，目前最流行的云计算框架 OpenStack 也是基于 Python 开发的。Python 热度很高的另一个主要原因是 Python 拥有强大到无法想象的标准库和第三方库，无论想进行哪个方向的编程，几乎都能找到相应库的支持。本章将首先介绍 Python 的安装、配置和基础语法，然后介绍 Python 的字符串、数据结构、控制语句和函数等，最后介绍文件读写操作和 Python 面向对象的知识。

1.1 Python 3.6 的安装与配置

Python 官方同时发行和维护着 Python 2.x 和 Python 3.x 两个不同系列的版本，这两个系列的版本之间的很多用法是不兼容的，除了基本的输入、输出方式有所不同，很多内置函数和标准库的用法也有非常大的区别。Python 3.x 的设计理念更加合理、高效和人性化，代码开发和运行效率更高，2015 年底就已经出现 Python 3.x 全面普及和应用的趋势。

本节将分别介绍 Windows、Linux 和 macOS 下的 Python 3.6.x 的安装方法。

1.1.1 Windows 下的安装

1）在网站 https://www.python.org/downloads/windows/ 中找到需要的 Python 版本，本书选择的版本是 3.6.5，版本信息如图 1-1 所示。

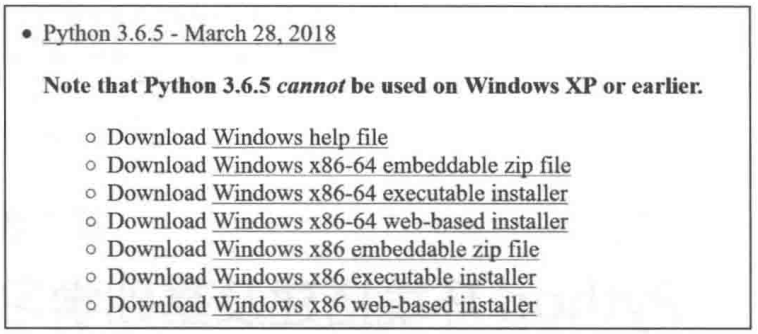


图 1-1 Python 3.6.5 版本信息

2）运行下载的文件，出现初始安装界面如图 1-2 所示，在安装界面中选择 Customize installation 选项。

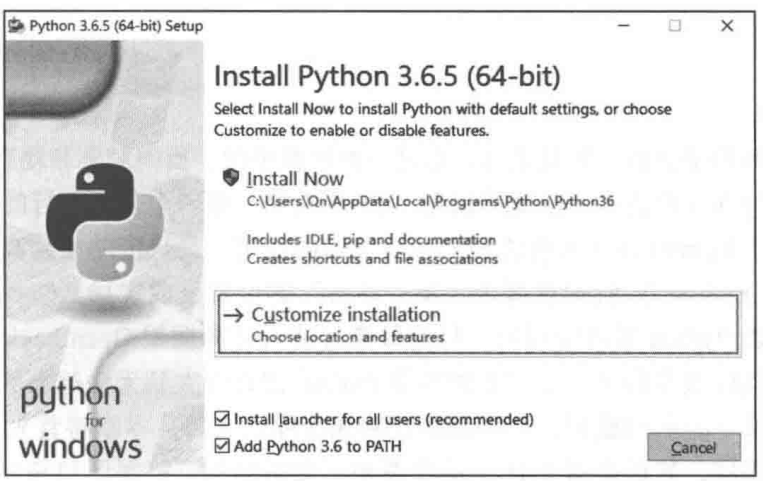


图 1-2 Python 安装界面

具体安装过程如图 1-3、图 1-4 和图 1-5 所示。

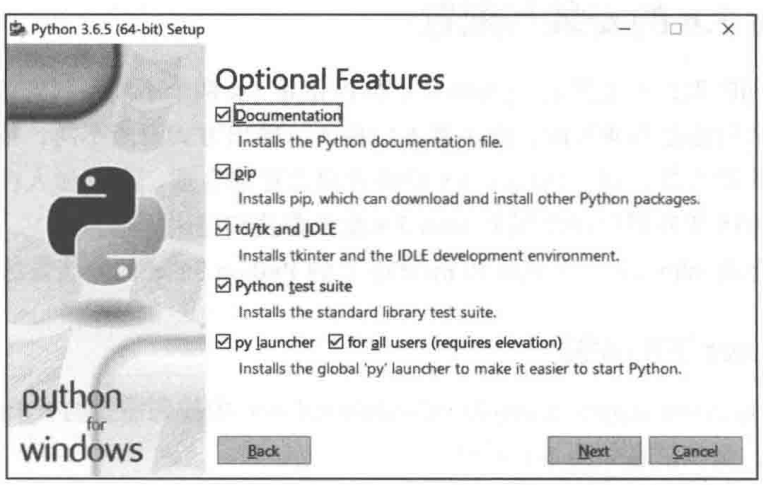


图 1-3 单击“Next”按钮

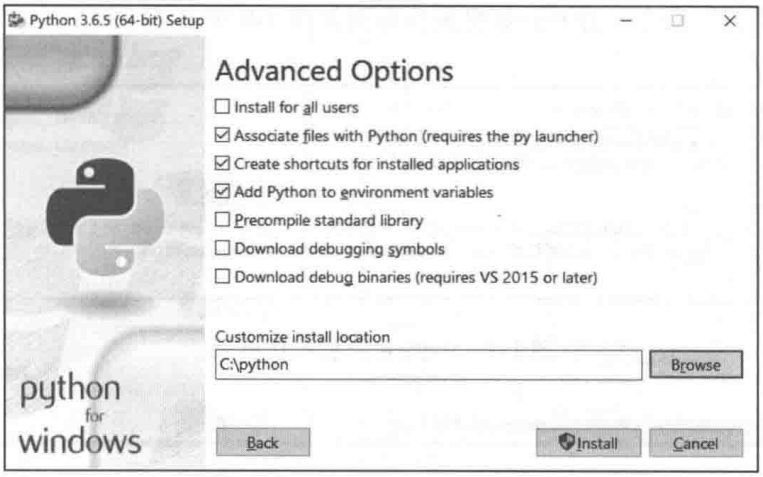


图 1-4　单击“Install”按钮

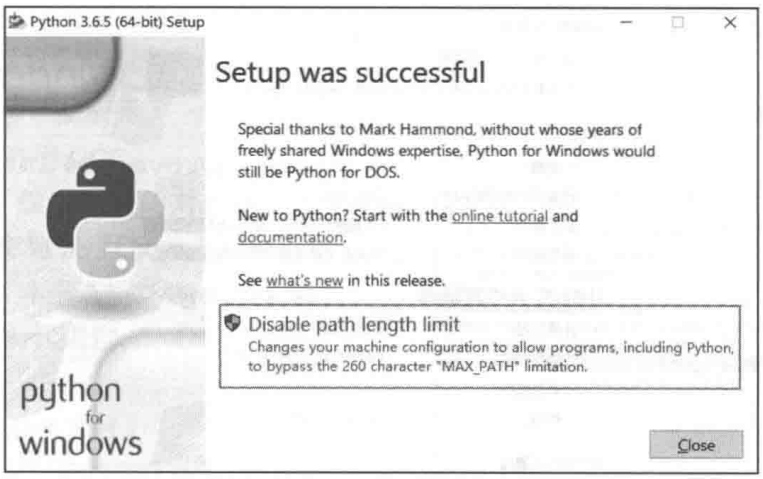


图 1-5　Python 安装完成

3）运行软件。单击“开始”按钮，在弹出的输入框中输入“cmd”后按 Enter 键，如图 1-6 所示。

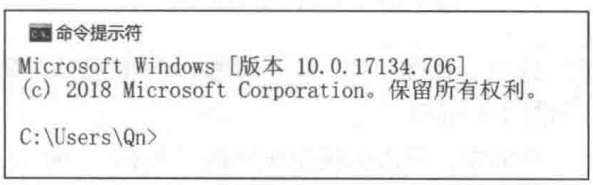


图 1-6　cmd 界面

在 cmd 命令界面中输入“python”后按 Enter 键，成功安装界面如图 1-7 所示。

4）若没有在初始安装界面（图 1-2）中选择“Add Python 3.6 to Path”，那么程序在 cmd 命令框中就无法正常运行。此时可以手动配置环境变量。鼠标右键单击桌面“计算机”图标，

在菜单中选择“属性”命令，打开计算机属性界面如图 1-8 所示。

图 1-7 Python 安装成功界面

图 1-8 计算机属性界面

单击“高级系统设置”选项，在弹出的“系统属性”界面选择“高级”选项卡，单击右下角的“环境变量”按钮，如图 1-9 所示。

在弹出的“环境变量”界面中，双击系统变量中的“Path”，如图 1-10 所示。

在弹出的“编辑环境变量”界面，单击“新建”按钮，在输入框中输入 Python 的安装路径（如 C:\Python\），如图 1-11 所示。单击“确定”按钮以保存添加的环境变量，之后重复第 3 步验证即可。至此，在 Windows 上安装 Python 的操作已完成。

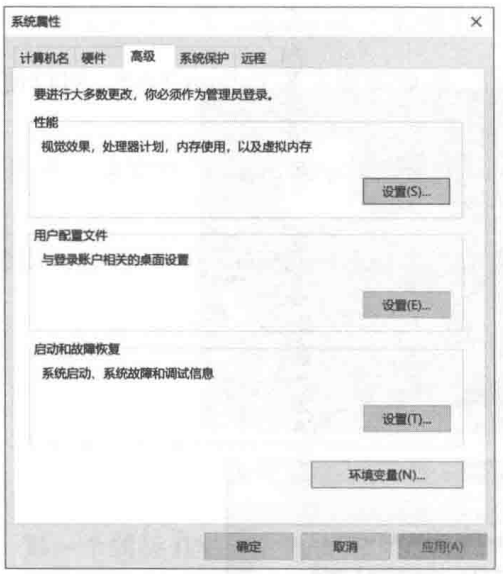


图 1-9　系统属性界面

图 1-10　环境变量界面

1.1.2　Linux 下的安装

1）准备编译环境 GCC。

2）去官网下载要安装的对应版本的 Python 的源代码。下载地址为 https://www.python.org/downloads/source/，本书选择 Python 3.6.5 版本。

3）解压下载的代码包。

4）配置。

① 查找 configure 文件。

```
find . -name configure
cd /usr/local/python-3.6.5/
```

② 进行配置。

```
./configure
```

5）编译。

```
make
make install
```

图 1-11　编辑环境变量界面

6）替换以前的 Python 默认版本（创建新的软链接）。

```
cd /usr/bin/
rm -rf python
ln -s /usr/local/Python-3.6.5/bin/python ./python
```

至此，在 Linux 下安装 Python 的操作已完成。