



Python 基础与 大数据应用

丁辉 ◎ 主编

商俊燕 伍转华 林雪纲 ◎ 副主编

Python Basic and Big Data Application

- ◆ 内容安排遵循学生的认知规律，由浅入深、循序渐进
- ◆ 所选的教学案例，均来自实际应用领域
- ◆ 代码完整，配套资源丰富



中国工信出版集团



人民邮电出版社
POSTS & TELECOM PRESS



Python 基础与 大数据应用

免 / 费 / 提 / 供

PPT等教学相关资料



人邮教育
www.ryjiaoyu.com

教材服务热线: 010-81055256

反馈 / 投稿 / 推荐信箱: 315@ptpress.com.cn

人民邮电出版社教育服务与资源下载社区: www.ryjiaoyu.com

美术编辑: 董志桢

封面设计: 秦延新

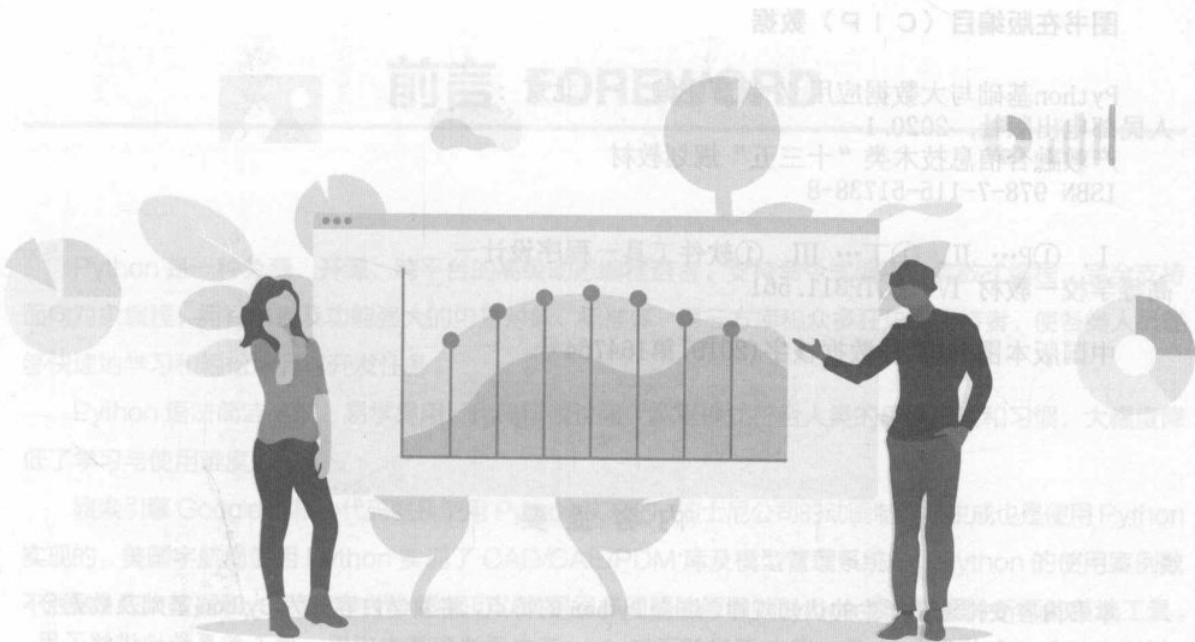
ISBN 978-7-115-51738-8



9 787115 517388 >

ISBN 978-7-115-51738-8

定价: 59.80 元



Python 基础与 大数据应用

丁辉 主编

商俊燕 伍转华 林雪纲 副主编

Python Basic and
Big Data Application

人民邮电出版社
北京

图书在版编目（CIP）数据

Python基础与大数据应用 / 丁辉主编. — 北京 :
人民邮电出版社, 2020.1
产教融合信息技术类“十三五”规划教材
ISBN 978-7-115-51738-8

I. ①P… II. ①丁… III. ①软件工具—程序设计—
高等学校—教材 IV. ①TP311.561

中国版本图书馆CIP数据核字(2019)第164764号

内 容 提 要

本书内容安排遵循学生的认知规律,结合Python的特点,将教学内容分为Python基础及数据分析两大部分。全书分为11章,前6章讲解了Python基本语法和基本应用,后5章系统地讲解了用Python爬取数据、处理数据、分析数据的方法与过程。每章除了丰富的实例外,还设计了来源于实践的项目训练及拓展训练项目,引导读者学以致用。

本书可以作为高校计算机类专业和相关专业的教材,也可作为Python爱好者的自学用书。

-
- ◆ 主 编 丁 辉
 - 副 主 编 商俊燕 伍转华 林雪纲
 - 责任编辑 左仲海
 - 责任印制 马振武
 - ◆ 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>
 - 北京鑫正大印刷有限公司印刷
 - ◆ 开本: 787×1092 1/16
 - 印张: 20.5 2020年1月第1版
 - 字数: 592千字 2020年1月北京第1次印刷
-

定价: 59.80 元

读者服务热线: (010)81055256 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

前言 FOREWORD

Python 是一种免费、开源、跨平台的高级动态编程语言，支持命令式编程、函数式编程，完全支持面向对象编程，拥有丰富及功能强大的内置对象、标准库、第三方库和众多狂热的支持者，使各类人员能够快速学习和轻松地完成开发任务。

Python 语法简洁清晰、易学易用，代码可读性强，编程模式符合人类的思维方式和习惯，大幅度降低了学习与使用难度。

搜索引擎 Google 的核心代码就是使用 Python 实现的，迪士尼公司的动画制作与生成也是使用 Python 实现的，美国宇航局使用 Python 实现了 CAD/CAE/PDM 库及模型管理系统等，Python 的使用案例数不胜数。在大数据和人工智能时代，数据可以说明很多问题的原因，Python 成为数据分析师的首选工具。

本书针对高职高专学生的认知规律、Python 的特点和数据分析过程，以培养学生逻辑思维能力、编程解决实际问题能力、程序调试能力为目标，精心设计教学内容、项目训练、练习题、拓展训练项目等。全书内容由 Python 基础和数据分析两大部分组成，教学时，可以根据需要进行选取。本书主要有以下几个特色。

(1) 教学案例均来自实际应用领域。丰富的、有意义的真实案例使学生实现知识与技能、经验、素养的综合提升。

(2) 语言精练，实践性强。全书语言精练，与 Python 简洁、优雅的特点相匹配，同时所有的命令和案例都是可操作的，学习者完全可以自己通过实践来体验，在实践中掌握 Python 的语法规则和数据分析方法。

(3) 内容丰富，代码完整。全书共 11 章，前 6 章为 Python 基本语法及基本应用部分，后 5 章为数据分析部分。Python 基本语法及基本应用部分内容全面，数据分析部分可根据教学目标灵活选择。全书代码完整，实践性强。

丁辉为本书主编，并且编写了第 1~5 章、第 10 和 11 章；商俊燕、伍转华和林雪纲为副主编，商俊燕编写了第 6~8 章，伍转华编写了第 9 章；其中，全部的拓展训练项目由北京西普阳光教育科技股份有限公司的林雪纲编写。在本书的编写过程中，北京西普阳光教育科技股份有限公司给予了大力支持，在此表示感谢。

Python 及其应用发展迅速，加之编者水平有限，书中难免有不足和遗漏之处，欢迎读者提出宝贵的意见和建议。编者电子邮箱为 dinghui@czili.edu.cn。

编者

2019 年 6 月

目录 CONTENTS

第1章

Python 环境搭建	1
1.1 Python 版本概述及下载 Python 安装文件	1
1.1.1 Python 版本概述	1
1.1.2 下载 Python 安装文件	2
1.2 安装 Python	5
1.2.1 解压下载的文件	5
1.2.2 运行 Python 安装文件	5
1.3 Python 交互模式	6
1.4 iPython 3 和 PyCharm 概述	7
1.4.1 iPython 3 概述	7
1.4.2 PyCharm 概述	9
1.5 项目训练: Python 的安装与使用	13
1.6 本章小结	15
1.7 练习	15
1.8 拓展训练项目	16

第2章

Python 编程基础	17
2.1 变量	17
2.2 数值	19
2.2.1 整型	19
2.2.2 浮点型	19
2.2.3 复数型	20
2.2.4 布尔型	20
2.3 字符串	21
2.3.1 转义字符	21
2.3.2 字符串运算	21
2.3.3 字符串操作方法	22
2.4 列表	23
2.4.1 列表的创建与删除	23

2.4.2	列表操作方法	24
2.4.3	列表切片操作	25
2.5	元组	27
2.6	字典	28
2.6.1	字典的创建与访问	28
2.6.2	字典元素的修改	28
2.6.3	字典操作方法	29
2.7	运算符	30
2.7.1	算术运算符	30
2.7.2	位运算符	31
2.7.3	逻辑运算符	31
2.7.4	比较运算符	32
2.7.5	赋值运算符	32
2.7.6	其他运算符	33
2.8	Python 代码编写规范	34
2.9	控制流	35
2.9.1	顺序结构程序	35
2.9.2	分支结构程序	40
2.9.3	循环结构程序	45
2.10	项目训练：个人所得税计算	53
2.11	本章小结	56
2.12	练习	56
2.13	拓展训练项目	58
2.13.1	Python 数值型变量的定义与赋值	58
2.13.2	Python 控制流和运算符	59
2.13.3	列表的基本操作	59
2.13.4	元组的基本操作	59
2.13.5	字典的基本操作	59
2.13.6	字符串的基本操作	60

第3章

函数	61
3.1 自定义函数	61
3.1.1 函数定义格式	61
3.1.2 函数的设计	62
3.1.3 lambda 表达式	62

3.2	函数调用	63
3.3	函数参数	64
3.3.1	位置参数	64
3.3.2	默认参数	64
3.3.3	关键参数	65
3.3.4	可变长度参数	66
3.4	变量作用域	68
3.4.1	局部变量	68
3.4.2	全局变量	69
3.5	异常	70
3.5.1	Python 标准异常类	70
3.5.2	异常处理	71
3.6	项目训练：哥德巴赫狂想——任何大于 2 的偶数总可以分解成两个素数的和	75
3.7	本章小结	76
3.8	练习	77
3.9	拓展训练项目	78
3.9.1	用函数实现乘法口诀	78
3.9.2	Python 函数参数	78
3.9.3	Python 局部变量和全局变量	78
3.9.4	Python 异常捕获与处理	78

第 4 章

面向对象编程基础	79
4.1 类和对象	79
4.1.1 类	79
4.1.2 对象	80
4.2 属性与方法	80
4.2.1 属性	80
4.2.2 方法	83
4.3 继承	84
4.4 多态	86
4.5 项目训练：简单学生成绩管理系统	88
4.6 本章小结	91
4.7 练习	92
4.8 拓展训练项目	93
4.8.1 Python 类与对象	93

4.8.2	类方法、实例方法和静态方法	93
4.8.3	类继承、组合	94
4.8.4	类的多重继承	94

第 5 章

模块	95
5.1 模块的创建和命名空间	95
5.1.1 模块的创建	95
5.1.2 命名空间	96
5.2 模块的导入和路径	97
5.2.1 模块的导入	97
5.2.2 模块的路径	98
5.3 包	100
5.4 Python 内置模块	100
5.4.1 math 模块	100
5.4.2 random 模块	101
5.4.3 time 模块	101
5.4.4 datetime 模块	103
5.4.5 calendar 模块	104
5.4.6 sys 模块	105
5.4.7 zipfile 模块	106
5.5 项目训练：日历	108
5.6 本章小结	110
5.7 练习	111
5.8 拓展训练项目	111
5.8.1 Python 模块导入	111
5.8.2 zipfile 模块的使用	112
5.8.3 Python 模块的属性	112
5.8.4 Python 模块内置函数	112

第 6 章

Python 文件和数据库	113
6.1 文件的基本操作	113
6.1.1 内置函数 open()	113
6.1.2 文件对象常用的属性和方法	114

6.1.3 文件操作案例	116
6.2 文件系统的基本操作	119
6.3 MySQL 数据库	121
6.3.1 MySQL 简介	121
6.3.2 安装 MySQL	122
6.3.3 使用 Python 连接 MySQL 数据库	126
6.3.4 MySQL 的基本操作	127
6.4 项目训练：使用 Python 完成课程表和学生信息表的创建	128
6.5 本章小结	131
6.6 练习	132
6.7 拓展训练项目	133
6.7.1 安装 MySQL 数据库和 Python 连接数据库	133
6.7.2 使用 Python 实现 MySQL 增查改删	133
6.7.3 Python 文件的基本操作	133
6.7.4 Python 文件目录的基本操作	134

第 7 章

Python 爬虫基础	135
7.1 网络爬虫概述及其结构	135
7.1.1 网络爬虫概述	135
7.1.2 网络爬虫结构	136
7.2 urllib 库	137
7.2.1 urllib.request 模块	137
7.2.2 urllib.parse 模块	138
7.2.3 urllib.error 模块	140
7.3 使用 urllib 爬取网页	141
7.4 浏览器的模拟与实战	142
7.5 正则表达式	143
7.6 图片爬虫实战	147
7.7 项目训练：用 urllib 库爬取百度贴吧	148
7.8 本章小结	152
7.9 练习	152
7.10 拓展训练项目	153
7.10.1 urllib 库的使用	153
7.10.2 百度贴吧网页爬虫	153
7.10.3 淘宝网站图片爬虫	153

第 8 章

Python 爬虫框架	154
8.1 常见爬虫框架	154
8.2 Scrapy 爬虫框架的安装	155
8.3 Scrapy 爬虫框架简介	156
8.4 Scrapy 常用工具命令	157
8.4.1 创建一个 Scrapy 项目	157
8.4.2 Scrapy 全局命令	158
8.4.3 Scrapy 项目命令	160
8.5 Scrapy 爬虫实战	161
8.6 项目训练：用 Scrapy 爬取豆瓣图书	167
8.7 本章小结	171
8.8 练习	171
8.9 拓展训练项目	171
8.9.1 Scrapy 框架的安装及使用	171
8.9.2 Scrapy 命令行工具	172

第 9 章

数据分析基础	173
9.1 numpy 模块	173
9.1.1 ndarray 类型数组	174
9.1.2 matrix 类型矩阵	182
9.1.3 matrix 类型和 array 类型的区别	189
9.2 pandas 模块	193
9.2.1 pandas 模块基础	193
9.2.2 pandas 模块数据清洗	199
9.2.3 pandas 模块数据预处理	221
9.2.4 pandas 模块数据提取	230
9.2.5 pandas 模块数据筛选	234
9.2.6 pandas 模块数据汇总	235
9.2.7 pandas 模块数据统计	237
9.2.8 pandas 模块综合应用示例	239
9.3 项目训练：清洗和预处理 8.6 节中爬取的 doubanread.csv 文件	245
9.4 本章小结	247
9.5 练习	247
9.6 拓展训练项目	249

9.6.1 pandas 基本功能实验	249
9.6.2 pandas 汇总和计算实验	249
9.6.3 pandas 缺失数据处理	249
9.6.4 pandas 构建层次化索引	249

第 10 章

pandas 数据分析	251
10.1 pandas 文件读写基础	251
10.1.1 CSV 文件的读写	251
10.1.2 Excel 文件的读写	254
10.2 pandas 与 MySQL 数据库的交互	256
10.2.1 pandas 与 MySQL 连接的步骤	256
10.2.2 pandas 与 MySQL 交互	257
10.3 pandas 字符串处理	259
10.4 pandas 数据分组与聚合	265
10.4.1 使用内置的聚合函数进行聚合运算	265
10.4.2 分组与聚合过程	267
10.4.3 agg()和 apply()聚合函数	268
10.5 项目训练：电影数据统计	271
10.6 本章小结	274
10.7 练习	274
10.8 拓展训练项目	274
10.8.1 pandas 文件读写	274
10.8.2 pandas 数据库读写	275
10.8.3 pandas 数据处理	275
10.8.4 pandas 数据聚合和组迭代	275

第 11 章

Python 可视化与可视化工具	276
11.1 Python 可视化与可视化工具介绍	276
11.2 pandas 基本图形绘制	278
11.2.1 折线图	278
11.2.2 柱状图	282
11.2.3 直方图	285
11.2.4 散点图	285
11.2.5 面积图	287

11.2.6 饼图	287
11.2.7 密度图	290
11.3 matplotlib 绘图	291
11.3.1 matplotlib 绘图基础	291
11.3.2 matplotlib 交互绘图	291
11.4 matplotlib.pyplot 的使用	294
11.4.1 pyplot 绘图基础	294
11.4.2 多种类型图的绘制	297
11.5 项目训练：电影数据信息分析	306
11.6 本章小结	310
11.7 练习	310
11.8 拓展训练项目	310
11.8.1 pandas 绘图	310
11.8.2 matplotlib 交互式绘图实践	310
11.8.3 pyplot 绘图元素的设置	311
11.8.4 子图的绘制	311

附录	312
----------	-----

参考文献	316
------------	-----

第 8 章 数据可视化	317
8.1 本章学习目标	317
8.2 数据可视化	317
8.2.1 数据可视化的概念	317
8.2.2 数据可视化的重要性	317
8.2.3 数据可视化的方法	317
8.2.4 数据可视化的工具	317
8.2.5 数据可视化的应用	317
8.3 数据可视化的基础	317
8.3.1 数据可视化的基础概念	317
8.3.2 数据可视化的基础方法	317
8.3.3 数据可视化的基础工具	317
8.3.4 数据可视化的基础应用	317
8.4 数据可视化的进阶	317
8.4.1 数据可视化的进阶概念	317
8.4.2 数据可视化的进阶方法	317
8.4.3 数据可视化的进阶工具	317
8.4.4 数据可视化的进阶应用	317
8.5 数据可视化的案例	317
8.5.1 数据可视化的案例概念	317
8.5.2 数据可视化的案例方法	317
8.5.3 数据可视化的案例工具	317
8.5.4 数据可视化的案例应用	317
8.6 数据可视化的总结	317
8.7 本章小结	317
8.8 练习	317
8.9 拓展训练项目	317

第1章

Python环境搭建

01

学习目标

- ① 了解 Python 的发展情况;
- ② 熟悉 Python 的官网内容, 会下载 Python 的安装程序;
- ③ 会使用 Linux 的基本操作命令安装 Python 3.7.0;
- ④ 理解 Python 的交互模式, 会在交互模式下编写简单的程序;
- ⑤ 了解 iPython 和 PyCharm 的功能, 并基本掌握安装和使用方法。

Python 是 Guido van Rossum (吉多·范罗苏姆) 于 1989 年年底开始开发的, 第一个版本发行于 1991 年, 比 Java 早 4 年。Python 推出后不久就迅速得到了各行业人士的青睐, 经过 20 多年的发展, Python 已经渗透到计算机科学与技术、统计分析、移动终端开发、科学计算与可视化、逆向工程与软件分析、图形图像处理、人工智能、游戏设计与策划、网站开发、数据提取与大数据处理、密码学、系统运维、音乐处理、计算机辅助教育、医药辅助设计、天文信息处理、化学、生物、电子电路设计等专业和领域, 在黑客领域更是一直拥有霸主地位。Python 连续多年在 TIOBE 网站编程语言排行榜中排名第 7、8 位, 2019 年 5 月排名上升至第 4, 可见其发展之迅速。

Python 是一种跨平台、开源、免费的解释型高级动态编程语言。它支持命令式编程、面向对象编程和函数式编程, 包含了完善且易于理解的标准库, 还有非常丰富的第三方开源库, 使用户能轻松地完成开发任务。

1.1 Python 版本概述及下载 Python 安装文件

除机器语言外的任何一种计算机语言都需要有相应的编译器或者解释器, Python 也不例外。“工欲善其事, 必先利其器”, 学习 Python 的第一件事就是下载及安装相关的程序, 准备好工具才能顺利进行学习。

1.1.1 Python 版本概述

Python 官网目前同时发行 Python 2.X 和 Python 3.X 两个不同的版本 (其下载地址为 <https://www.python.org/downloads/source/>), 如图 1-1 所示 (2018 年 10 月 1 日界面)。这两个版本互不兼容, 除了输入及输出方式有所不同外, 很多内置函数的实现和使用方式也有较大的区别。Python 3.X 对 Python 2.X 的标准库也进行了一定程度的重新拆分和整合。对于初学者来说, 在选择使用 Python 2.X 还是 Python 3.X 时, 一定要考虑清楚学习 Python 的目的是什么, 打算做哪方面的开发, 有哪些扩展库可用, 这些扩展库最高支持哪个版本的 Python 等。确定了这些内容后, 再做出选择, 才能事半功倍,

而不至于在版本的选择上纠结。另外还要注意，较新的版本推出后，不要急于更新和替换，而应该在自己必须使用的扩展库更新之后再行 Python 的更新。

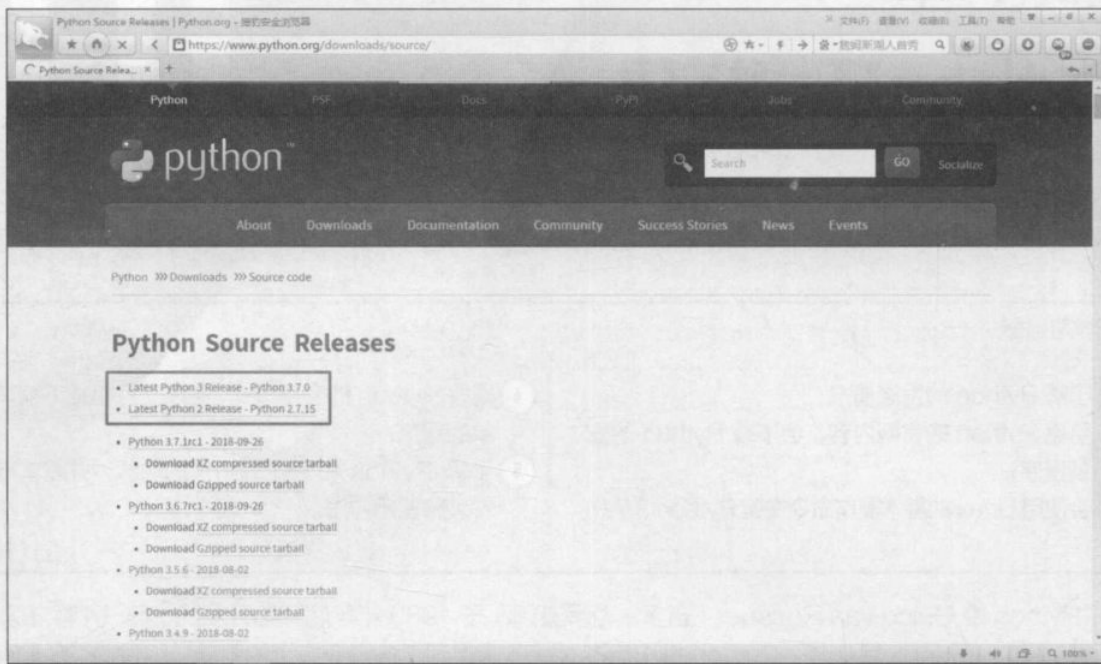


图 1-1 Python 官网下载页

目前看来，Python 3.X 的使用是大势所趋，如果读者是初学者，还没有明确应用领域，建议选择 Python 3.X 系列的最高版本。

1.1.2 下载 Python 安装文件

1. Linux 系统下的 Python 安装文件下载

Python 的官网地址是 <http://www.python.org>，在 Ubuntu 16.04 的虚拟机系统中，使用 Firefox 浏览器打开 Python 官网，单击“Downloads”菜单，展开子菜单，选择“Source code”选项，进入图 1-1 的下载页面。单击图 1-1 中的“Download Gzipped source tarball”选项，出现如图 1-2 所示的下载提示，单击“保存文件”按钮，开始进行下载。



图 1-2 下载提示

下载过程中，为了观察下载的进度，可以单击界面右上角的  按钮，将出现如图 1-3 所示的下载列表和下载进度指示。

下载的文件 Python-3.7.0.tgz 默认存放在“/home/用户名/下载”文件夹下。“用户名”是登录 Linux 的用户名，如“ding”，因此，下载的文件存放在“/home/ding/下载”文件夹中。下载后也可单击图 1-3 中的下载文件名，打开“归档管理器”，如图 1-4 所示。

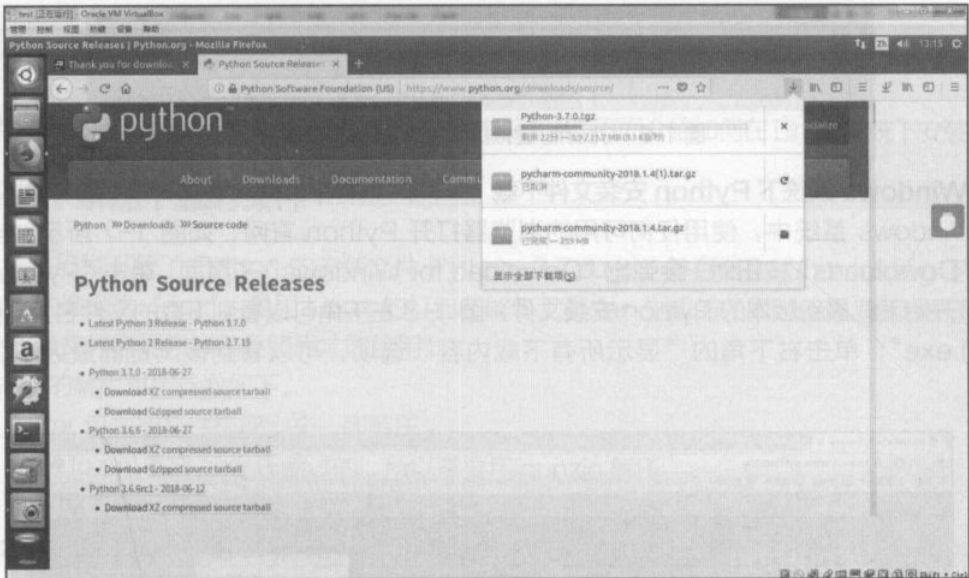


图 1-3 下载列表及下载进度



图 1-4 归档管理器

单击图 1-4 中的 **+** 按钮，打开如图 1-5 所示的“添加文件”窗口，在“位置”框中可以看到所下载文件的路径。

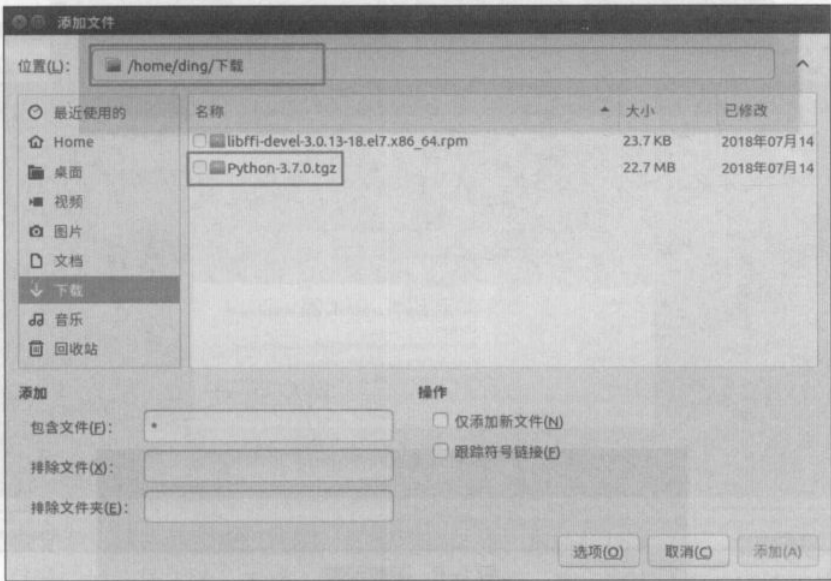


图 1-5 “添加文件”窗口

至此，Linux 操作系统下的 Python 3.7.0 安装文件下载完毕。用户可以在 Ubuntu 虚拟机的终端依次输入“cd /home/ding/下载”和“ls”命令，查看 Python-3.7.0.tgz 文件，如图 1-6 所示。



图 1-6 Ubuntu 虚拟机终端查看到的下载文件

2. Windows 系统下 Python 安装文件下载

在 Windows 系统中，使用任何可用的浏览器打开 Python 官网，如图 1-7 所示，当鼠标指针指向“Downloads”按钮时，会弹出“Download for Windows”子页面。单击“Python 3.7.0”按钮即可开始下载最新版本的 Python 安装文件。图 1-8 左下角可以看到下载的文件名为“python-3.7.0(1).exe”；单击右下角的“显示所有下载内容”选项，可以看到该浏览器最近下载的文件列表。

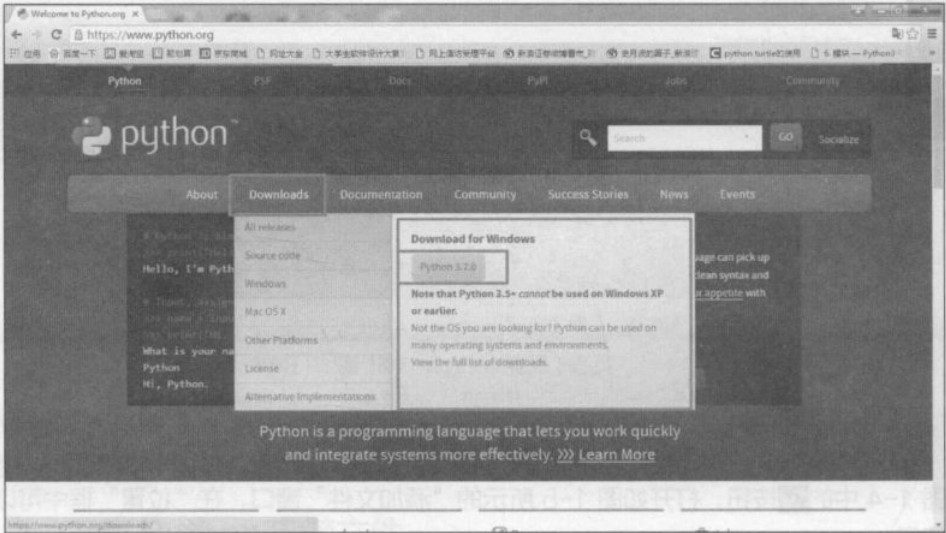


图 1-7 在 Windows 系统中打开 Python 官网

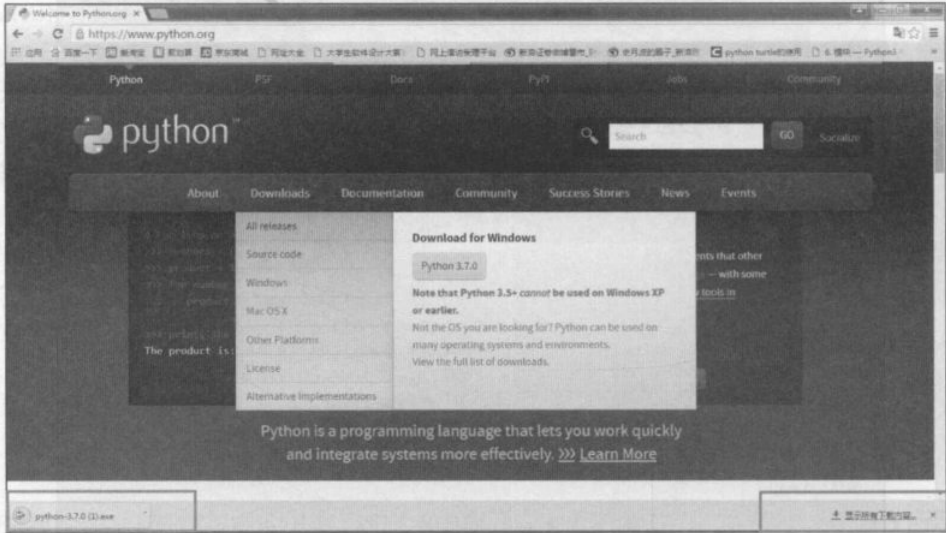


图 1-8 下载状态

Windows 中下载文件名实际为“python-3.7.0.exe”，名称中出现的“(1)”表示的是本机已经下载过该文件，为了不重名而加上数字序号。