版权相关注意事项：

1、书籍版权归原者和出版社所有

2、本PDF未得子子己的信息平台，经过整理而成

3、本PDF仅用于非商业用途或者个人交流研究学习使用

4、本PDF获得者不得在互联网上以任何目的进行传播，违规者造成的法律责任和后果，违规者自负

5、如果您将书籍内容摘录，请一定购买正版实体书，多多支持编写高质量图书的作者和出版的出版社！当然，如果图书内容不堪入目，质量低下，你也可以选择跳读清晰复杂PDF

6、技术类书籍是非采购取问题的，不是非卖收藏的，你得到了书籍不意味着你得到了版权，所以请不要得到书籍后就觉得法以自喜，要经常关顾！！经常顾顾

7、请于下载PDF后24小时内研究使用并删除本PDF

版权相关注意事项：

1、书籍版权归原者和出版社所有

2、本PDF未得子子己的信息平台，经过整理而成

3、本PDF仅用于非商业用途或者个人交流研究学习使用

4、本PDF获得者不得在互联网上以任何目的进行传播，违规者造成的法律责任和后果，违规者自负

5、如果您将书籍内容摘录，请一定购买正版实体书，多多支持编写高质量图书的作者和出版的出版社！当然，如果图书内容不堪入目，质量低下，你也可以选择跳读清晰复杂PDF

6、技术类书籍是非采购取问题的，不是非卖收藏的，你得到了书籍不意味着你得到了版权，所以请不要得到书籍后就觉得法以自喜，要经常关顾！！经常顾顾

7、请于下载PDF后24小时内研究使用并删除本PDF

数据分析必备开源语言包括R、Python、Perl、Ruby、Lua、Tcl、Scala、Clojure、Haskell、Erlang、Lisp、Ocaml、Octave、Scilab、Gnuplot、Processing、Protovis、Flot、D3.js、Exhibit、Jsplot、Jsplot、Jsplot

仅供非商业用途或交流学习使用

CONTENTS

目录

第1部分 基 础 篇

第1章 Python 语言基础 /2

1.0 引子 /2

1.1 工欲善其事，必先利其器（安装 Python） /3

1.2 学跑得先学走（语法基础） /9

1.3 程序结构 /11

1.3.1 Hello World ! /11

1.3.2 运算符介绍 /12

1.3.3 顺序结构 /14

1.3.4 判断结构 /17

1.3.5 循环结构 /18

1.3.6 异常 /20

1.4 函数 /24

1.4.1 基本函数结构 /24

仅供非商业用途或交流学习使用

仅供非商业用途或交流学习使用

1.4.2 参数结构 /25
1.4.3 回调函数 /28
1.4.4 函数的递归与嵌套 /28
1.4.5 闭包 /31
1.4.6 匿名函数 lambda /32
1.4.7 关键字 yield /32
1.5 数据结构 /35
1.5.1 列表（list） /35
1.5.2 元组（tuple） /38
1.5.3 集合（set） /39
1.5.4 字典（dict） /40
1.5.5 集合的操作 /41
1.5.6 学以致用 /45
1.6 3 个函数（map、filter、reduce） /47
1.6.1 遍历函数（map） /47
1.6.2 筛选函数（filter） /48
1.6.3 累计函数（reduce） /48
1.7 面向对象编程基础 /50
1.7.1 类 /50
1.7.2 类和实例 /51
1.7.3 数据封装 /52
1.7.4 私有变量与私有方法 /53
本章小结 /54

仅供非商业用途或交流学习使用

仅供非商业用途或交流学习使用

第 2 章

数据处理 /60

2.1 Anaconda 简介 /60

2.2 Numpy 简介 /66

2.3 关于 Pandas /68

2.3.1 什么是 Pandas /68

2.3.2 Pandas 中的数据结构 /68

2.4 数据准备 /68

2.4.1 数据类型 /68

2.4.2 数据结构 /69

2.4.3 数据导入 /79

2.4.4 数据导出 /86

2.5 数据处理 /88

2.5.1 数据清洗 /89

2.5.2 数据抽取 /97

2.5.3 插入记录 /114

2.5.4 修改记录 /117

2.5.5 交换行或列 /120

2.5.6 排名索引 /122

2.5.7 数据合并 /131

2.5.8 数据计算 /137

2.5.9 数据分组 /141

2.5.10 日期处理 /143

仅供非商业用途或交流学习使用

带你飞（数据处理案例） /148
本章小结 /160

第 3 章
数据分析 /165
3.1 基本统计分析 /165
3.2 分组分析 /169
3.3 分布分析 /171
3.4 交叉分析 /173
3.5 结构分析 /174
3.6 相关分析 /176
小试牛刀（相关分析案例：电商数据分析） /178
本章小结 /180

第 4 章
数据可视化 /181
4.1 使用 Python 对数据进行可视化处理 /181
4.1.1 准备工作 /181
4.1.2 Matplotlib 绘图示例 /186
4.1.3 Seabon 中的图例 /198
4.1.4 pandas 的一些可视化功能 /212
4.1.5 文本数据可视化 /217
4.1.6 networkx 网络图 /218
4.1.7 folium 绘制地图 /220
4.2 Python 图像处理基础 /221

试读结束：需要全本请在线购买： www.ertongbook.com

4.2.1 PIL 图库 /221
4.2.2 OpenCV 图库 /224
本章小结 /226

第 5 章
字符串处理与网络爬虫 /228
5.1 字符串处理 /228
5.1.1 字符串处理函数 /228
5.1.2 正则表达式 /230
5.1.3 编码处理 /237
5.2 网络爬虫 /240
5.2.1 获取网页源码 /240
5.2.2 从源码中提取信息 /241
5.2.3 数据存储 /246
5.2.4 网络爬虫从这里开始 /248
本章小结 /260

第 2 部分 实战案例篇

第 6 章
词云 /262
6.1 安装文件包 /263
6.2 jieba 功能用法 /264
6.2.1 cut 用法 /264

6.2.2 词频与分词字典 /265
6.3 文本词云图 /269
6.4 背景轮廓词云图的制作 /271
6.4.1 数据准备 /271
6.4.2 分词 /272
6.4.3 构建词云 /273
本章小结 /278

第 7 章
航空客户分类 /279
7.1 问题的提出 /279
7.2 聚类分析相关概念 /280
7.3 模型的建立 /281
7.4 Python 实现代码 /281
7.5 分类结果展示与分析 /284
本章小结 /287

第 8 章
《红楼梦》文本分析 /288
8.1 准备工作 /289
8.2 分词 /291
8.2.1 读取数据 /291
8.2.2 数据预处理 /293
8.2.3 对红楼梦进行分词 /301
8.2.4 制作词云 /303

8.3 文本聚类分析 /312
8.3.1 构建分词 TF-IDF 矩阵 /312
8.3.2 使用 TF-IDF 矩阵对章节进行聚类 /314
8.4 LDA 主题模型 /322
8.5 人物社交网络分析 /328
本章小结 /334

第 3 部分 拓展与延伸

第 9 章
Python 字符串格式化 /336
9.1 使用%符号进行格式化 /336
9.2 使用 format()方法进行格式化 /339
9.3 使用 f 方法进行格式化 /341
本章小结 /342

第 10 章
在 Python 中操作 MySQL 数据库 /343
10.1 对 MySQL 的连接与访问 /344
10.2 对 MySQL 的增、删、改、查操作 /345
10.2.1 查询操作 /345
10.2.2 插入操作 /346
10.2.3 更新操作 /347
10.2.4 删除操作 /347

10.3 创建数据库表 /348
本章小结 /349
第 11 章
fractal（分形）库的发布 /350
11.1 用 Python 绘制分形 /351
11.1.1 分形简介 /351
11.1.2 先睹为快 /351
11.1.3 绘制方法简介 /352
11.2 第三方库发布到 PyPi /364
本章小结 /369
参考文献 /370

第 1 部分

基础篇

Python 语言基础

数据处理

数据分析

数据可视化

字符串处理与网络爬虫

第 1 章

Python 语言基础

1.0 引子

第一次接触 Python 是在 2013 年，那年笔者刚好从加拿大访学回来，正想鼓捣一个自己的网站，看到很多网文说用 Django 搭建网站很不错，于是开始疯狂地搜索关于 Django 的学习资料。看着看着，就发现 Django 是基于 Python 基础之上的，于是回过头来学习 Python，从此就与 Python 结下了不解之缘。

现在回想起来，那时候学习 Python 确实吃了不少苦头。刚入手时，遇到的第一个问题就是选择 Python 的版本问题。那是 Python 2.7 风行的年代，买回来的书、网上能搜索到的资料，几乎都是用 Python 2.7 写的。好不容易摸到了 Python 的大门，却被那些还没有完善、兼容的库折磨得死去活来。看别人的代码轻松又简单，到自己输入代码时，处处提示错误，用百度搜索半天才从知乎上的英文答案里摸索出一点眉目——库不兼容！

试读结束：需要全本请在线购买： www.ertongbook.com

现在回过头来看看自己爬过的“坑”，也是一笔不小的财富。今天把这些写下来，为的是让那些跟笔者一样的“小白”们不要再重蹈覆辙，少走点弯路吧！

Python 自从 1991 年正式发布以来，由于其简洁易懂、扩展性强，确实受到很多程序员的追捧，他们编写了很多类库，使得 Python 的应用越来越方便，吸引了更多的人使用。近几年，谷歌等大型互联网公司开始使用 Python 语言来编写 AI 程序，在机器学习、神经网络、模式识别、人脸识别、定理证明、大数据等各个领域都产生了很多可以由 Python 直接引用的功能模块。现在最流行的深度学习框架也是用 Python 编写的，比如震惊世界的 AlphaGo，其大部分程序就是用 Python 编写的。随着人工智能的火爆，Python 几乎被推上了神坛，获得了“人工智能标配语言”的美誉，相关程序员的薪金也水涨船高。

随着 Python 3 的崛起和各种库的完善，Python 在数据分析、科学计算领域的应用越来越多。除了 Python 语言本身的特点，很多第三方库也很好用。某些场景下，Python 甚至可以取代 MATLAB。Python 常见的数据分析库有 Pandas、NumPy、SciPy 等。

Python 前后版本的不兼容确实让新、老学员感到头痛，许多人在为选择 Python 2.x 还是 Python 3.x 而发愁。不过现在应该不存在这个问题了，现在的主流版本是 Python 3.6，很多库都已经兼容或者升级到了 Python 3.x，而且有消息称 Python 2.7 到 2020 年就不再维护了，所以本书使用的版本是 Python 3.6。

如果读者已经有了一定的 Python 基础，那么完全可以越过本章进入下一章的学习。毕竟，基础知识的学习还是有点枯燥乏味的。但还是建议读者先浏览一遍本章小结，尤其是“坑点”提示内容。

1.1　工欲善其事，必先利其器（安装 Python）

Windows 用户可以访问 https://Python.org/download，从中下载最新版本的 Python，如图 1-1 所示。其大小约为 27MB 左右，与其他大多数语言安装软件相比，还是十分紧凑小巧的。安装过程与其他 Windows 软件类似。

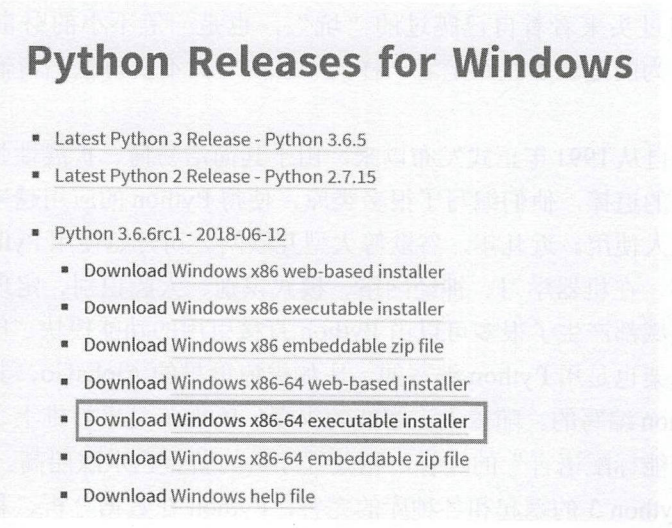
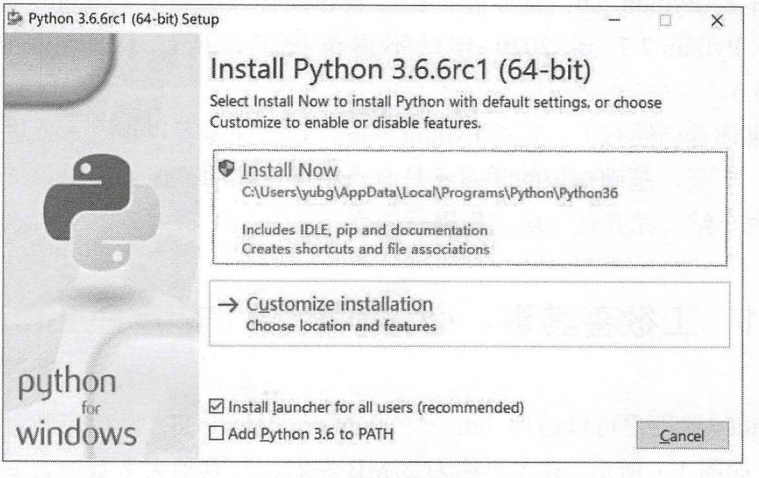


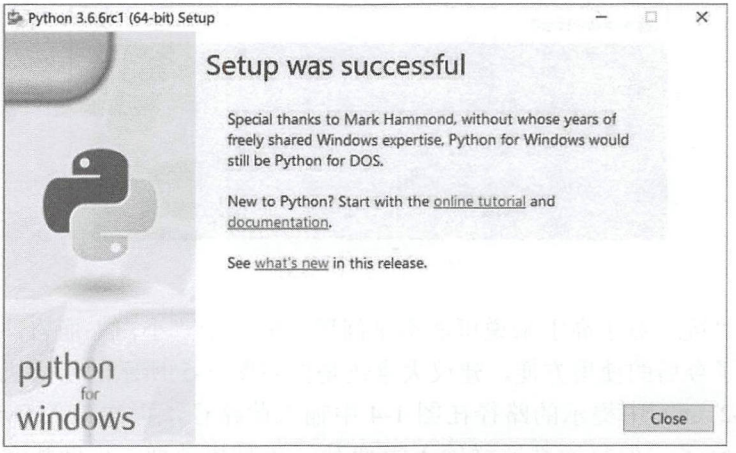
图 1-1　Python 安装版本

使用 Windows 系统的读者建议下载 executable installer 版本。笔者在此下载的版本为 Python 3.6.6，操作系统为 Windows 10。安装 Python 很简单，双击 Python-3.6.6.exe，选择默认选项，直接单击 Install Now 按钮即可，如图 1-2 所示。安装完毕后会显示安装成功，单击 Close 按钮即可使用。



（a）Python 安装界面 1

图 1-2　Python 安装界面



（b）Python 安装界面 2

图 1-2　Python 安装界面（续）

安装完成后，在桌面“开始”菜单中会显示已安装组件的目录，如图 1-3 所示。当我们要编写代码时，直接选择 IDLE 组件即可。

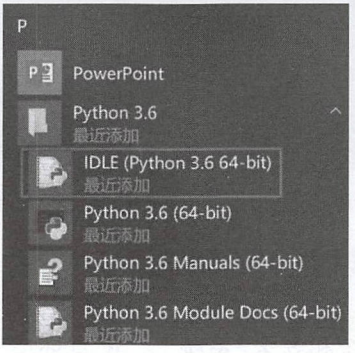


图 1-3　“开始”菜单

当然，为了讲得全面一点，介绍 Python 的设置还是有必要的。如果读者觉得这部分内容比较“高深”，也可以先跳过 Python 的设置，直接看第 2 章的 Anaconda 安装，毕竟后面的内容大多都是在 Anaconda 下编写程序代码及运行。这样做的最大好处是不用安装很多公共的包和库。

此时，若在 DOS Shell 命令提示符下输入 Python，很遗憾，运行将失败，提示信息如图 1-4 所示。

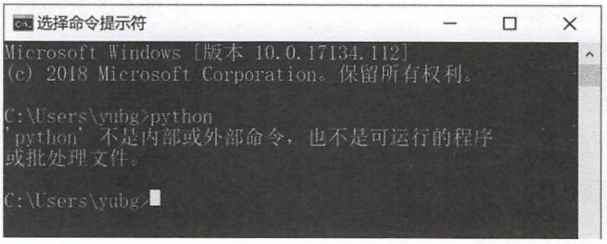


图 1-4　路径不对

这个“坑”对于高手来说可能不是问题，但对于“小白”而言，那就是拦路虎。为了今后的使用方便，建议大家还是按照图 1-5 所示的步骤执行一遍。先按图 1-2（a）中提示的路径在图 1-4 中输入此路径，再次输入 Python 就可以正常运行了。但每次都这样输入很费劲，还是提前把工作做扎实了吧，否则后面 pip 安装各类库还会遇到很多意想不到的问题。

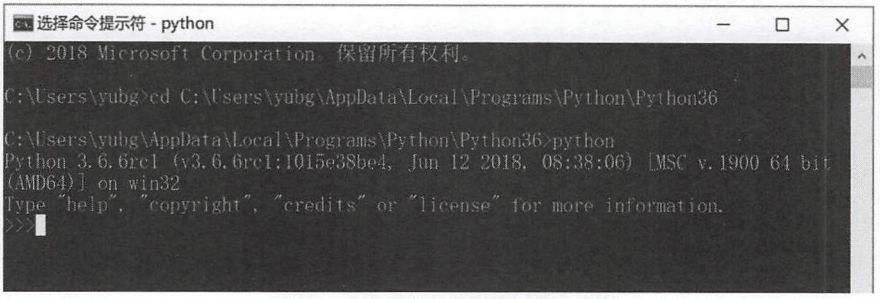


图 1-5　命令提示符下启动 Python

上面不能运行的问题主要是没有配置好环境变量。其实也不难，自己动手 DIY 吧。

打开“控制面板>系统和安全>系统”，配置环境变量（按图 1-6 中标号 A、B、C、D 的步骤操作即可）。如果找不到相关步骤，那就百度一下吧。

接下来添加地址变量。把 Python 的安装路径新增到 path 环境变量中去，如图 1-7 所示。在笔者的计算机中，Python 安装在 C:\Users\yubg\AppData\Local\Programs\Python\Python36 下，按照图 1-7 中标号 A、B、C 的步骤进行操作就可以了，完成后单击“确定”按钮即可。再在 DOS Shell 命令提示符下输入 Python，就会得到如图 1-8 所示的结果。以后直接运行 Python 就可以了，省去了输入 cd 及 Python 安装路径之苦。

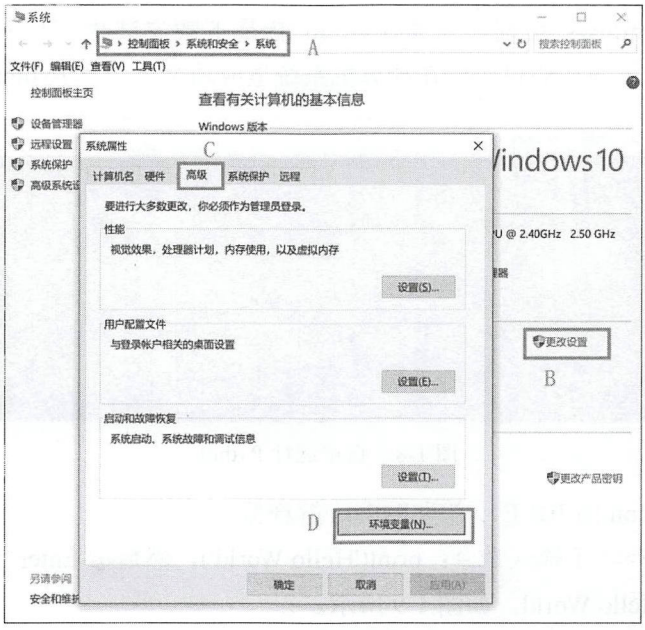


图 1-6　配置环境变量

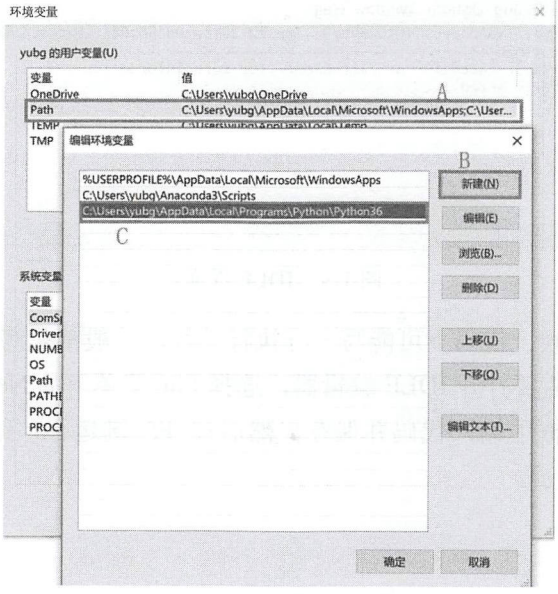


图 1-7　添加地址变量

对于 Windows 系统，安装 Python 就是下载安装程序后双击它那么简单。从现在起，我们假设已经在计算机系统里成功地安装了 Python 3.6。

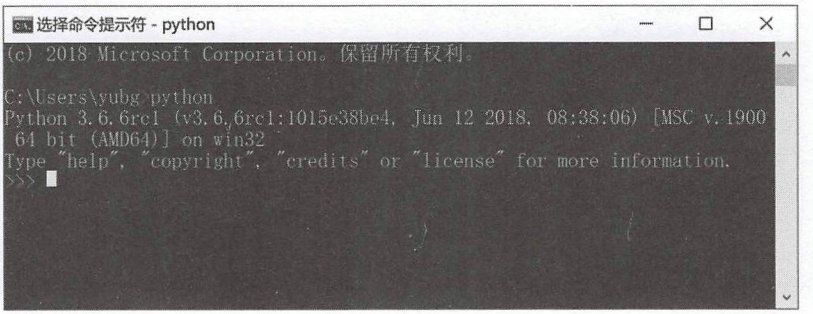


图 1-8　直接运行 Python

打开 Python 的 IDLE，启动 Python 解释器。

在提示符>>>下输入命令：print('Hello World')，然后按 Enter 键，可以看到输出的单词 Hello World，如图 1-9 所示。

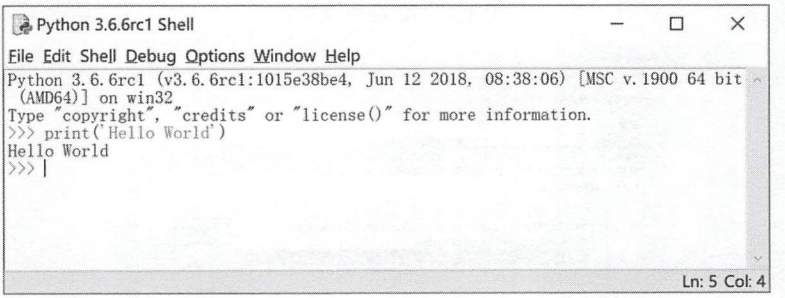


图 1-9　IDLE 界面

在更多的时候，代码不可能写一行执行一行，一般都是写完一个模块再去执行。这时候就需要打开 IDLE 编辑器，选择 File 菜单下的 New File 命令，在弹出的空白编辑框中输入代码并保存，然后按 F5 键运行，结果显示在 Shell 上，如图 1-10 所示。

试读结束：需要全本请在线购买： www.ertongbook.com