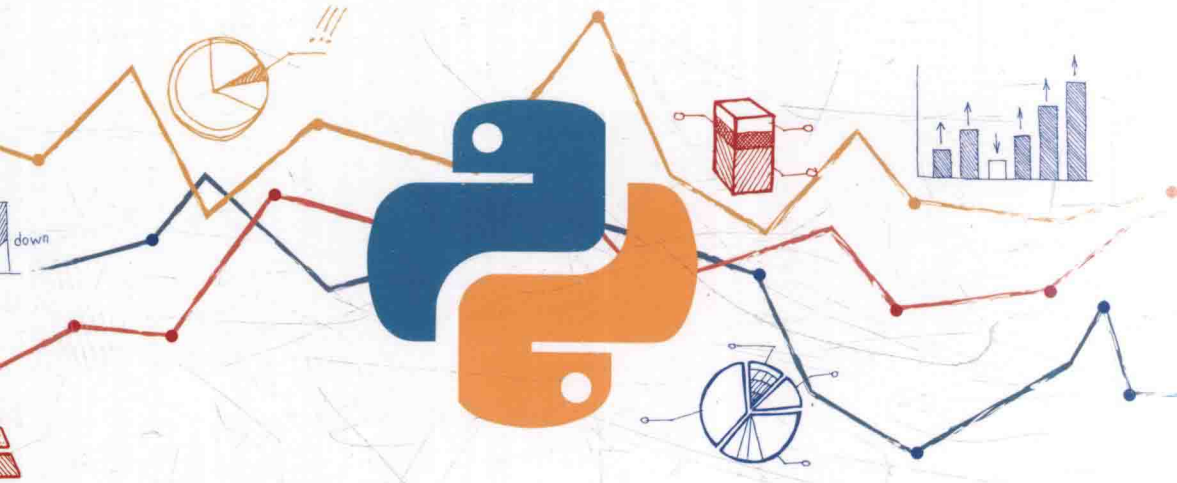
Broadview®
www.broadview.com.cn



Python数据分析入门
从数据获取到可视化

沈祥壮 著

应用Python 轻松实现数据分析和数据处理

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数据分析入门
从数据获取到可视化

沈祥壮 著

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

内 容 简 介

本书作为数据分析的入门图书，以 Python 语言为基础，介绍了数据分析的整个流程。本书内容涵盖数据的获取（即网络爬虫程序的设计）、前期数据的清洗和处理、运用机器学习算法进行建模分析，以及使用可视化的方法展示数据及结果。首先，书中不会涉及过于高级的语法，不过还是希望读者有一定的语法基础，这样可以更好地理解本书的内容。其次，本书重点在于应用 Python 来完成一些数据分析和数据处理的工作，即如何使用 Python 来完成工作而非专注于 Python 语言语法等原理的讲解。本书的目的是让初学者不论对数据分析流程本身还是 Python 语言，都能有一个十分直观的感受，为以后的深入学习打下基础。最后，读者不必须按顺序通读本书，因为各个章节层次比较分明，可以根据兴趣或者需要来自行安排。例如第 5 章介绍了一些实战的小项目，有趣且难度不大，大家可以在学习前面内容之余来阅读这部分内容。

图书在版编目（CIP）数据

Python 数据分析入门：从数据获取到可视化 / 沈祥壮著. —北京：电子工业出版社，2018.3
ISBN 978-7-121-33653-9

Ⅰ. ①P… Ⅱ. ①沈… 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 024600 号

策划编辑：石　倩
责任编辑：牛　勇
印　　刷：三河市华成印务有限公司
装　　订：三河市华成印务有限公司
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：720×1000　1/16　印张：16.5　字数：290 千字
版　　次：2018 年 3 月第 1 版
印　　次：2018 年 3 月第 1 次印刷
印　　数：2500 册　定价：59.00 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。

前言

Python 作为一门优秀的编程语言，近年来受到很多编程爱好者的青睐。一是因为 Python 本身具有简捷优美、易学易用的特点；二是由于互联网的飞速发展，我们正迎来大数据的时代，而 Python 无论是在数据的采集与处理方面，还是在数据分析与可视化方面都有独特的优势。我们可以利用 Python 便捷地开展与数据相关的项目，以很低的学习成本快速完成项目的研究。本书本着实用性的目的，着眼于整个数据分析的流程，介绍了从数据采集到可视化的大致流程。希望借此为 Python 初学者打开数据分析领域的大门，初窥数据分析的奥秘。

本书的主要内容

第 1 章主要讲解了在 Ubuntu 和 Windows 系统下，Python 集成开发环境的搭建。考虑到初学者容易为安装第三方库犯难，又介绍了三种简单实用的方法来安装这些常见的库。接着对几个后面要用到的高级语法进行了简单介绍，为之后的应用打下基础。

第 2 章集中讲解了数据采集的流程，即网络爬虫程序的设计与实现。首先本章没有拘泥于使用 Python 的内置库 urllib 库进行实现，而是直接介绍了 requests 和其他更加简捷强大的库来完成程序的设计。在进阶内容中，对常见的编码问题、

异常处理、代理 IP、验证码、机器人协议、模拟登录，以及多线程等相关问题给出了解决的方案。

第 3 章讲解数据的清洗问题。在具体讲解清洗数据之前，先介绍了 TXT、XLSX、JSON、CSV 等各种文件的导入和导出的方法，并介绍了 Python 与 MySQL 数据库交互的方式。接着介绍了 NumPy 和 pandas 库的基本使用方法，这是我们用于数据处理和科学计算的两个强大的工具。最后综合以上的学习介绍了数据的去重、缺失值的填补等经典的数据清洗方法。

第 4 章首先讲解探索性数据分析的应用，并且简单介绍了机器学习基本知识。然后演示如何应用 sklearn 库提供的决策树和最邻近算法来处理分类问题，并尝试根据算法原理手动实现最邻近算法。最后介绍如何使用 pandas、matplotlib 和 seaborn 这三个库来实现数据的可视化。

第 5 章是综合性学习的章节，讲解了三个小项目的完整实现过程，旨在通过操作生活中真正的数据来强化前面基础内容的学习。

本书的读者对象

本书面向想从事数据工作的 Python 初学者。由于本书并不对 Python 的基础语法做详细的讲解，所以希望读者有一定的语法基础。

测试环境及代码

我们使用的语法是基于 Python 3 的，具体是 Python 3.6，用到的第三方库也已经全面支持此版本，所以读者不必担心相关的版本问题；测试环境为 Ubuntu 16.04 LTS 64-Bit。本书中使用的全部代码及相关数据已经托管至 Github，读者可以进入 https://github.com/shenxiangzhuang/PythonDataAnalysis 进行下载。

联系作者

虽然本书只是入门级图书，但是限于笔者水平有限，难免会存在一些错误，有些地方的表述可能也不是那么准确。非常欢迎读者指出本书的不当之处或提出

建设性的意见。笔者的电子邮件地址是 datahonor@gmail.com。

致谢

在本书的撰写过程中受到过很多人的帮助，这里特别感谢刘松学长，感谢学长对笔者本人长久以来的帮助，从他那里我学到了很多关于 Python 语言、机器学习以及计算机视觉等相关知识。另外，特别感谢 IT 工作者谢满锐先生对本书的细心审校，也感谢他为本书的进一步修改提出建设性意见。同时，感谢电子工业出版社石倩、杨嘉媛编辑的帮助。最后，本书参阅了大量的国内外的文献，这里对有关作者表示衷心的感谢。

读者服务

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。

- **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/33653*

目录

准备

学习目标

- 完成 Linux 或 Windows 系统下 Python 集成开发环境的搭建
- 了解 Python 基础知识

本章作为学习 Python 前的准备环节，主要分为开发环境搭建和 Python 基础语法介绍两部分。这里将介绍在 Linux（实际测试环境为 Ubuntu 16.04 LTS 64Bit）和 Windows 系统下（实际测试环境为 Windows 10）搭建 Python 集成开发环境的详细步骤，以及 Anaconda（附带 Spyder 编辑器）和 PyCharm 的安装与配置。鉴于本书面向 Python 初学者，所以操作步骤比较详细，已经完成安装的读者可以略过此环境配置部分。但是建议大家阅读使用 conda 和 pip 安装第三方库的部分，本书中讲到的所有库均通过这两种包管理器安装。此外本章将介绍一些 Python 的基本语法。当然，这里只是粗略地介绍一些初学者难以理解的内容，例如装饰器和列表解析等。

1.1 开发环境搭建

本书所有的代码都是在 Ubuntu 16.04 LTS 64 Bit 英文版系统下完成的，因此首先将介绍在 Ubuntu 下环境的搭建。当然，考虑到有很多读者使用的是 Windows 系统，所以也会对 Windows 下环境的搭建进行相应的介绍，这里我们选用的是 Windows 10。对于编程学习者来说，Linux 是一个很好的系统，它有多种开源的工具可以给我们的开发带来便利，同时其活跃的社区环境和网络上大量的资料使得日常遇到的大部分问题可以迅速得到解决。所以建议读者在 Linux 系统环境下进行开发，开始的时候可以尝试在虚拟机中使用，之后考虑在本机上直接安装。当然，在 Windows 系统下开发并不影响学习本书的内容。

1.1.1 在 Ubuntu 系统下搭建 Python 集成开发环境

1. Anaconda（Spyder）安装与配置

Ubuntu 是自带 Python 环境的（Python 2），按下快捷键[Ctrl + Alt + T]（或者在桌面空白处单击右键，在弹出的快捷菜单中选择“open terminal”命令）打开终端，输入 `python` 即可，如图 1-1 所示。

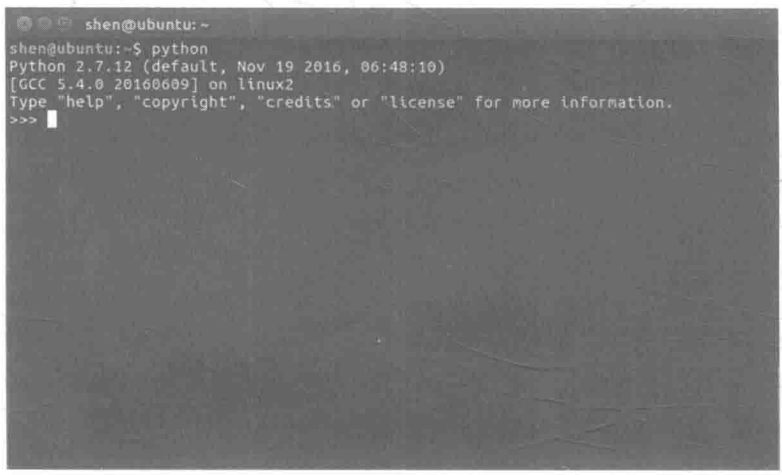


图 1-1

由于我们学习的是 Python 3，所以将使用 Anaconda 完成 Python 3 的环境配置。

“Anaconda 是用于大规模数据处理、预测分析和科学计算的 Python 和 R 编程语言的免费平台，旨在简化包管理和部署”[①]。第三方库的安装对于初学者来说可能是一件比较头疼的事，但是它集成了很多用于数据处理和科学计算的第三方库，使得我们不用额外再去安装。同时，Anaconda 提供了强大的安装包管理功能，这点会在后面详细介绍。Anaconda 还自带一款十分优秀的编辑器——Spyder，它的界面和使用方法与 MATLAB 和 RStudio 十分相像，其特点在于中间变量的储存。下面介绍具体的安装步骤。

> 注意：下面的命令均是在终端执行的，并且要根据需要切换到特定目录后再执行对应的命令。此外随着版本的更新，下载文件的文件名可能会改变，请读者仔细查看并在必要时对命令做出修改。

首先，进入 Anaconda 官网（https://www.anaconda.com/download）下载对应版本的安装文件，这里选择 Python 3.6 version 64-BIT(X86)INSTALLER(499M)，默认下载到 Downloads 文件夹。之后，通过快捷键[Ctrl + Alt + T]（或者在桌面空白处单击鼠标右键，在弹出的快捷菜单中选择“open terminal”命令）打开终端，输入命令 `cd Downloads/` 后切换到包含下载文件的目录下，运行命令 `bash Anaconda3-4.4.0-Linux-x86_64.sh`，开始安装（Anaconda3-4.4.0-Linux-x86_64.sh 是下载的文件名），如图 1-2 所示。

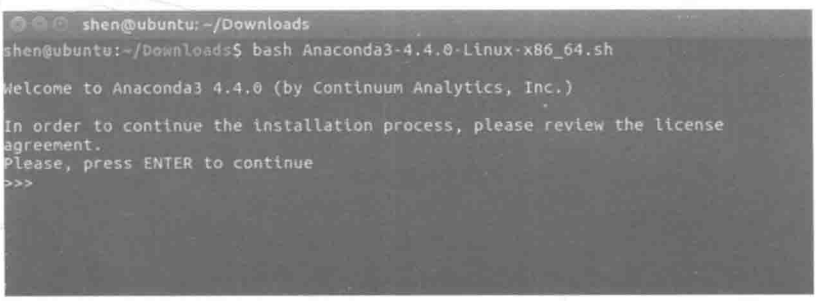


图 1-2

① 参考维基百科：https://en.wikipedia.org/wiki/Anaconda。

然后按回车键确认，继续运行安装程序。接着会出现与协议相关的确认信息，如图 1-3 所示。

```
openssl
The OpenSSL Project is a collaborative effort to develop a robust,
commercial-grade, full-featured, and Open Source toolkit implementing the
Transport Layer Security (TLS) and Secure Sockets Layer (SSL) protocols as
well as a full-strength general purpose cryptography library.

pycrypto
A collection of both secure hash functions (such as SHA256 and RIPEMD160),
and various encryption algorithms (AES, DES, RSA, ElGamal, etc.).

pyopenssl
A thin Python wrapper around (a subset of) the OpenSSL library.

kerberos (krb5, non-Windows platforms)
A network authentication protocol designed to provide strong authentication
for client/server applications by using secret-key cryptography.

cryptography
A Python library which exposes cryptographic recipes and primitives.

Do you approve the license terms? [yes|no]
>>> Please answer 'yes' or 'no':
```

图 1-3

输入 yes 并按回车键继续安装，接下来选择安装路径，这里直接按回车键选择默认的路径，如图 1-4 所示。

```
Please answer 'yes' or 'no':
>>> yes

Anaconda3 will now be installed into this location:
/home/shen/anaconda3

  - Press ENTER to confirm the location
  - Press CTRL-C to abort the installation
  - Or specify a different location below

[/home/shen/anaconda3] >>>
PREFIX=/home/shen/anaconda3
installing: python-3.6.1-2 ...
installing: _license-1.1-py36_1 ...
installing: alabaster-0.7.10-py36_0 ...
installing: anaconda-client-1.6.3-py36_0 ...
installing: anaconda-navigator-1.6.2-py36_0 ...
```

图 1-4

接着会进行一系列的配置，稍后询问是否将 Anaconda 加入环境变量，这里选择加入，如图 1-5 所示。

```
Python 3.6.1 :: Continuum Analytics, Inc.
creating default environment...
installation finished.
Do you wish the installer to prepend the Anaconda3 install location
to PATH in your /home/shen/.bashrc ? [yes|no]
[no] >>> yes

Prepending PATH=/home/shen/anaconda3/bin to PATH in /home/shen/.bashrc
A backup will be made to: /home/shen/.bashrc-anaconda3.bak


For this change to become active, you have to open a new terminal.

Thank you for installing Anaconda3!

Share your notebooks and packages on Anaconda Cloud!
Sign up for free: https://anaconda.org
```

图 1-5

按照提示，打开一个新的终端测试是否成功安装，如图 1-6 所示。

```
shen@ubuntu: ~
shen@ubuntu:~$ conda list
# packages in environment at /home/shen/anaconda3:
#
_license                  1.1                      py36_1
alabaster                 0.7.10                   py36_0
anaconda                  4.4.0                    np112py36_0
anaconda-client           1.6.3                    py36_0
anaconda-navigator        1.6.2                    py36_0
anaconda-project          0.6.0                    py36_0
asn1crypto                0.22.0                   py36_0
```

图 1-6

从图 1-6 中可以看出，`conda list` 命令给出了已经安装的第三方库的列表，表示安装成功。

由于已经将其加入环境变量，所以此时默认的 Python 版本即为 Python 3.6，由 Anaconda 提供，而原来的 Python 2 也可以正常使用，调用方法如图 1-7 所示。

```
shen@ubuntu:~$ python
Python 3.6.1 |Anaconda 4.4.0 (64-bit)| (default, May 11 2017, 13:09:58)
[GCC 4.4.7 20120313 (Red Hat 4.4.7-1)] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>>
[6]+  Stopped                 python
shen@ubuntu:~$ python3
Python 3.6.1 |Anaconda 4.4.0 (64-bit)| (default, May 11 2017, 13:09:58)
[GCC 4.4.7 20120313 (Red Hat 4.4.7-1)] on linux
Type "help", "copyright", "credits" or "license" for more information.
>>>
[7]+  Stopped                 python3
shen@ubuntu:~$ python2
Python 2.7.12 (default, Nov 19 2016, 06:48:10)
[GCC 5.4.0 20160609] on linux2
Type "help", "copyright", "credits" or "license" for more information.
>>>
```

图 1-7

我们也可以在终端通过 `ipython` 命令来使用 IPython 这一优秀的交互式环境，如图 1-8 所示。本书大部分的代码便是在此进行测试。

```
shen@ubuntu:~$ ipython
Python 3.6.1 |Anaconda 4.4.0 (64-bit)| (default, May 11 2017, 13:09:58)
Type "copyright", "credits" or "license" for more information.

IPython 5.3.0 -- An enhanced Interactive Python.
?         -> Introduction and overview of IPython's features.
%quickref -> Quick reference.
help      -> Python's own help system.
object?   -> Details about 'object', use 'object??' for extra details.

In [1]:
```

图 1-8

此外，可以通过 `spyder` 命令启动自带的编辑器 Spyder，如图 1-9 所示。

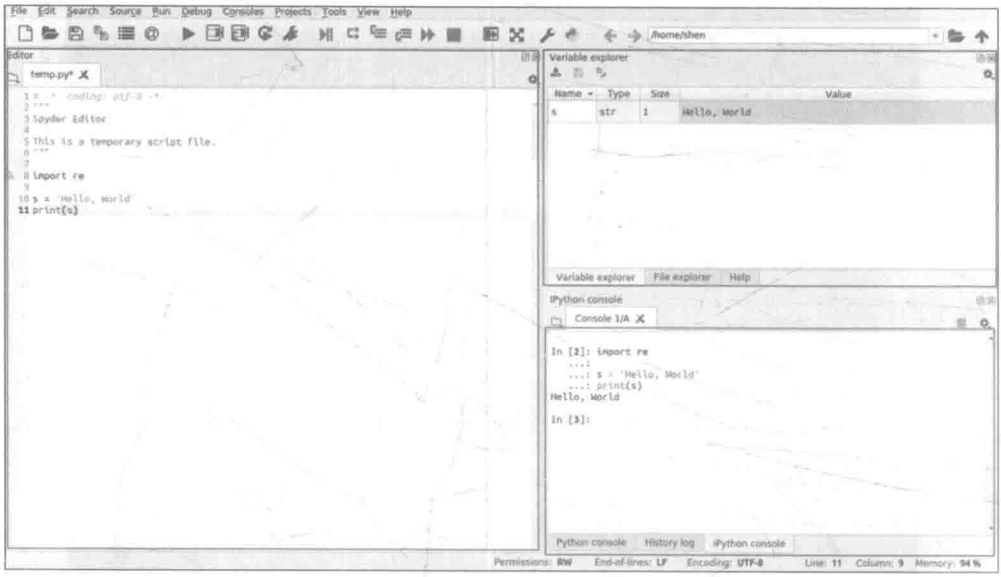


图 1-9

其界面与 MATLAB 和 RStudio 十分类似，主要分为三部分：左侧的编辑区，右上方的变量查看和文件管理区，以及右侧下方的 IPython 交互式命令行区。可以通过选中左侧部分代码，再按快捷键[Ctrl+Enter]测试部分代码，十分方便。

作为专注于数据的开源工具，新版 Anaconda 提供导航器（navigator），其包

含更加丰富的内容，可以通过命令 `anaconda-navigator` 启动，界面如图 1-10 所示。

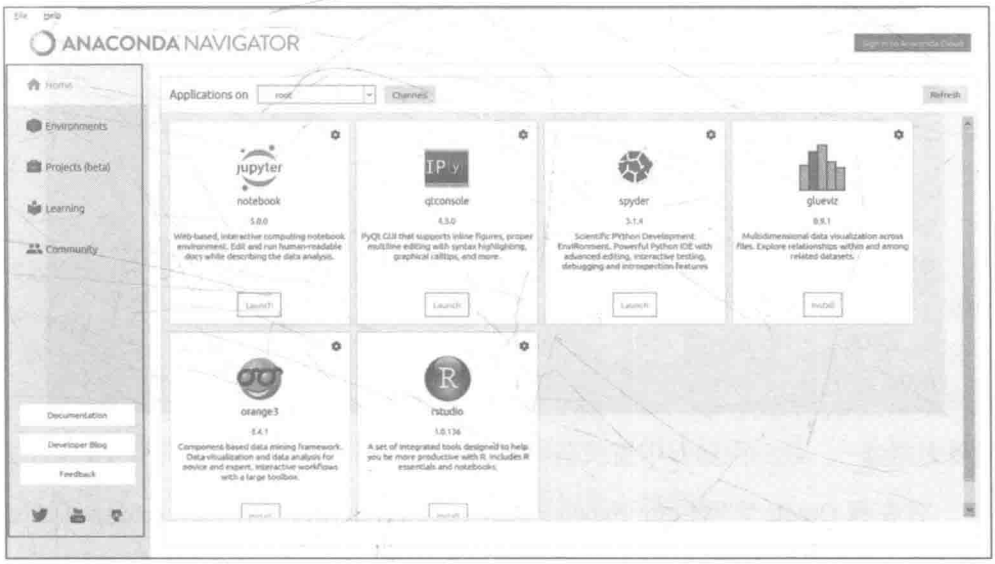


图 1-10

这里可以便捷地启动一系列的工具，此外，还包含其他有用的内容，读者可自行探索。

至此，完成了对 Anaconda 安装和基本使用方法的介绍。接下来就可以使用 Spyder 编辑器开发了，不过在此之前先为大家介绍另一款更加优秀的编辑器——PyCharm。不过 PyCharm 需要 Java 环境，所以接下来首先介绍 Java 环境的搭建。

2. Java 环境搭建

Ubuntu 默认是没有 Java 环境的，可以通过命令 `java-version` 查看，如图 1-11 所示。

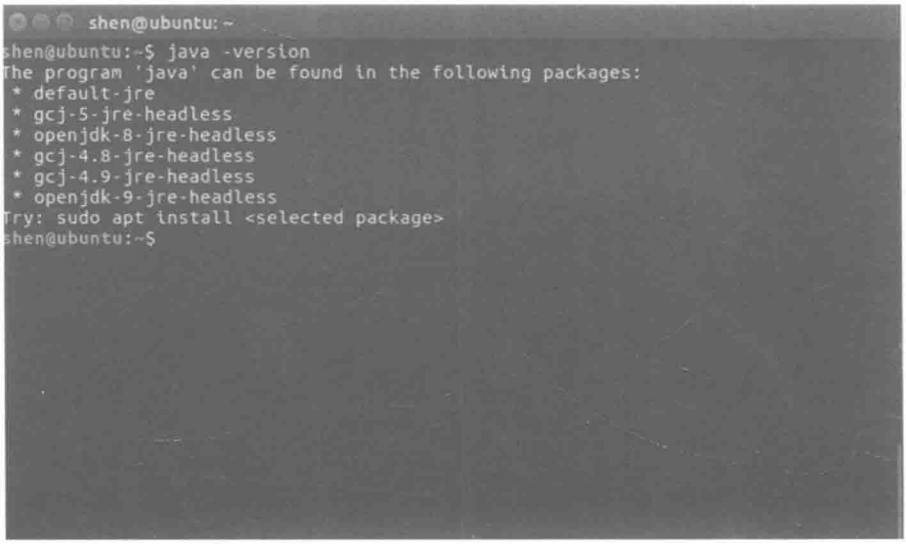


图 1-11

首先到 Oracle 官网（http://www.oracle.com/technetwork/java/javase/downloads/jdk8-downloads-2133151.html）找到对应的下载文件，这里选择 jdk-8u131-linux-x64.tar.gz。默认下载到 Downloads 文件夹。之后通过命令 `sudo su`，输入密码，切换到管理员身份进行下面的配置；然后输入命令 `cd '/usr'`，切换到 usr 文件夹下，再通过 `tar -zxvf '/home/shen/Downloads/jdk-8u131-linux-x64.tar.gz'` 将压缩包解压（注意：这里的路径是压缩包的绝对路径）。运行 `mv jdk1.8.0_131 jdk-8`，将文件重命名为 jdk-8。

接下来修改系统配置文件，运行 `gedit /etc/profile`，打开文件，在文件最后加入以下代码。

```
export JAVA_HOME=/usr/jdk-8
export JRE_HOME=$JAVA_HOME/jre
export CLASSPATH=.:$CLASSPATH:$JAVA_HOME/lib:$JRE_HOME/lib
export PATH=$PATH:$JAVA_HOME/bin:$JRE_HOME/bin
```

最后运行 `source /etc/profile`，使配置文件生效。再次查看 Java 环境版本，可以看到安装成功，如图 1-12 所示。