项目1

基础开发环境的安装与配置

项目目标

- 掌握 Anaconda 开发环境的安装方法。
- 学会 jupyter notebook 的基本使用方法。
- 了解数据采集所具备的 Python 基础知识。

本项目旨在为让读者掌握 Python 开发环境的搭建技能，学会对 Anaconda3 进行安装和配置是初学者必须掌握的技能。

项目描述

项目开发环境的搭建是学习一门编程语言的开始，如使用 C 语言开发应用，需要安装 C 语言编译器，将编程语言编译成可执行的文件，然后运行。本书使用的 Python 语言是一门解释型语言，在学习阶段，由于不必考虑性能问题，所有无须编译，安装 Python 解释器即可执行。

在数据科学领域，最佳的开发环境解决方案是 Anaconda3，它让数据科学的开发过程变得更加简单。Anaconda3 不仅包含了 Python 解释器，还包含 1000 多个开源库，以及包管理工具 conda，同时，内置了基于网页的、用于交互计算的应用程序 jupyter notebook，它可以在网页页面中直接编写代码和运行代码。

项目实施

（1）安装 Python 基础环境。

（2）安装 Anaconda3，配置环境变量。

（3）启动 jupyter notebook，运行 Python 入门程序“Hello，world.”。

课程思政要求

本项目的思政要求是让读者通过掌握互联网数据采集开发环境的搭建下技能，掌握先进信息技术工具的使用，培养其实际动手能力和精益求精、严谨治学的态度，为未来走向工作岗位打好基础。

知识链接

1. 大数据与互联网数据采集概述

我们处在信息爆炸的时代，过去几年中，全球产生的数据量超过了过去几十年数据量的总和。根据 IDC（International data Corporation，国际数据公司）近期做出的估测，数据在以每年 50% 的速度增长，也就是每两年多就增长一倍，这也成了大数据的摩尔定律。

数据的种类繁多，主要由结构化数据、非结构化数据和半结构化数据组成。约 10% 的结构化数据存储在数据库中，其余 90% 的非结构化数据与半结构化数据，分布在人类的工作、学习、生活中。

数据主要来源于以下几个方面。

1）科学研究产生的数据

- 基因组。
- LHC（Large hadron collider，大型强子对撞机）加速器。
- 地球与空间探测。

2）企业应用产生的数据

- E-mail、文档、文件。
- 应用日志。
- 交易记录。
- 银行、基金、股票、期货等金融数据。

3）Web 1.0 数据

- 文本数据。
- 图像数据。
- 视频数据。

4）Web 2.0 数据

- 查询日志、点击流产生的数据。
- 微博、贴吧、社交网络上产生的数据。
- 百度百科、维基百科。

• 抖音、快手、视频号等短视频数据。

5）Web 3.0 数据

• 数字货币、虚拟货币等数据。

• 区块链产生的数据。

• 元宇宙产生的数据。

• DAO、NFT 等数据。

大数据有四个特征，也称 4V 特征，分别为数据量大（volume）、种类多（variety）、高速处理（velocity）和价值密度低（value）。以前，由于成本的限制，如此繁多的数据无法长期保存。如今，存储设备的成本在逐渐降低，之前价值不高的数据也开始被保存、挖掘。大数据为政府、科研、企业、个人等贡献了巨大的价值。

对大数据的研究，主要分为数据采集、数据预处理、数据存储、数据分析与挖掘等几个方面。其中，数据采集是大数据分析与挖掘项目的基础，它又细分为互联网应用数据采集、数据库数据采集、日志采集等。互联网应用数据采集是初级数据采集从业人员的基本技能，也是本书重点研究的对象。

2. 互联网数据采集的主要技术

互联网数据采集技术主要有以下两种。

（1）使用软件工具进行数据采集，如“八爪鱼采集器”“后羿采集器”等。这些采集器大多数的原理为通过抓包技术抓取 HTTP 请求，然后通过正则表达式获取关键信息，部分软件工具还可以模拟单击按钮查看隐藏内容等。

使用软件工具的优点：简单，不需要会使用编程语言；可以直接使用别人做好的采集流程（大多数为收费）来采集特定网站的信息。

使用软件工具的缺点：依赖使用环境，如 Linux 环境下无法使用；操作人工化，自由度低；采集大数据集时有性能缺陷；无法通过编写代码的方式与数据分析挖掘、数据存储等工作进行实时对接。

（2）通过编写代码进行数据采集，编程语言可以使用 Java、Python 等。其中，Python 语言简单易学，与后期的数据预处理部分所使用的 Numpy、Pandas 等第三方库可以进行无缝衔接，是最佳的选择。

使用 Python 语言编写代码采集数据是本书的主要内容。要想成为一名专业的数据采集人员，最基本的编程知识是要掌握的，如列表和字典的基本使用等。

本书使用 Anaconda3 作为 Python 开发的基础环境。Anaconda 是一个开源的 Python 发行版本，在全球拥有超过 2500 万用户。使用 Anaconda 开源个人版（分发版）是在单台机器上执行 Python/R 数据科学和机器学习的最简单方法。Anaconda 是为独立开发者开发的工具包，包含了数千个开源包和库，免去了独立安装 Python 基础环境，再依次安装包的过程。Anaconda 还包含了一个开源免费的网页编辑代码工具 jupyter notebook，它是数据科学领域的最佳交互式工具。

任务 1.1 通过 Anaconda3 安装基础开发环境

步骤 1　下载 Anaconda3。

打开 Anaconda 官方产品下载网站，如图 1.1 所示。

图 1.1　Anaconda 官网产品下载页面

读者可以根据已有的操作系统，选择单击 Download 按钮下方的 Windows、MacOS 或者 Linux 图标按钮，进入如图 1.2 所示的下载链接页面（本书选择 Windows 操作系统）。

图 1.2　Anaconda 下载链接

步骤 2　安装 Anaconda3。

（1）找到上述操作下载完成的 Anaconda-2022.05-Windows-x86_64.exe 文件，右击安装包，在菜单中选择“以管理员身份运行”标签，打开后出现欢迎界面，单击 Next 按钮，进入下一步，如图 1.3 所示。

图 1.3 单击 Next 按钮

（2）进入安装使用协议界面，单击 I Agree 按钮，同意使用协议，如图 1.4 所示。

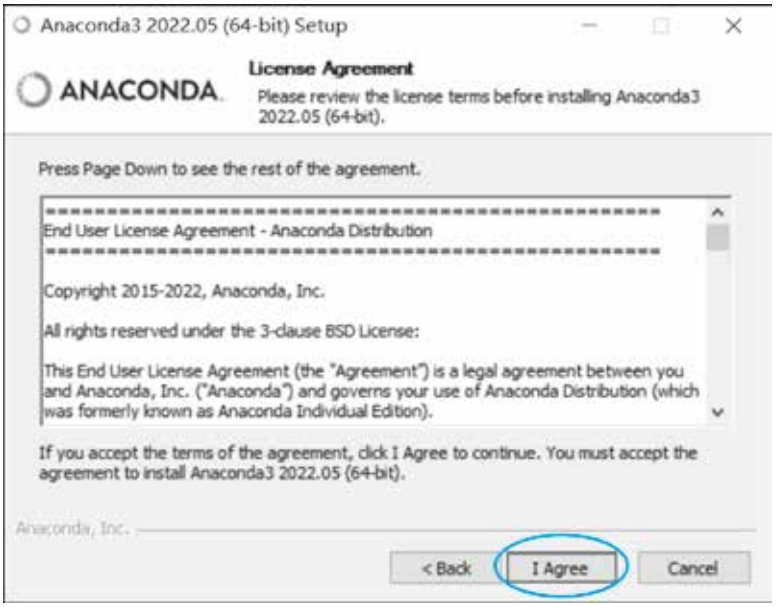

图 1.4 安装使用协议界面

（3）选择安装类型界面有两个选项：一个是仅为当前用户安装，另一个为所有用户安装。选择下方的 All Users（为所有用户安装）后单击 Next 按钮，如图 1.5 所示。

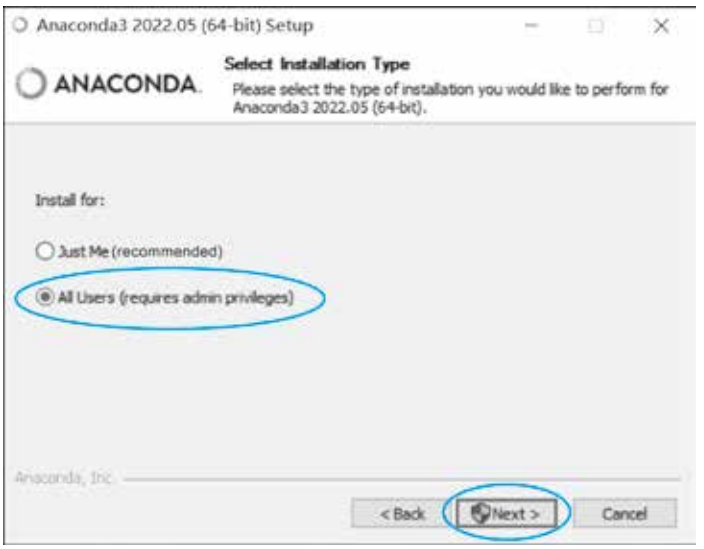

图 1.5　为所有用户安装

（4）这时，Windows 10 用户会弹出“安装程序会更改用户设置，是否同意”的选项，选择“是”按钮。下一步选择用户安装路径，我们选择安装在 C:\ProgramData\Anaconda3 目录下或 D:\ProgramData\Anaconda3 目录下，然后单击 Next 按钮，如图 1.6 所示。

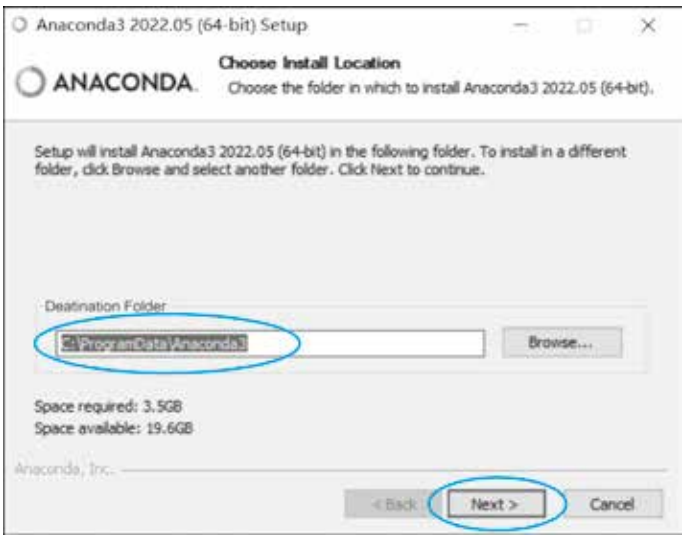

图 1.6　选择安装路径

注意：此时，如果出现警告“Directory ‘C:\Programedata\...’ is not empty.”，说明安装目录不为空，需要修改一下安装路径，如改为 C:\Anaconda3 或 D:\Anaconda3 就可以解决。

（5）选择安装路径后，进入添加环境变量操作环节。如图 1.7 所示，界面中有两个选项，上面的选项是将 Anaconda 添加到系统的环境变量中，下面的选项是将 Anaconda 注册

为系统的 Python 3.9，勾选“Register Anaconda3 as the system Python 3.9”，安装完成再手动添加系统环境变量，单击 Install 按钮。

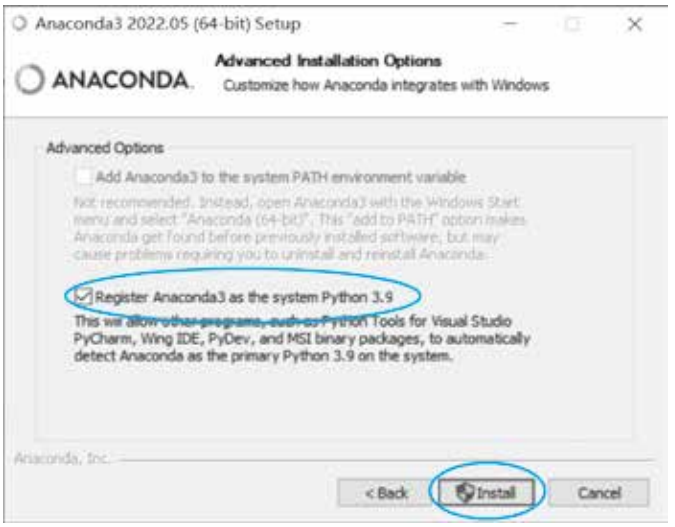

图 1.7　添加环境变量

（6）添加完环境变量后，系统需要花费一段时间安装，如图 1.8 所示。

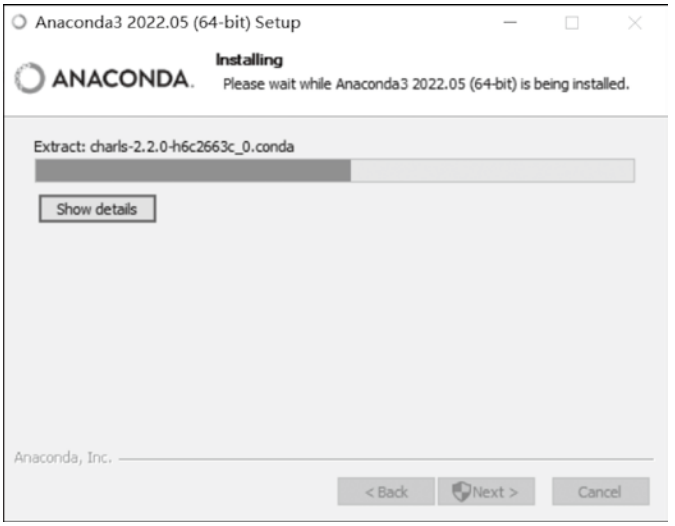

图 1.8　Anaconda 安装中

注意：在安装过程中，如果计算机安装了腾讯电脑管家、360 安全卫士、火绒等安全软件，会弹出窗口，这时切记要选择同意所有操作选项，千万不能选择关闭或者拒绝。

根据计算机配置，安装需要 5～15 分钟，要耐心等待。如图 1.9 所示，当出现 Completed 时，即代表已经安装成功，单击 Next 按钮，进入推荐下载页面。

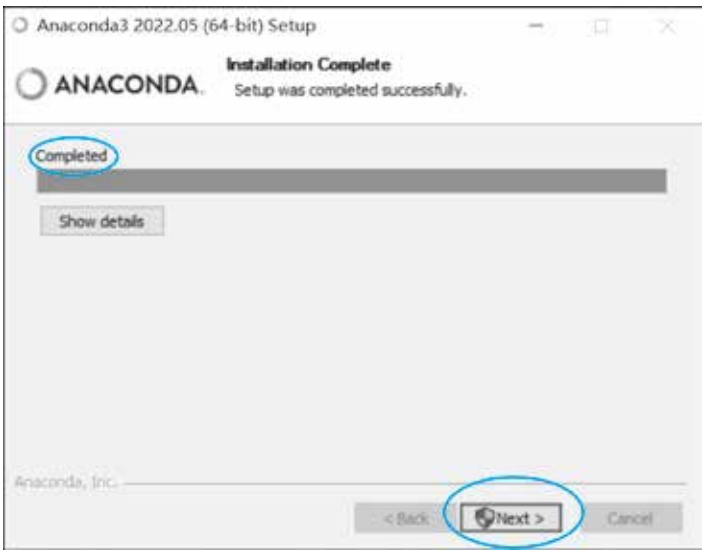

图 1.9　Anaconda 安装成功

（7）Anaconda 推荐下载 DataSpell，它是 Jetbrains 公司的最新 IDE（integrated development enviroment）工具，但本书使用 jupyter notebook 作为编辑工具，故忽略该步骤，直接单击 Next 按钮进入下一步。

（8）完成安装页面，取消勾选“Anacoda Individual Edition Tutorial”和“Getting Started with Anaconda”两个选项，单击 Finish 按钮，完成安装，如图 1.10 所示。

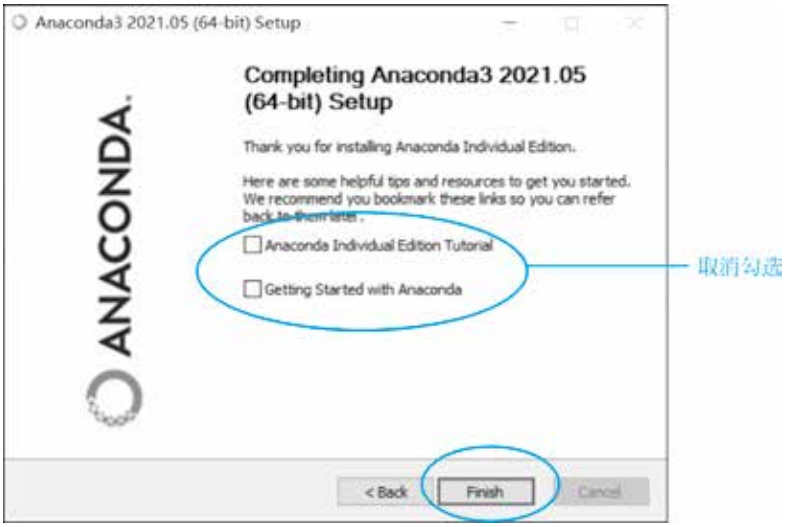

图 1.10　Anaconda 安装完成

（9）添加环境变量。由于安装过程中并没有勾选自动添加环境变量的选项，所以需要将 Anaconda 的安装路径手工添加到 Windows 系统环境变量中。右击 Windows 桌面上“此电脑”图标，在弹出的窗口中选择“属性”选项，在打开的界面中单击左侧“高级系统

设置”标签，在打开的“系统属性”窗口中，单击“高级”选项卡，单击“环境变量”按钮，打开“环境变量”窗口，如图 1.11 所示，向下拖动“系统变量”右侧的滚动条，找到 Path 选项并双击。

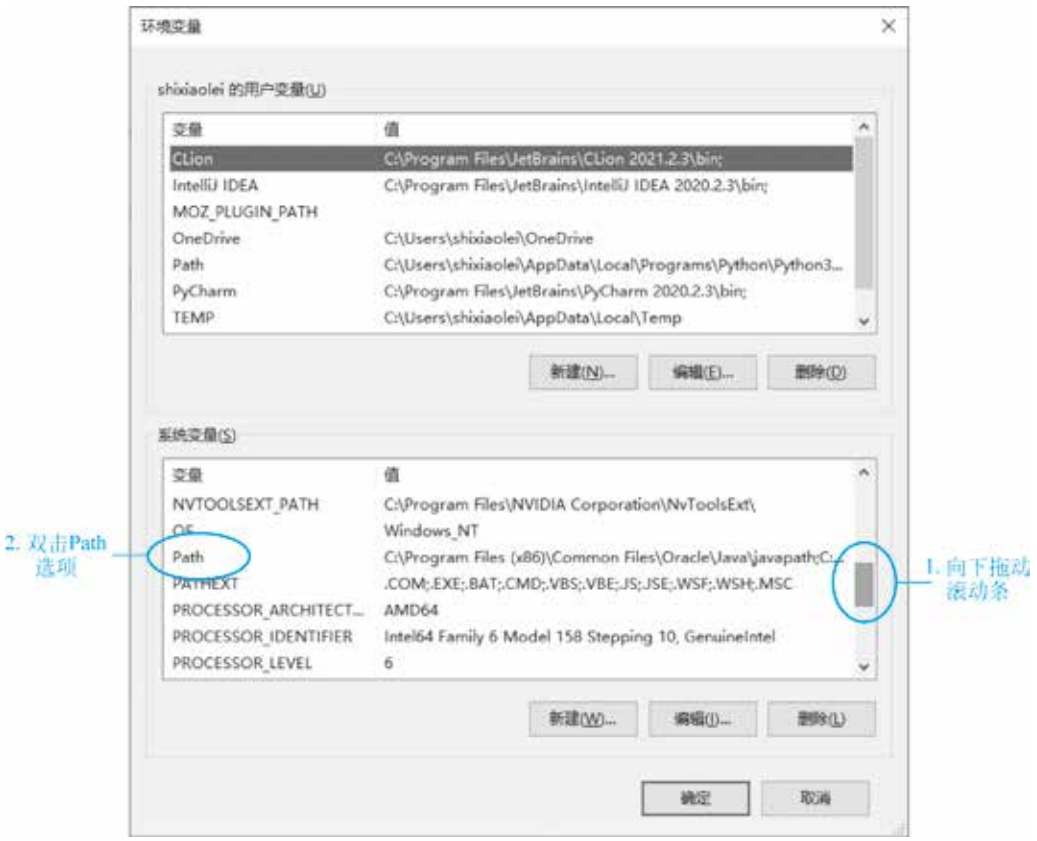

图 1.11　环境变量窗口

（10）打开“编辑环境变量”窗口后，单击“新建”按钮，依次在输入框中添加如下三个路径：“C:\ProgramData\Anaconda3”“C:\ProgramData\Anaconda3\Library\bin” 和“C:\ProgramData\Anaconda3\Scripts”。由于读者在安装过程中可能选择了不同的安装路径，所以需要先自行找到 Anaconda3 的具体安装路径。完成后单击“确定”按钮，如图 1.12 所示。

（11）返回“环境变量”窗口（如图 1.11 所示的环境变量窗口）后，再次单击“确定”按钮。最后，重启计算机，让环境变量的配置生效。

步骤 3　验证安装。

Anaconda 的安装过程包含了 Python 3.9 解释器的安装，并且已经将其注册为系统默认的 Python 解释器。

先验证一下 Python 是否正常使用。按 Win + R 组合键，打开运行窗口，输入 cmd，然后回车，打开命令提示符窗口，如图 1.13 所示。

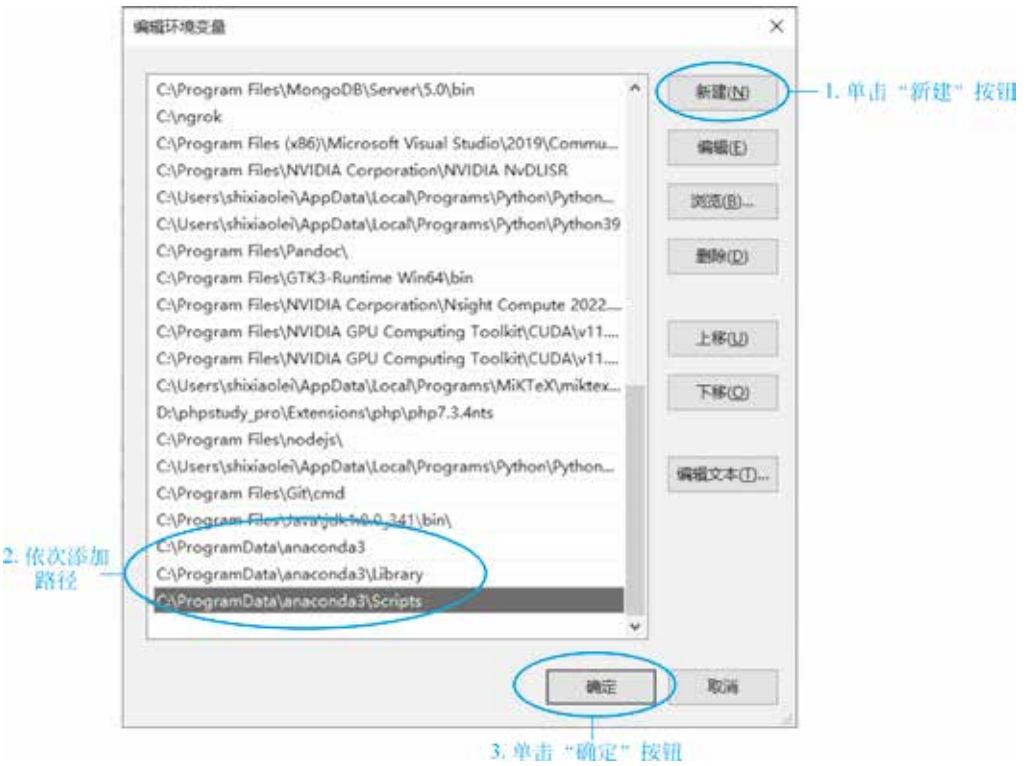

图 1.12　编辑环境变量

图 1.13　打开命令提示符窗口

在命令提示符窗口中输入 Python，然后回车，出现 Python 版本号 3.X.X 字样，最下方出现“>>>”，即代表 Python 安装成功，如图 1.14 所示。

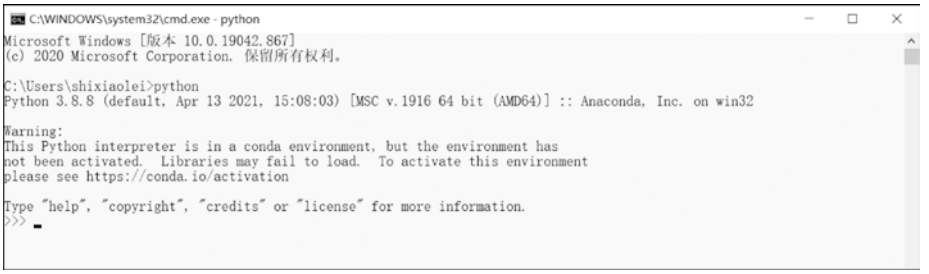

图 1.14　Python 正常运行