大数据工程师的Python快速入门书
Scrapy、Beautiful Soup 网络爬虫实战
Mechanize、Selenium 模拟浏览器实战

Web Scraping with Python

Python 网络爬虫实战

胡松涛 编著

本书示例源代码下载

清華大學出版社

Python
网络爬虫实战

胡松涛 编著

清華大學出版社
北京

本书从 Python 的安装开始，详细讲解了 Python 从简单程序延伸到 Python 网络爬虫的全过程。本书从实战出发，根据不同的需求选取不同的爬虫，有针对性地讲解了几种 Python 网络爬虫。

本书共 8 章，涵盖的内容有 Python 语言的基本语法、Python 常用 IDE 的使用、Python 第三方模块的导入使用、Python 爬虫常用模块、Scrapy 爬虫、Beautiful Soup 爬虫、Mechanize 模拟浏览器和 Selenium 模拟浏览器。本书所有源代码已上传网盘供读者下载。

本书内容丰富，实例典型，实用性强。适合 Python 网络爬虫初学者、数据分析与挖掘技术初学者，以及高校及培训学校相关专业的师生阅读。

图书在版编目（CIP）数据

Python 网络爬虫实战 / 胡松涛编著. — 北京：清华大学出版社，2017
ISBN 978-7-302-45787-9

I. ①P… II. ①胡… III. ①软件工具—程序设计 IV. ①TP311.56

中国版本图书馆 CIP 数据核字（2016）第 290543 号

责任编辑：夏非彼
封面设计：王　翔
责任校对：闫秀华
责任印制：杨　艳

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 刷 者：北京富博印刷有限公司
装 订 者：北京市密云县京文制本装订厂
经　　销：全国新华书店
开　　本：190mm×260mm　　**印　　张**：19　　**字　　数**：486 千字
版　　次：2017 年 1 月第 1 版　　**印　　次**：2017 年 1 月第1 次印刷
印　　数：1～3000
定　　价：59.00 元

产品编号：071193-01

前　言

计算机技术飞速发展，人们对计算机使用技能的要求也越来越高。在编写软件时，大家既希望有超高的效率，又希望这门语言简单易用。这种鱼与熊掌皆得的要求的确很高，Python编程语言恰好符合这么苛刻的要求。

Python的执行效率仅比效率之王C略差一筹，在简单易用方面Python也名列三甲。可以说Python在效率和简单之间达到了平衡。另外，Python还是一门胶水语言，可以将其他编程语言的优点融合在一起，达到1+1>2的效果。这也是Python如今使用人数越来越多的原因。

Python 语言发展迅速，在各行各业都发挥独特的作用。在各大企业、学校、机关都运行着Python明星程序。但就个人而言，运用Python最多的还是网络爬虫（这里的爬虫仅涉及从网页提取数据，不涉及深度、广度算法爬虫搜索）。在网络上经常更新的数据，无须每次都打开网页浏览，使用爬虫程序，一键获取数据，下载保存后分析。考虑到Python爬虫在网络上的资料虽多，但大多都不成系统，难以提供系统有效的学习。因此笔者抛砖引玉，编写了这本有关Python网络爬虫的书，以供读者学习参考。

Python简单易学，Python爬虫也不复杂。只需要了解了Python的基本操作即可自行编写。本书中介绍了几种不同类型的Python爬虫，可以针对不同情况的站点进行数据收集。

本书特色

- 附带全部源代码

为了便于读者理解本书内容，作者已将全部的源代码上传到网络，供读者下载使用。读者通过代码学习开发思路，精简优化代码。

- 涵盖了Linux&Windows上模块的安装配置

本书包含了Python模块源的配置、模块的安装，以及常用IDE的使用。

- 实战实例

通过常用的实例，详细说明网络爬虫的编写过程。

本书结构

本书共 8 章，前面 4 章简单地介绍了 Python 的基本用法和简单 Python 程序的编写。第 5 章的 Scrapy 爬虫框架主要针对一般无须登录的网站，在爬取大量数据时使用 Scrapy 会很方便。第 6 章的 Beautiful Soup 爬虫可以算作爬虫的“个人版”。Beautiful Soup 爬虫主要针对一些爬取数据比较少的，结构简单的网站。第 7 章的 Mechanize 模块，主要功能是模拟浏览器。它的作用主要是针对那些需要登录验证的网站。第 8 章的 Selenium 模块，主要功能也是模拟浏览器，它的作用主要是针对 JavaScript 返回数据的网站。

本书读者与作者

- Python 网络爬虫初学者
- 数据分析与挖掘技术初学者
- 高校和培训学校相关专业的师生

本书由胡松涛主笔，其他参与编写的有宋士伟、张倩、彭霁、杨旺功、邹瑛、王铁民、殷龙、李春城、张兴瑜、刘祥淼、李柯泉、林龙、赵殿华、牛晓云。

本书代码下载

本书代码下载地址（注意数字和字母大小写）为 http://pan.baidu.com/s/1miTmq5y。如果下载有问题，请电子邮件联系booksaga@163.com，邮件主题为“网络爬虫代码”。

编　者

2016 年 11 月

目 录

第 1 章

◂Python环境配置▸

为什么选择 Python 来写网络爬虫？

众所周知 Python 的速度并不是最快的，比不上 Java，比不上 C++，更比不上传说中的速度效率之王 C 了。学习资料的完备也不在三甲之内，市面上讲解 C&C++的书籍绝对是 Python 的几倍甚至几十倍。使用的人数也不是最多，比不上 Java、C、C++。

那么，为什么会选择 Python？

首先是它简单易学。简单到没有学过任何编程语言的人稍微看下资料，再看几个示例就可以编写出可用的程序；其次它是一门解释型编程语言，编写完毕后可直接执行，无须编译，发现 Bug 后立即修改，省下了无数的编译时间；还有它的代码重用性高，可以把包含某个功能的程序当成模块代入其他程序中使用，因而 Python 的模块库庞大到恐怖，几乎是无所不包；最后就是因为它的跨平台性，几乎所有的 Python 程序，都可以不加修改地运行在不同的操作平台，都能得到同样的结果。这么多的优点都集中在这个语言中，因此最好的选择就是 Python。

1.1 Python 简介

了解一门语言，我们先从它的历史说起。Python 的应用越来越广泛，它最初是用来做什么用的，之后又如何发展的，了解这些，我们就更能了解 Python。

1.1.1 Python 的历史由来

Python 是一种开源的面向对象的脚本语言，它起源于 1989 年末，当时，CWI（阿姆斯特丹国家数学和计算机科学研究所）的研究员 Guido van Rossum 需要一种高级脚本编程语言，为其研究小组的 Amoeba 分布式操作系统执行管理任务。为创建新语言，他从高级数学语言 ABC（ALL BASIC CODE）汲取了大量语法，并从系统编程语言 Modula-3 借鉴了错误处理机制。Van Rossum 把这种新的语言命名为 Python（大蟒蛇）——来源于 BBC 当时正在热播的喜剧连续剧 Monty Python。

ABC 是由 Guido 参加设计的一种教学语言。就 Guido 本人看来，ABC 这种语言非常优

美和强大，是专门为非专业程序员设计的。但是 ABC 语言并没有成功，究其原因，Guido 认为是非开放造成的。Guido 决心在 Python 中避免这一错误。同时，他还想实现在 ABC 中闪现过但未曾实现的东西。

就这样，Python 在 Guido 手中诞生了。可以说，Python 是从 ABC 发展起来，并且结合了 Unix shell 和 C 的习惯。Python 源代码遵循 GPL（GNU General Public License）协议。所以任何个人用户都可以免费使用。

1.1.2 Python 的现状

Python 于 1991 年初公开发行，由于功能强大和采用开源方式发行，Python 发展得很快，用户越来越多，形成了一个强大的社区力量。2001 年，Python 的核心开发团队移师 Digital Creations 公司，该公司是 Zope（一个用 Python 编写的 Web 应用服务器）的创始者。大家可到 http://www.python.org/上了解最新的 Python 动态和资料 。

如今，Python 已经成为最受欢迎的程序设计语言之一。2011 年 1 月，它被 TIOBE 编程语言排行榜评为 2010 年度语言。自从 2004 年以后，Python 的使用率是呈线性增长 。

1.1.3 Python 的应用

Python 应用广泛，特别适用与以下几个方面。

- 系统编程：提供 API（Application Programming Interface，应用程序编程接口），能方便地进行系统维护和管理，Linux 下标志性语言之一，是很多系统管理员理想的编程工具 。
- 图形处理：有 PIL、Tkinter 等图形库支持，能方便进行图形处理。
- 数学处理：NumPy 扩展提供大量与许多标准数学库的接口。
- 文本处理：Python 提供的 re 模块能支持正则表达式，还提供 SGML、XML 分析模块，许多程序员利用 Python 进行 XML 程序的开发。
- 数据库编程：程序员可通过遵循 Python DB-API（数据库应用程序编程接口）规范的模块与 Microsoft SQL Server、Oracle、Sybase、DB2、MySQL、SQLite 等数据库通信。Python 自带有一个 Gadfly 模块，提供了一个完整的 SQL 环境。
- 网络编程：提供丰富的模块支持 sockets 编程，能方便快速地开发分布式应用程序。很多大规模软件开发计划，例如 Zope、Mnet 及 BitTorrent、Google 都在广泛地使用它。
- Web 编程：应用的开发语言，支持最新的 XML 技术。
- 多媒体应用：Python 的 PyOpenGL 模块封装了 OpenGL 应用程序编程接口，能进行二维和三维图像处理。PyGame 模块可用于编写游戏软件。
- PYMO 引擎：PYMO 全称为 Python Memories Off，是一款运行于 Symbian S60V3、Symbian3、S60V5、Symbian3、Android 系统上的 AVG 游戏引擎。因其基于 Python2.0 平台开发，并且适用于创建秋之回忆（memories off）风格的 AVG 游戏，

故命名为 PYMO。

不只个人用户推崇 Python，企业用户也对 Python 青睐有加，以下是明星企业的应用项目：

- Reddit：社交分享网站，最早用 Lisp 开发，在 2005 年转为 python。
- Dropbox：文件分享服务。
- 豆瓣网：图书、唱片、电影等文化产品的资料数据库网站。
- Django：鼓励快速开发的 Web 应用框架。
- Fabric：用于管理成百上千台 Linux 主机的程序库。
- EVE：网络游戏 EVE 大量使用 Python 进行开发。
- Blender：以 C 与 Python 开发的开源 3D 绘图软件。
- BitTorrent：bt 下载软件客户端。
- Ubuntu Software Center：Ubuntu 9.10 版本后自带的图形化包管理器。
- YUM：用于 RPM 兼容的 Linux 系统上的包管理器。
- Civilization IV：游戏《文明 4》。
- Battlefield 2：游戏《战地 2》。
- Google：谷歌在很多项目中用 python 作为网络应用的后端，如 Google Groups、Gmail、Google Maps 等，Google App Engine 支持 python 作为开发语言。
- NASA：美国宇航局，从 1994 年起把 python 作为主要开发语言。
- Industrial Light & Magic：工业光魔，乔治·卢卡斯创立的电影特效公司。
- Yahoo! Groups：雅虎推出的群组交流平台。
- YouTube：视频分享网站，在某些功能上使用到 python。
- Cinema 4D：一套整合 3D 模型、动画与绘图的高级三维绘图软件，以其高速的运算和强大的渲染插件著称。
- Autodesk Maya：3D 建模软件，支持 python 作为脚本语言。
- gedit：Linux 平台的文本编辑器。
- GIMP：Linux 平台的图像处理软件。
- Minecraft: Pi Edition：游戏《Minecraft》的树莓派版本。
- MySQL Workbench：可视化数据库管理工具。
- Digg：社交新闻分享网站。
- Mozilla：为支持和领导开源的 Mozilla 项目而设立的一个非营利组织。
- Quora：社交问答网站。
- Path：私密社交应用。
- Pinterest：图片社交分享网站。
- SlideShare：幻灯片存储、展示、分享的网站。
- Yelp：美国商户点评网站。
- Slide：社交游戏/应用开发公司，被谷歌收购。

还有很多企业级的应用这里就不一一列举了。Python 适用于不同的场合、不同的人群，

是适应性非常强的一门语言。

1.2 Python 开发环境配置

Python 在 PC 三大主流平台（Windows、Linux 和 OS X）都可使用。在这里只讲解 Windows 和 Linux 下的开发环境配置。Windows 平台以 Windows 7 为例，Linux 平台以 Debian 8 系统为例。Python 目前主要有两个版本，Python 2 和 Python 3。目前，Python 2 的最终版本是 Python 2.7.11，Python 3 的最终版本是 Python 3.5.1。Python 3 虽然功能更加强大，但暂时 Python 2 的使用人数更多，本书中全部选择 Python 2.7 为例。

1.2.1 Windows 下安装 Python

（1）打开 Chrome 浏览器，在地址栏输入 Python 官网地址 www.python.org，如图 1-1 所示。

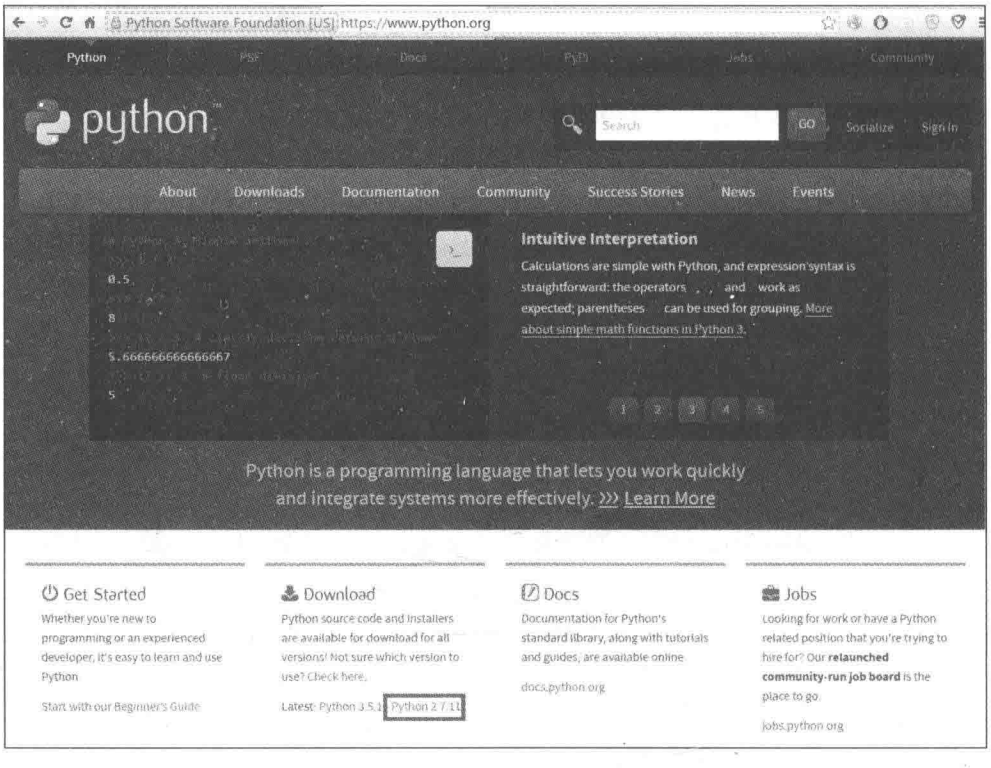

图 1-1 Python 官网

（2）单击 Python 2.7.11，进入 Python 2.7.11 的下载页面，如图 1-2 所示。

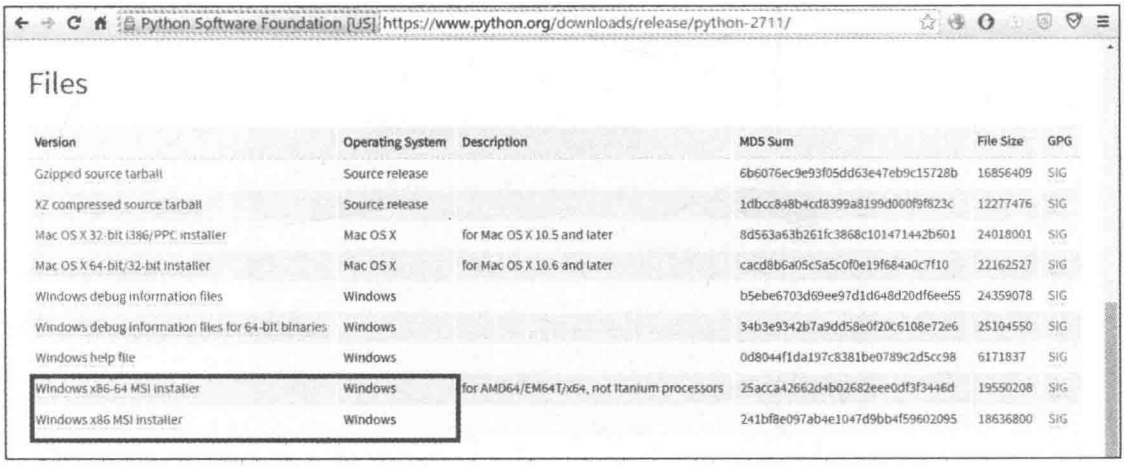

图 1-2　Python 下载

（3）按照安装的 Windows 系统选择下载的安装文件。示例系统是 Windows 7 64 位版本，所以在此下载的是 Windows x86-64 installer。

（4）下载完毕，得到安装文件 python-2.7.11.amd64.msi。双击该文件图标，开始安装 Python 2.7，如图 1-3 所示。

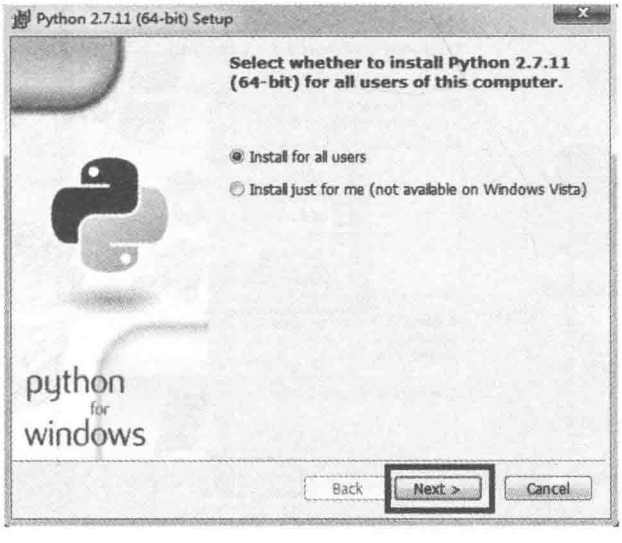

图 1-3　安装 Python

（5）单击 Next 按钮，设置 Python 安装路径，如图 1-4 所示。

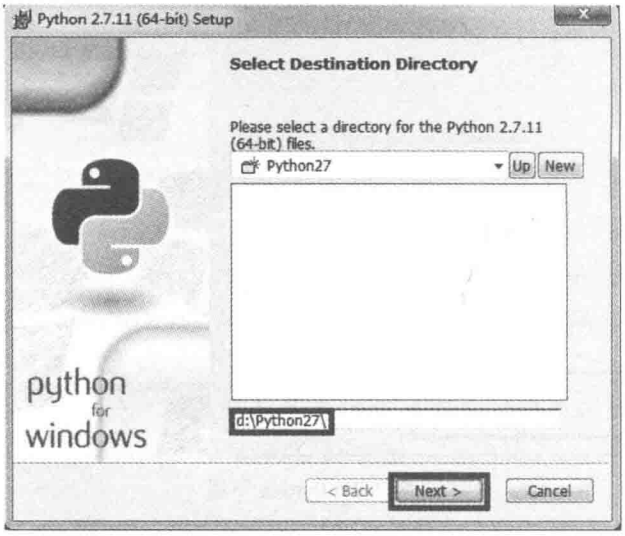

图 1-4　设置 Python 安装路径

（6）选择或者填入 Python 的安装路径后，单击 Next 按钮，进入 Python 组件设置，如图 1-5 所示。

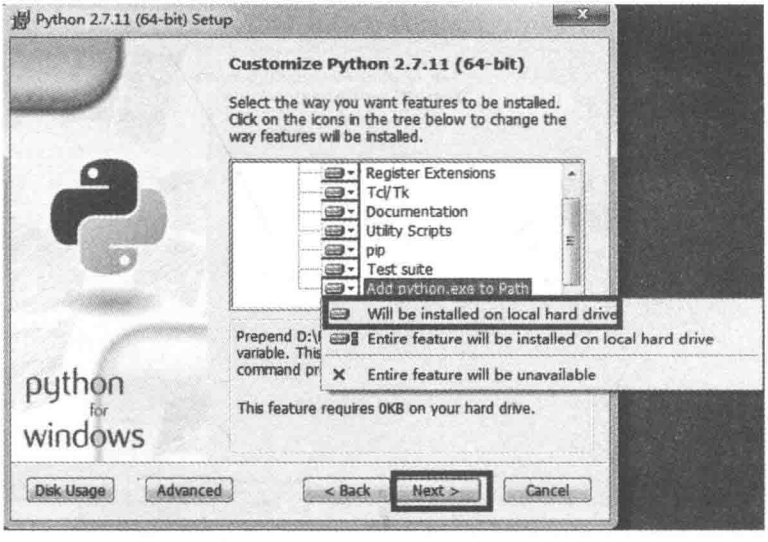

图 1-5　Python 组件选择

默认情况下 Add python.exe to Path 这个组件是未选择的，它的作用是将 Python 的路径加入系统环境中。请将它选择上，单击 Next 按钮，开始安装 Python，如图 1-6 所示。