TURING 图灵原创

Python 3 网络爬虫开发实战

崔庆才 著

中国工信出版集团
人民邮电出版社
POSTS & TELECOM PRESS

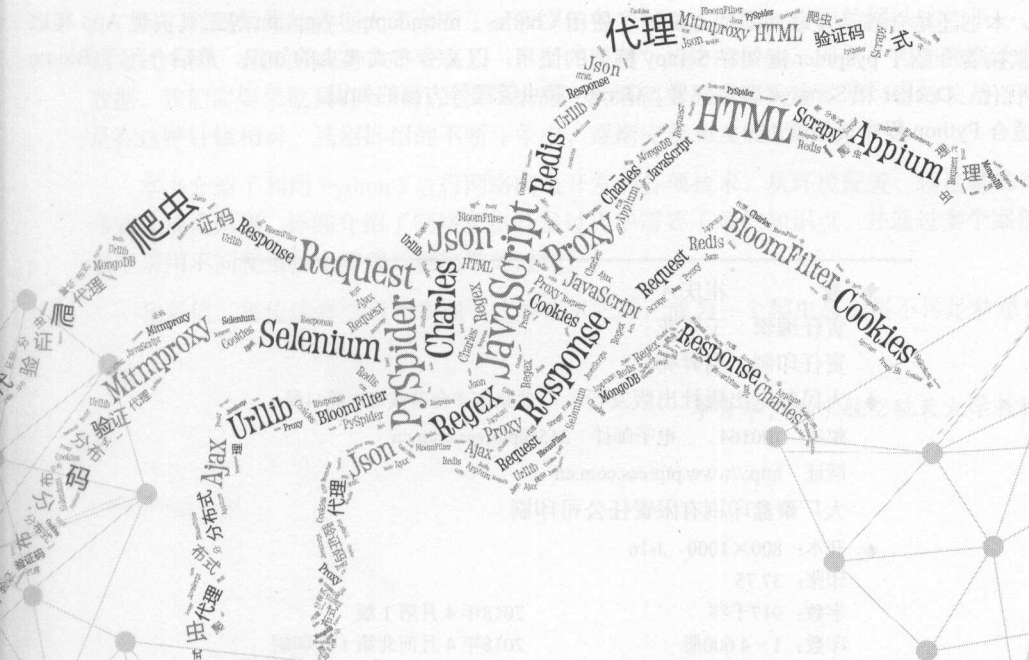
崔庆才

北京航空航天大学硕士，静觅博客（https://cuiqingcai.com/）博主，爬虫博文访问量已过百万，喜欢钻研，热爱生活，乐于分享。欢迎关注个人微信公众号“进击的Coder”，二维码如下：

图灵原创

Python 3 网络爬虫开发实战

崔庆才 著



人民邮电出版社
北 京

图书在版编目（CIP）数据

Python 3网络爬虫开发实战 / 崔庆才著. -- 北京 : 人民邮电出版社, 2018.4
（图灵原创）
ISBN 978-7-115-48034-7

Ⅰ. ①P… Ⅱ. ①崔… 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(2018)第042370号

内 容 提 要

本书介绍了如何利用 Python 3 开发网络爬虫。书中首先详细介绍了环境配置过程和爬虫基础知识；然后讨论了 urllib、requests 等请求库，Beautiful Soup、XPath、pyquery 等解析库以及文本和各类数据库的存储方法；接着通过多个案例介绍了如何进行 Ajax 数据爬取，如何使用 Selenium 和 Splash 进行动态网站爬取；再后介绍了爬虫的一些技巧，比如使用代理爬取和维护动态代理池的方法，ADSL 拨号代理的使用，图形、极验、点触、宫格等各类验证码的破解方法，模拟登录网站爬取的方法及 Cookies 池的维护。

此外，本书还结合移动互联网的特点探讨了使用 Charles、mitmdump、Appium 等工具实现 App 爬取的方法，紧接着介绍了 pyspider 框架和 Scrapy 框架的使用，以及分布式爬虫的知识，最后介绍了 Bloom Filter 效率优化、Docker 和 Scrapyd 爬虫部署、Gerapy 爬虫管理等方面的知识。

本书适合 Python 程序员阅读。

◆ 著　　崔庆才
责任编辑　王军花
责任印制　周昇亮

◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路11号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
大厂聚鑫印刷有限责任公司印刷

◆ 开本：800×1000　1/16
印张：37.75
字数：917千字　　2018年4月第1版
印数：1－4 000册　　2018年4月河北第1次印刷

定价：99.00元

读者服务热线：(010)51095186转600　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

序　一

人类社会已经进入大数据时代，大数据深刻改变着我们的工作和生活。随着互联网、移动互联网、社交网络等的迅猛发展，各种数量庞大、种类繁多、随时随地产生和更新的大数据，蕴含着前所未有的社会价值和商业价值。大数据成为 21 世纪最为重要的经济资源之一。正如马云所言：未来最大的能源不是石油而是大数据。对大数据的获取、处理与分析，以及基于大数据的智能应用，已成为提高未来竞争力的关键要素。

但如何获取这些宝贵数据呢？网络爬虫就是一种高效的信息采集利器，利用它可以快速、准确地采集我们想要的各种数据资源。因此，可以说，网络爬虫技术几乎已成为大数据时代 IT 从业者的必修课程。

我们需要采集的数据大多来源于互联网的各个网站。然而，不同的网站结构不一、布局复杂、渲染方式多样，有的网站还专门采取了一系列“反爬”的防范措施。因此，为准确高效地采集到需要的数据，我们需要采取具有针对性的反制措施。网络爬虫与反爬措施是矛与盾的关系，网络爬虫技术就是在这种针锋相对、见招拆招的不断斗争中，逐渐完善和发展起来的。

本书介绍了利用 Python 3 进行网络爬虫开发的各项技术，从环境配置、理论基础到进阶实战、分布式大规模采集，详细介绍了网络爬虫开发过程中需要了解的知识点，并通过多个案例介绍了不同场景下采用不同爬虫技术实现数据爬取的过程。

我坚信，每位读者学习和掌握了这些技术之后，成为一个爬虫高手将不再是梦想！

李舟军，北京航空航天大学教授，博士生导师

2017 年 10 月

序　二

众所周知，人工智能的这次浪潮和深度学习技术的突破密不可分，却很少有人会谈论另一位幕后英雄，即数据。如果不是网络上有如此多的图片，李飞飞教授也无法构建近千万的标注图片集合ImageNet，从而成就深度学习技术在图像识别领域的突破。如果不是在网络上有了如此多的聊天数据，小冰也不会学习到人类的情商，在聊天中带给人类惊喜、欢笑和抚慰。人工智能的进步离不开数据和算法的结合，人类无意间产生的数据却能够让机器学习到超乎想象的“智慧”，反过来服务人类。

在互联网时代，强大的爬虫技术造就了很多伟大的搜索引擎公司，让人类的记忆搜索能力得到巨大的延展。今天在移动互联网时代，爬虫技术仍然是支撑一些信息融合应用（如今日头条）的关键技术。但是，今天爬虫技术面临着更大的挑战。与互联网的共享机制不同，很多资源只有在登录之后才能访问，还采取了各种反爬虫措施，这就让爬虫不那么容易访问这些资源。无论是产品还是研究，都需要大量的优质数据来让机器更加智能。因此，在这个时代，大量的从业者急需一本全面介绍爬虫技术的书。如果你需要了解全面和前沿的爬虫技术，而且想迅速地上手实战，这本书就是首选。

我很荣幸认识崔庆才先生，他目前还是一名北京航空航天大学在读研究生，正处在一个对技术狂热追求的年纪。我听他讲了一些修炼爬虫技术的故事，很有意思。他在本科的时候因为一个项目开始接触爬虫，之后他用爬虫竟然得到了所在学校同学的照片，还帮助他的哥们儿追其他系的女孩。我问他是否也是用这些信息找到了女友，他甩了下头发，酷酷地说：“需要吗？”

崔庆才是个非常擅长学习的人，他玩什么都能玩到精通。他有一个很好的习惯，就是边学边写，他早期学习爬虫技术的时候，就开了博客，边学边分享他学到并实际操作过的经验，圈粉无数。我很受启发，这样的学习模式很高效，要教给别人之前自己必须弄得特别清楚。另一方面，互联网上的互动也给了他继续学习和精益求精的动力。

除了网络，图书是最成体系的经验分享。本书记录了崔庆才先生对爬虫实战技术最精华的部分。我已经迫不及待地想买一本，也一定会把它推荐给更多的朋友。

宋睿华，微软小冰首席科学家

2017 年 10 月

前　言

为什么写这本书

在这个大数据时代，尤其是人工智能浪潮兴起的时代，不论是工程领域还是研究领域，数据已经成为必不可少的一部分，而数据的获取很大程度上依赖于爬虫的爬取，所以爬虫也逐渐变得火爆起来。我是在 2015 年开始接触爬虫的，当时爬虫其实并没有这么火，我当时觉得能够把想要的数据抓取下来就是一件非常有成就感的事情，而且也可以顺便熟悉 Python，一举两得。在学习期间，我将学到的内容做好总结，发表到博客上。随着我发表的内容越来越多，博客的浏览量也越来越多，很多读者对我的博文给予了肯定的评价，这也给我的爬虫学习之路增添了很多动力。在学习的过程中，困难其实还是非常多的，最早学习时使用的是 Python 2，当时因为编码问题搞得焦头烂额。另外，那时候相关的中文资料还比较少，很多情况下还得自己慢慢去啃官方文档，走了不少弯路。随着学习的进行，我发现爬虫这部分内容涉及的知识点太多、太杂了。网页的结构、渲染方式不同，我们就得换不同的爬取方案来进行针对性的爬取。另外，网页信息的提取、爬取结果的保存也有五花八门的方案。随着移动互联网的兴起，App 的爬取也成了一个热点，而为了提高爬取速度又需要考虑并行爬取、分布式爬取方面的内容，爬虫的通用性、易用性、架构都需要好好优化。这么多杂糅的知识点对于一个爬虫初学者来说，学习的挑战性会非常高，同时学习过程中大家或许也会走我之前走过的弯路，浪费很多时间。后来有一天，图灵的王编辑联系了我，问我有没有意向写一本爬虫方面的书，我听到之后充满了欣喜和期待，这样既能把自己学过的知识点做一个系统整理，又可以跟广大爬虫爱好者分享自己的学习经验，还可以出版自己的作品，于是我很快就答应约稿了。

一开始觉得写书并不是一件那么难的事，后来真正写了才发现其中包含的艰辛。书相比博客来说，用词的严谨性要高很多，而且逻辑需要更加缜密，很多细节必须考虑得非常周全。前前后后写了大半年的时间，审稿和修改又花费了几个月的时间，一路走来甚是不易，不过最后看到书稿成型，觉得这一切都是值得的。在书中，我把我学习爬虫的很多经验都写了进去。环境配置是学习的第一步，环境配置不好，其他工作就没法开展，甚至可能很大程度上打击学习的积极性，所以我在第 1 章中着重介绍了环境的配置过程。而因为操作系统的不同，环境配置过程又各有不同，所以我把每个系统（Windows、Linux、Mac）的环境配置过程都亲自实践了一遍，并梳理记录下来，希望为各位读者在环境配置时多提供一些帮助。后面我又针对爬虫网站的不同情形分门别类地进行了说明，如 Ajax 分析爬取、动态渲染页面爬取、App 爬取、使用代理爬取、模拟登录爬取等知识，每个知识点我都选取了一些典型案例来说明，以便于读者更好地理解整个过程和用法。为了提高代码编写和爬取的效率，还可以使用一些爬虫框架辅助爬取，所以本书后面又介绍了两个流行的爬虫框架的用法，最后又介绍

了一些分布式爬虫及部署方面的知识。总体来说，本书根据我个人觉得比较理想的学习路径介绍了学习爬虫的相关知识，并通过一些实战案例帮助读者更好地理解其中的原理。

本书内容

本书一共分为 15 章，归纳如下。

- 第 1 章介绍了本书所涉及的所有环境的配置详细流程，兼顾 Windows、Linux、Mac 三大平台。本章不用逐节阅读，需要的时候查阅即可。
- 第 2 章介绍了学习爬虫之前需要了解的基础知识，如 HTTP、爬虫、代理的基本原理、网页基本结构等内容，对爬虫没有任何了解的读者建议好好了解这一章的知识。
- 第 3 章介绍了最基本的爬虫操作，一般学习爬虫都是从这一步学起的。这一章介绍了最基本的两个请求库（urllib 和 requests）和正则表达式的基本用法。学会了这一章，就可以掌握最基本的爬虫技术了。
- 第 4 章介绍了页解析库的基本用法，包括 Beautiful Soup、XPath、pyquery 的基本使用方法，它们可以使得信息的提取更加方便、快捷，是爬虫必备利器。
- 第 5 章介绍了数据存储的常见形式及存储操作，包括 TXT、JSON、CSV 各种文件的存储，以及关系型数据库 MySQL 和非关系型数据库 MongoDB、Redis 存储的基本存储操作。学会了这些内容，我们可以灵活方便地保存爬取下来的数据。
- 第 6 章介绍了 Ajax 数据爬取的过程，一些网页的数据可能是通过 Ajax 请求 API 接口的方式加载的，用常规方法无法爬取，本章介绍了使用 Ajax 进行数据爬取的方法。
- 第 7 章介绍了动态渲染页面的爬取，现在越来越多的网站内容是经过 JavaScript 渲染得到的，而原始 HTML 文本可能不包含任何有效内容，而且渲染过程可能涉及某些 JavaScript 加密算法，可以使用 Selenium、Splash 等工具来实现模拟浏览器进行数据爬取的方法。
- 第 8 章介绍了验证码的相关处理方法。验证码是网站反爬虫的重要措施，我们可以通过本章了解到各类验证码的应对方案，包括图形验证码、极验验证码、点触验证码、微博宫格验证码的识别。
- 第 9 章介绍了代理的使用方法，限制 IP 的访问也是网站反爬虫的重要措施。另外，我们也可以使用代理来伪装爬虫的真实 IP，使用代理可以有效解决这个问题。通过本章，我们了解到代理的使用方法，还学习了代理池的维护方法，以及 ADSL 拨号代理的使用方法。
- 第 10 章介绍了模拟登录爬取的方法，某些网站需要登录才可以看到需要的内容，这时就需要用爬虫模拟登录网站再进行爬取了。本章介绍了最基本的模拟登录方法以及维护一个 Cookies 池的方法。
- 第 11 章介绍了 App 的爬取方法，包括基本的 Charles、mitmproxy 抓包软件的使用。此外，还介绍了 mitmdump 对接 Python 脚本进行实时抓取的方法，以及使用 Appium 完全模拟手机 App 的操作进行爬取的方法。
- 第 12 章介绍了 pyspider 爬虫框架及用法，该框架简洁易用、功能强大，可以节省大量开发爬虫的时间。本章结合案例介绍了使用该框架进行爬虫开发的方法。

- 第 13 章介绍了 Scrapy 爬虫框架及用法。Scrapy 是目前使用最广泛的爬虫框架，本章介绍了它的基本架构、原理及各个组件的使用方法，另外还介绍了 Scrapy 通用化配置、对接 Docker 的一些方法。
- 第 14 章介绍了分布式爬虫的基本原理及实现方法。为了提高爬取效率，分布式爬虫是必不可少的，本章介绍了使用 Scrapy 和 Redis 实现分布式爬虫的方法。
- 第 15 章介绍了分布式爬虫的部署及管理方法。方便快速地完成爬虫的分布式部署，可以节省开发者大量的时间。本章结合 Scrapy、Scrapyd、Docker、Gerapy 等工具介绍了分布式爬虫部署和管理的实现。

致谢

感谢我的父母、导师，没有他们创造的环境，我不可能完成此书的写作。

感谢我的女朋友李园，在我写书期间给了我很多的支持和鼓励。同时她还主导设计了本书的封面，正是她的理解和付出才使本书得以完善。

感谢在我学习过程中与我探讨技术的各位朋友，特别感谢汪海洋先生在我初学爬虫过程中给我提供的指导，特别感谢崔弦毅、苟桃、时猛先生在我写书过程中为我提供的思路和建议。

感谢为本书撰写推荐语的李舟军老师、宋睿华老师、梁斌老师、施水才老师（排名不分先后），感谢你们对本书的支持和推荐。

感谢王军花、陈兴璐编辑，在书稿的审核过程中给我提供了非常多的建议，没有你们的策划和敦促，我也难以顺利完成此书。

感谢为本书做出贡献的每一个人！

相关资源

本书中的所有代码都放在了 GitHub（详见 https://github.com/Python3WebSpider），书中每个实例对应的章节末也有说明。

本人的个人博客也会更新爬虫相关文章，欢迎读者访问交流，博客地址：https://cuiqingcai.com/。

崔庆才

2018 年 1 月

目　录

第 1 章 开发环境配置

工欲善其事，必先利其器！

编写和运行程序之前，我们必须先把开发环境配置好。只有配置好了环境并且有了更方便的开发工具，我们才能更加高效地用程序实现相应的功能。然而很多情况下，我们可能在最开始就卡在环境配置上，如果这个过程花费了太多时间，学习的兴趣可能就下降了大半，所以本章专门对本书中所有的环境配置做一下说明。

本章将讲解书中使用的所有库及工具的安装过程。为了使书的条理更加清晰，本书将环境配置的过程统一合并为一章。本章不必逐节阅读，可以在需要的时候查阅。

在介绍安装过程时，我们会尽量兼顾各个平台。另外，书中也会指出一些常见的安装错误，以便快速高效地搭建好编程环境。

1.1 Python 3 的安装

既然要用 Python 3 开发爬虫，那么第一步一定是安装 Python 3。这里会介绍 Windows、Linux 和 Mac 三大平台下的安装过程。相关链接如下。

- 官方网站：http://python.org
- 下载地址：https://www.python.org/downloads
- 第三方库：https://pypi.python.org/pypi
- 官方文档：https://docs.python.org/3
- 中文教程：http://www.runoob.com/python3/python3-tutorial.html
- Awesome Python：https://github.com/vinta/awesome-python
- Awesome Python 中文版：https://github.com/jobbole/awesome-python-cn

1.1.1 Windows 下的安装

在 Windows 下安装 Python 3 的方式有两种。

- 一种是通过 Anaconda 安装，它提供了 Python 的科学计算环境，里面自带了 Python 以及常用的库。如果选用了这种方式，后面的环境配置方式会更加简便。
- 另一种是直接下载安装包安装，即标准的安装方式。

下面我们依次介绍这两种安装方式，任选其一即可。

1. Anaconda 安装

Anaconda 的官方下载链接为 https://www.continuum.io/downloads，选择 Python 3 版本的安装包下载即可，如图 1-1 所示。

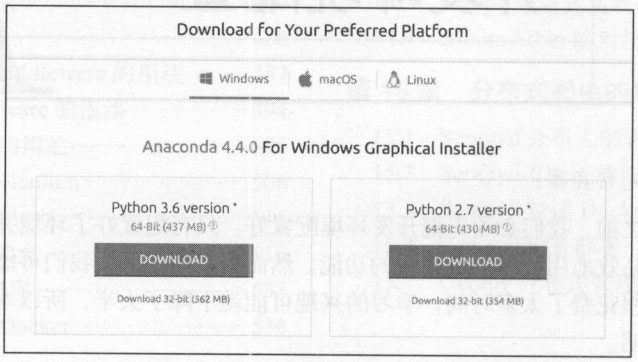


图 1-1　Anaconda Windows 下载页面

如果下载速度过慢，可以选择使用清华大学镜像，下载列表链接为 https://mirrors.tuna.tsinghua.edu.cn/anaconda/archive/，使用说明链接为 https://mirrors.tuna.tsinghua.edu.cn/help/anaconda/。

下载完成之后，直接双击安装包安装即可。安装完成之后，Python 3 的环境就配置好了。

2. 安装包安装

我们推荐直接下载安装包来安装，此时可以直接到官方网站下载 Python 3 的安装包：https://www.python.org/downloads/。

写书时，Python 的最新版本[①]是 3.6.2，其下载链接为 https://www.python.org/downloads/release/python-362/，下载页面如图 1-2 所示。需要说明的是，实际的 Python 最新版本以官网为准。

Files

Version	Operating System	Description	MD5 Sum	File Size	GPG
Gzipped source tarball	Source release		2d0fc9f3a5940707590e07f03ecb08b9	22540566	SIG
XZ compressed source tarball	Source release		692b4fc3a2ba0d54d1495d4ead5b0b5c	16872064	SIG
Mac OS X 64-bit/32-bit installer	Mac OS X	for Mac OS X 10.6 and later	6dd08e7027d2a1b3a2c02cfacbe611ef	27511848	SIG
Windows help file	Windows		69082441d723060fb333dcda8815105e	7986690	SIG
Windows x86-64 embeddable zip file	Windows	for AMD64/EM64T/x64, not itanium processors	708496ebbe9a730d19d5d288afd216f1	6926999	SIG
Windows x86-64 executable installer	Windows	for AMD64/EM64T/x64, not itanium processors	ad69fdacde90f2ce8286c279b11ca188	31392272	SIG
Windows x86-64 web-based installer	Windows	for AMD64/EM64T/x64, not itanium processors	a055a1a0e938e74c712a1c495261ae6c	1312520	SIG
Windows x86 embeddable zip file	Windows		8dff09a1b19b7a7dcb915765328484cf	6320763	SIG
Windows x86 executable installer	Windows		3773db079c173bd6d8a631896c72a88f	30453192	SIG
Windows x86 web-based installer	Windows		f58f01933f39e0b45a0ae68027888d7	1287064	SIG

图 1-2　Python 下载页面

① 若无特别说明，书中的最新版本均为作者写书时的情况，后面不再一一说明。

64 位系统可以下载 Windows x86-64 executable installer，32 位系统可以下载 Windows x86 executable installer。

下载完成之后，直接双击 Python 安装包，然后通过图形界面安装，接着设置 Python 的安装路径，完成后将 Python 3 和 Python 3 的 Scripts 目录配置到环境变量即可。

关于环境变量的配置，此处以 Windows 10 系统为例进行演示。

假如安装后的 Python 3 路径为 C:\Python36，从资源管理器中打开该路径，如图 1-3 所示。

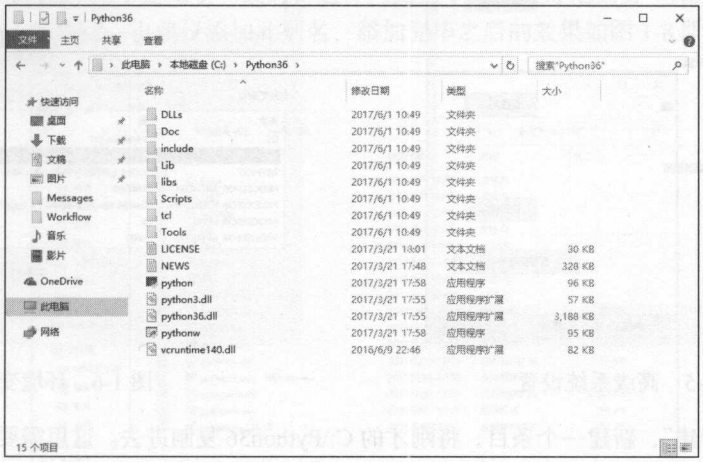


图 1-3 Python 安装目录

将该路径复制下来。

随后，右击“计算机”，从中选择“属性”，此时将打开系统属性窗口，如图 1-4 所示。

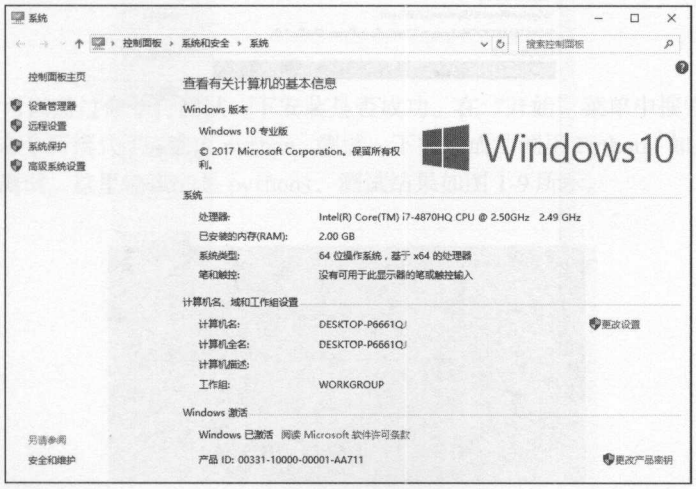


图 1-4 系统属性

点击左侧的“高级系统设置”，即可在弹出的对话框下方看到“环境变量”按钮，如图 1-5 所示。

点击“环境变量”按钮，找到系统变量下的 Path 变量，随后点击“编辑”按钮，如图 1-6 所示。

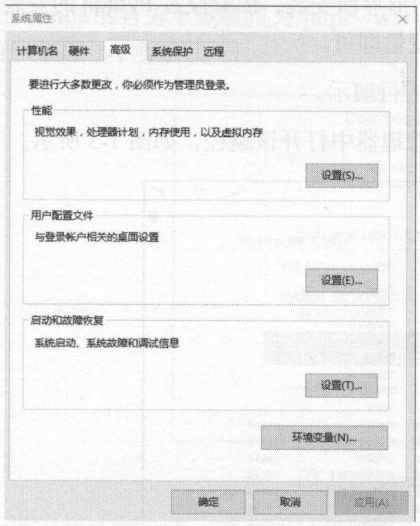


图 1-5　高级系统设置

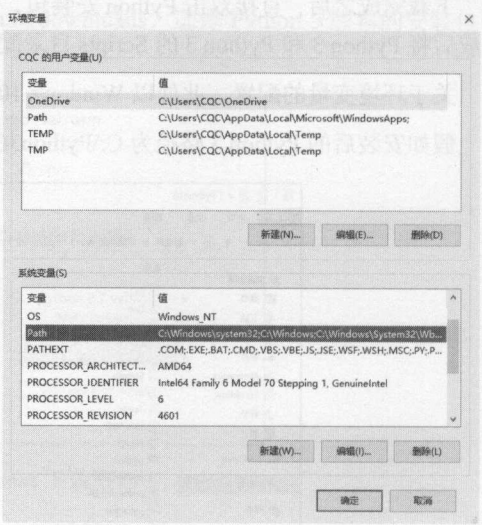


图 1-6　环境变量

随后点击“新建”，新建一个条目，将刚才的 C:\Python36 复制进去。这里需要说明的是，此处的路径就是你的 Python 3 安装目录，请自行替换。然后，再把 C:\Python36\Scripts 路径复制进去，如图 1-7 所示。

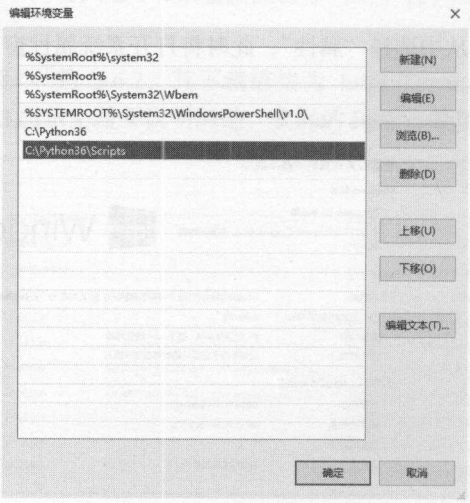


图 1-7　编辑环境变量

最后，点击“确定”按钮即可完成环境变量的配置。

配置好环境变量后，我们就可以在命令行中直接执行环境变量路径下的可执行文件了，如 python、pip 等命令。

3. 添加别名

上面这两种安装方式任选其一即可完成安装，但如果之前安装过 Python 2 的话，可能会导致版本冲突问题，比如在命令行下输入 python 就不知道是调用的 Python 2 还是 Python 3 了。为了解决这个问题，建议将安装目录中的 python.exe 复制一份，命名为 python3.exe，这样便可以调用 python3 命令了。实际上，它和 python 命令是完全一致的，这样只是为了可以更好地区分 Python 版本。当然，如果没有安装过 Python 2 的话，也建议添加此别名，添加完毕之后的效果如图 1-8 所示。

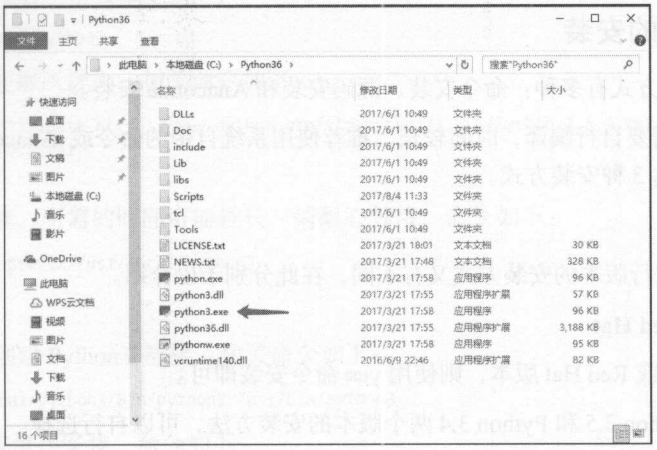


图 1-8 添加别名

对于 pip 来说，安装包中自带了 pip3.exe 可执行文件，我们也可以直接使用 pip3 命令，无需额外配置。

4. 测试验证

安装完成后，可以通过命令行测试一下安装是否成功。在“开始”菜单中搜索 cmd，找到命令提示符，此时就进入命令行模式了。输入 python，测试一下能否成功调用 Python。如果添加了别名的话，可以输入 python3 测试。这里输入的是 python3，测试结果如图 1-9 所示。

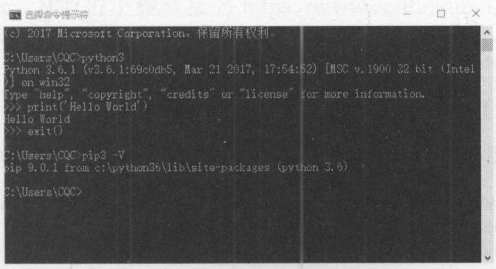


图 1-9 测试验证页面