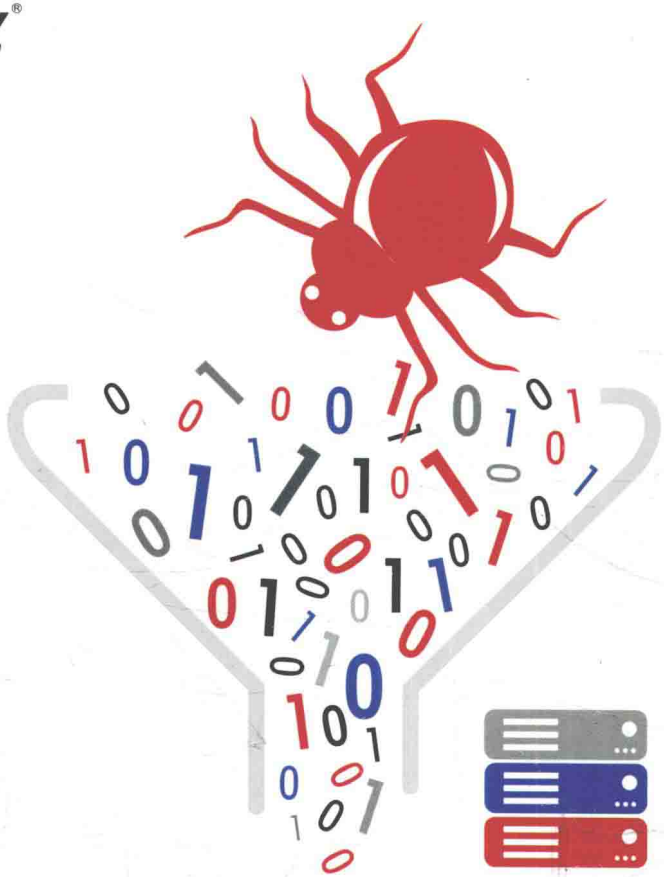
Broadview®
www.broadview.com.cn



Python 3

爬虫、数据清洗与可视化实战

零一 韩要宾 黄园园 著

Python技术的入门读物

通过实战教初学者学习爬取数据、清洗和组织数据进行分析和可视化

适合Python初学者、爱好者及高等院校的相关专业学生学习使用

中国工信出版集团

電子工業出版社
PUBLISHING HOUSE OF ELECTRONICS INDUSTRY
http://www.phei.com.cn

Python 3

爬虫、数据清洗与可视化实战

零一　韩要宾　黄园园　著

電子工業出版社
Publishing House of Electronics Industry
北京•BEIJING

内 容 简 介

本书是一本通过实战教初学者学习采集数据、清洗和组织数据进行分析及可视化的 Python 读物。书中案例均经过实战检验，笔者在实践过程中深感采集数据、清洗和组织数据的重要性，作为一名数据行业的“码农”，数据就是沃土，没有数据，我们将无田可耕。

本书共分 11 章，6 个核心主题：其一是 Python 基础入门，包括环境配置、基本操作、数据类型、语句和函数；其二是 Python 爬虫的构建，包括网页结构解析、爬虫流程设计、代码优化、效率优化、容错处理、反防爬虫、表单交互和模拟页面点击；其三是 Python 数据库应用，包括 MongoDB、MySQL 在 Python 中的连接与应用；其四是数据清洗和组织，包括 NumPy 数组知识、pandas 数据的读写、分组变形、缺失值异常值处理、时序数据处理和正则表达式的使用；其五是综合应用案例，帮助读者贯穿爬虫、数据清洗与组织的过程；最后是数据可视化，包括 Matplotlib 和 Pyecharts 两个库的使用，涉及饼图、柱形图、线图、词云图、地图等图形，帮助读者进入可视化的殿堂。

本书以实战为主，适合 Python 初学者及高等院校的相关专业学生，也适合 Python 培训机构作为实验教材使用。

未经许可，不得以任何方式复制或抄袭本书之部分或全部内容。
版权所有，侵权必究。

图书在版编目（CIP）数据

Python 3 爬虫、数据清洗与可视化实战 / 零一，韩要宾，黄园园著. —北京：电子工业出版社，2018.3
ISBN 978-7-121-33359-0

Ⅰ. ①P… Ⅱ. ①零… ②韩… ③黄… 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(2017)第 321885 号

策划编辑：张慧敏
责任编辑：牛　勇
印　　刷：三河市良远印务有限公司
装　　订：三河市良远印务有限公司
出版发行：电子工业出版社
　　北京市海淀区万寿路 173 信箱　　邮编：100036
开　　本：787×980　1/16　印张：13.25　字数：200 千字
版　　次：2018 年 3 月第 1 版
印　　次：2018 年 3 月第 1 次印刷
印　　数：3000 册　　定价：49.00 元

凡所购买电子工业出版社图书有缺损问题，请向购买书店调换。若书店售缺，请与本社发行部联系，联系及邮购电话：（010）88254888，88258888。

质量投诉请发邮件至 zlts@phei.com.cn，盗版侵权举报请发邮件至 dbqq@phei.com.cn。

本书咨询联系方式：010-51260888-819，faq@phei.com.cn。

Python 是一种解释型、面向对象的动态数据类型高级程序设计语言。从 20 世纪 90 年代初 Python 语言诞生至今，它逐渐被广泛应用于各个领域，比如桌面应用开发、游戏开发、Web 开发、网络爬虫、数据统计分析、自然语言处理、机器学习、深度学习、人工智能等。由于 Python 语言的简捷性、易读性及可扩展性，在国外用 Python 进行科学计算的研究机构日益增多。一些知名大学甚至采用 Python 语言教授程序设计课程，例如麻省理工学院的计算机科学及编程导论课程。

在数据科学领域，Python 的发展势头十分迅猛。一方面，Python 拥有各种开源的网络爬虫框架，可以帮助数据科学家快速收集数据；另一方面，Python 在机器学习和深度学习方面有很多成熟的拓展包，可以帮助数据科学家完成各类数据分析任务，无论是简单的线性回归，还是复杂的深度学习网络构建。

Stack Overflow 最新调查显示，Python 已经成为快速发展的主流编程语言，也是高收入国家网民访问 Stack Overflow 网站过程中，点击量最高的标签。由此可见，Python 将必成为各个领域的程序员需要掌握的技能之一。

黄志洪

著名数据分析网站炼数成金创始人

我最早是通过《电商数据分析——淘宝实战》一书接触零一的。在该书中，他用相当多的实务案例来告诉读者如何在电商的环境中，用数据做决策，从实践中学知识，令我印象深刻。后来通过 CDA 的活动认识了零一，发现他是一名数据分析爱好者。在跟他的谈话中，提到的更多的是实务的应用，令我钦佩不已。

人工智能的先驱者吴恩达曾说过，一家人工智能的公司必须具备三种能力：其一是有策略的数据采集，其二是集中式的数据仓库及统一的数据分析平台，最后是无所不在的自动化应用。零一的这本书就是教你如何系统化地采集数据、储存数据及应用数据。

这本书教大家如何利用 Python 撰写爬虫程序、清洗和组织数据、解析网页的内容，并将数据储存于数据库中。本书巨细无遗，帮助大家节省时间，是值得一读的好书！

李御玺（Yue-Shi Lee）

台湾大学资讯工程博士

铭传大学资讯工程学系教授

前　言

Python 是军刀型的开源工具，被广泛应用于 Web 开发、爬虫、数据清洗、自然语言处理、机器学习和人工智能等方面，而且 Python 的语法简洁易读，这让许多编程入门者不再望而却步，因此 Python 在最近几年非常受欢迎，各行各业的技术人员都开始使用 Python。

本书内容来自笔者在高校授课的内容，主要介绍如何运用 Python 工具获取电商平台的页面数据，并对数据进行清洗和存储。本书简化了 Python 基础部分，保证有足够的篇幅来介绍爬虫和数据清洗的内容。

本书采用的版本是 Python 3.6.2，是笔者写书时的最新版本，而且笔者习惯用的操作平台是 Windows 系统。虽然目前一些高校和开发者在使用 Python 2.7，但是 Python 团队将在 2020 年停止对 Python 2.7 的支持更新，Python 2.X 转向 Python 3.X 是大势所趋。

本书第 1 章简单介绍 Python 和相关的 IDE，如果读者完全没有 Python 基础，那么建议选购一本基础书作为辅助。第 2 ~ 6 章介绍爬虫的实例，实现从最简单的爬虫到相对比较复杂的爬虫。鉴于实例的限制，本书的爬虫内容没有涉及代理服务器和验证码处理等问题。第 7 章介绍在 Python 中如何连接并操作数据库。第 8 章介绍了 NumPy 及其用法。第 9 章详细介绍 pandas 的功能，pandas 是 Python 数据清洗和建模中非常重要的库。第 10 章用两个完整案例展示了从爬虫到建模的过程。第 11 章介绍 Python 的可视化，选用的库是 matplotlib 和 pyecharts，这里详细介绍了 pyecharts。

鉴于作者的水平有限，不足之处请读者不吝指教。

轻松注册成为博文视点社区用户（www.broadview.com.cn），扫码直达本书页面。

- **下载资源**：本书如提供示例代码及资源文件，均可在 下载资源 处下载。
- **提交勘误**：您对书中内容的修改意见可在 提交勘误 处提交，若被采纳，将获赠博文视点社区积分（在您购买电子书时，积分可用来抵扣相应金额）。
- **交流互动**：在页面下方 读者评论 处留下您的疑问或观点，与我们和其他读者一同学习交流。

页面入口：*http://www.broadview.com.cn/33359*

试读结束：需要全本请在线购买：　www.ertongbook.com

目　录

第 1 章

Python 基础

1.1 安装 Python 环境

1.1.1 Python 3.6.2 安装与配置

根据 Windows 版本（64 位/32 位）从 Python 官网安装下载对应的版本，如图 1-1 所示。

官方下载网址：https://www.python.org/

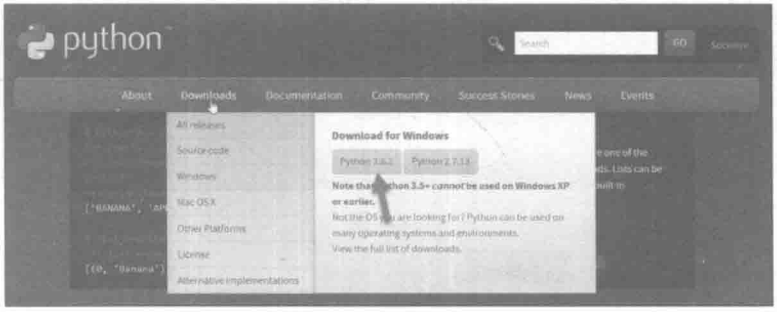


图 1-1

下载完成后，双击文件以运行安装程序安装 Python，如图 1-2 所示。

图 1-2

STEP 1：勾选“Add Python 3.6 to PATH”选项后单击“Customize installation”选项。

这个选项用于将 Python 3.6 加入系统路径，勾选该选项会使日后的操作非常方便；如果没有勾选这个选项就需要手动为系统的环境变量添加路径。

STEP 2：在弹出的选项卡中勾选所有的选项，并单击“Next”按钮，如图 1-3 所示。

选项“Documentation”表示安装 Python 的帮助文档；选项“pip”表示安装 Python 的第三方包管理工具；选项“tcl/tk and IDLE”表示安装 Python 的集成开发环境；选项“Python test suite ”表示安装 Python 的标准测试套件，后两个选项则表示允许版本更新。

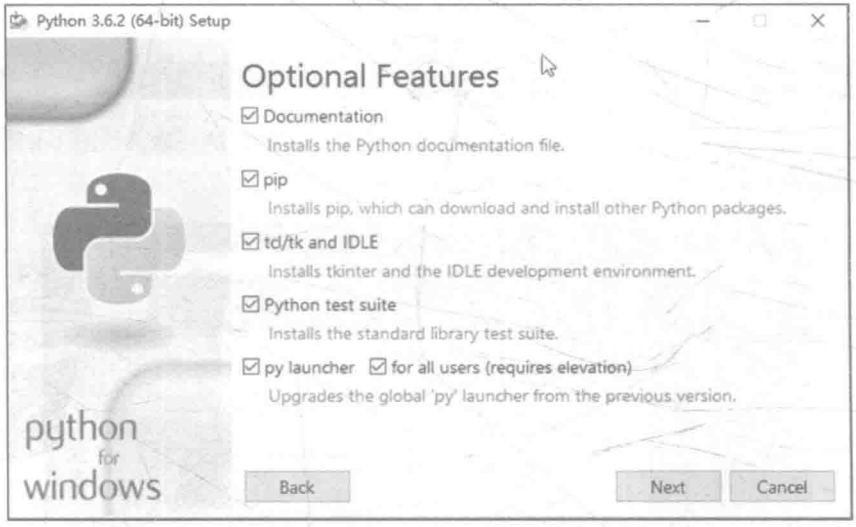


图 1-3

试读结束：需要全本请在线购买： www.ertongbook.com

STEP3：保持默认勾选状态，单击“Browse”按钮，选择安装路径，如图 1-4 所示。

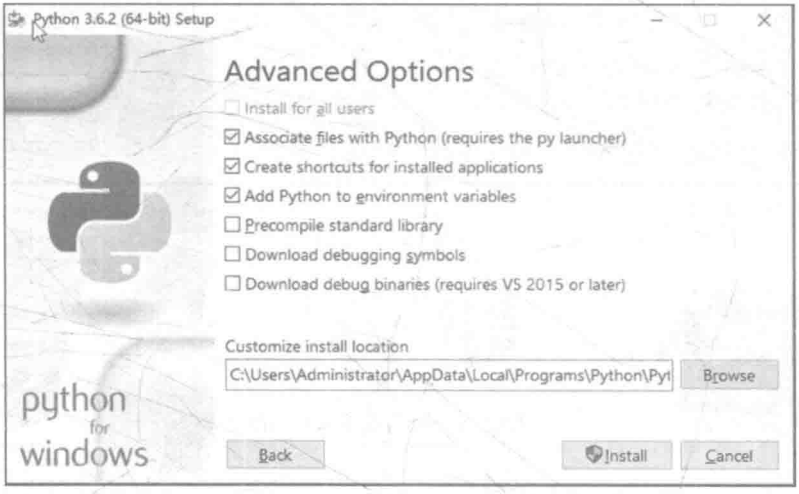


图 1-4

STEP4：单击“Install”按钮，直至完成安装。

安装好后，调出命令提示符，输入“python”，检查是否安装成功。如果 Python 安装成功，将出现如图 1-5 所示的界面，即输入“python”后，会看到“>>>”符号。

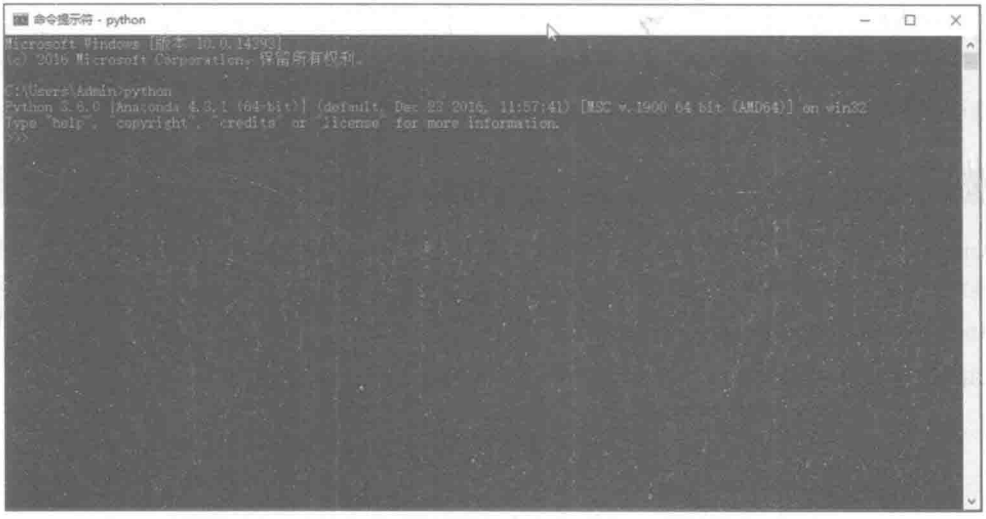


图 1-5

1.1.2 使用 IDE 工具——PyCharm

安装好环境后，还需要配置一个程序员专属工具，即 PyCharm，它是一个适合用于开发的多功能 IDE（集成开发环境），下载社区版（免费版）。

笔者使用的版本是 2017.2.2，发行日期是 2017 年 8 月 24 日，下载地址如下（参见图 1-6）。

http://www.jetbrains.com/pycharm/download/#section=windows

图 1-6

PyCharm 非常好用，通过 PyCharm 可以下载、安装和管理库。

1.1.3 使用 IDE 工具——Anaconda

Anaconda 是一个专门用于统计和机器学习的 IDE，它集成了 Python 和许多基础的库，如果业务场景是统计和机器学习，那么只要安装一个 Anaconda 就可以了，这样省去许多复杂的配置过程。

Anaconda 的官方下载地址如下（参见图 1-7）。

https://www.anaconda.com/download/

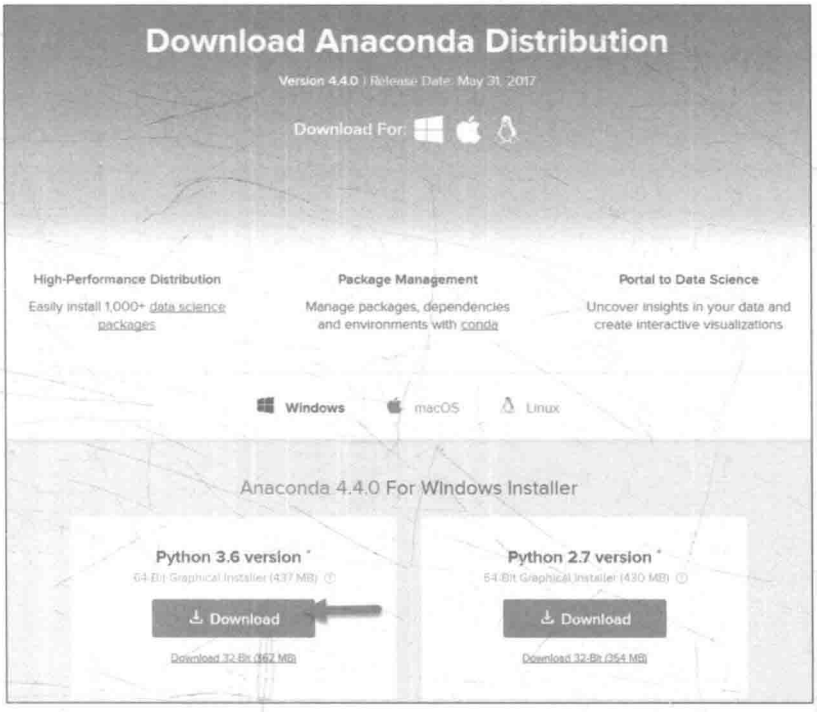


图 1-7

这里默认下载的是 64 位的版本，如果需要 32 位的版本，那么可以单击“Download”按钮下的文字链接。

使用 Anaconda 不需要提前安装 Python，安装后即可运行：通过快捷键【 Win+R 】调用运行窗口，输入“ipython jupyter”，然后单击“确定”按钮（参见图 1-8）。

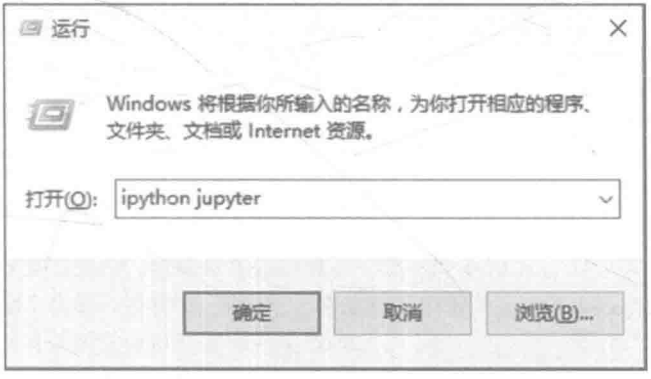


图 1-8

1.2 Python 操作入门

1.2.1 编写第一个 Python 代码

运行 PyCharm 后，需要先新建计划，单击“Create New Project”选项（参见图 1-9）。

图 1-9

设置 Location（路径）和 Interpreter（翻译器），笔者同时安装了 Python 和 Anaconda，所以图 1-10 中的翻译器有两个可选项，二者的区别在于 Anaconda 中有许多预置好的库，不用再配置库了。这里选择 Python 原版的翻译器，然后单击右下角的“Create”按钮。

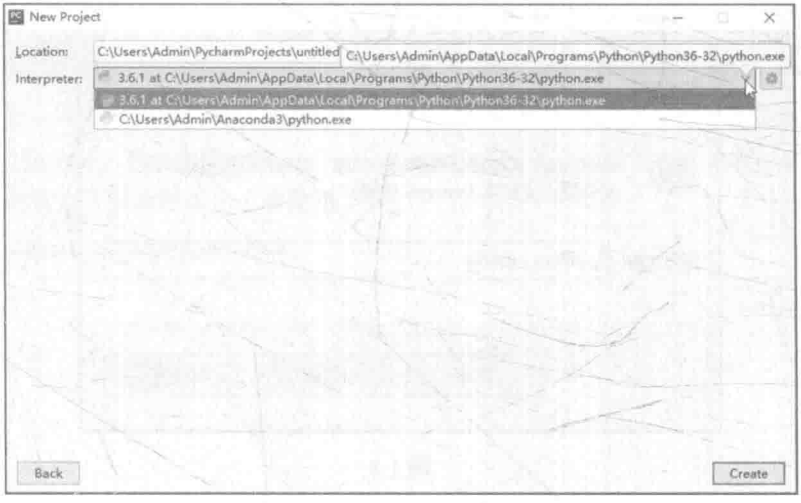


图 1-10

新建 Project（计划）后，在左侧的项目窗口，右击鼠标，在快捷菜单中选择“New”→“Python File”命令，新建 Python 文件（参见图 1-11）。

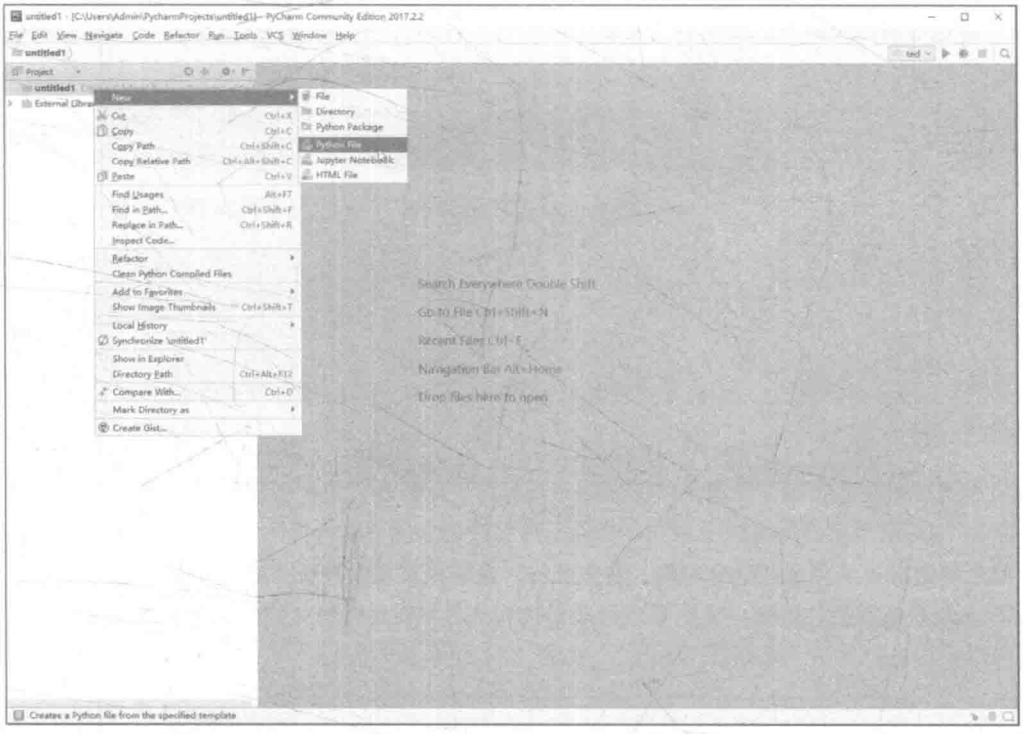


图 1-11

设置 Name（文件名），然后单击右下角的“OK”按钮（参见图 1-12）。

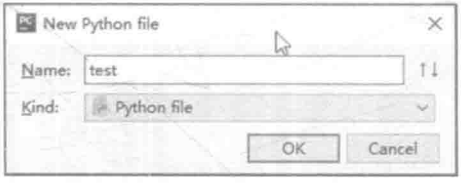


图 1-12

新建文件后，右侧的空白区域就是代码编辑区（参见图 1-13）。

从“Hello World（你好，世界）”开始吧！在编辑区中输入print('Hello, World!')，print()是一个打印函数，表示将括号中的文本打印在即时窗口中。

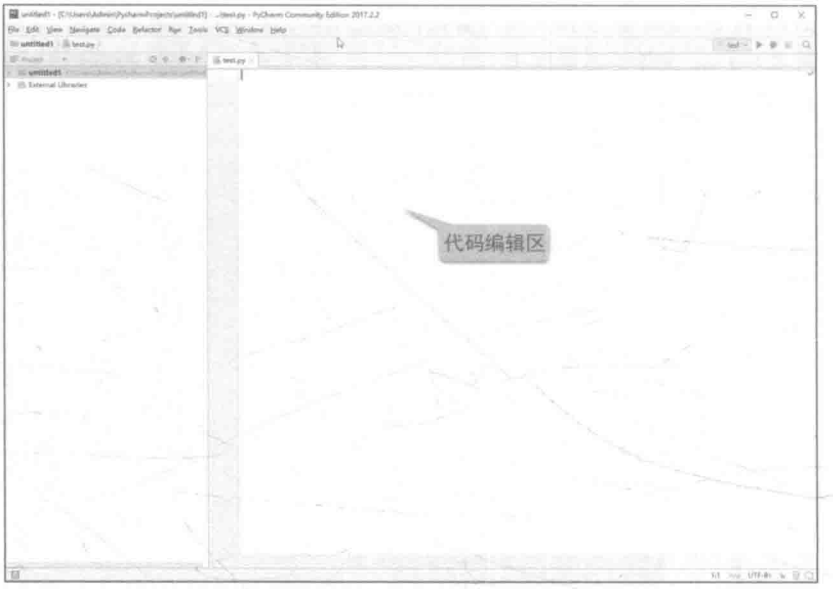


图 1-13

然后将鼠标光标停留在括号右侧，右击鼠标，在快捷菜单中选择“Run ‘test’”命令，其中单引号中的 test 是当前的文件名，一定要注意运行的文件名和要运行的文件名保持一致。运行后可以观察到即时窗口中打印出“Hello，World!”，如图 1-14 所示。

图 1-14

试读结束：需要全本请在线购买： www.ertongbook.com