

Python

网络爬虫开发

从入门到精通

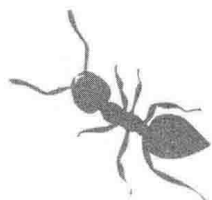
刘延林◎编著

学爬虫、抓数据、提内容、存数据、懂反爬、
学框架、会部署、重实战……

一书在手，精通Python网络爬虫！



北京大学出版社
PEKING UNIVERSITY PRESS



Python

网络爬虫开发

从入门到精通

刘延林◎编著



北京大学出版社
PEKING UNIVERSITY PRESS

内 容 提 要

本书共分3篇,针对Python爬虫初学者,从零开始,系统地讲解了如何利用Python进行常见的网络爬虫的程序开发。

第1篇快速入门篇(第1章~第9章):本篇主要介绍了Python环境的搭建和一些Python的基础语法知识等、Python爬虫入门知识及基本的使用方法、Ajax数据的分析和抓取、动态渲染页面数据的爬取、网站代理的设置与使用、验证码的识别与破解,以及App数据抓取、数据的存储方法等内容。

第2篇技能进阶篇(第10章~第12章):本篇主要介绍了PySpider和Scrapy两个常用爬虫框架的基本使用方法、分布式爬虫的实现思路,以及数据分析、数据清洗常用库的使用方法。

第3篇项目实战篇(第13章):本篇通过6个综合实战项目,详细地讲解了Python数据爬虫开始与实战应用。本篇对全书内容进行了总结回顾,强化读者的实操水平。

本书案例丰富,注重实战,既适合Python程序员和爬虫爱好者阅读学习,也适合作为广大职业院校相关专业的教学用书。

图书在版编目(CIP)数据

Python网络爬虫开发从入门到精通 / 刘延林编著. —北京:北京大学出版社, 2019.12
ISBN 978-7-301-30909-4

I. ①P… II. ①刘… 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(2019)第238679号

书 名 Python网络爬虫开发从入门到精通

PYTHON WANGLUO PACHONG KAIFA CONG RUMEN DAO JINGTONG

著作责任者 刘延林 编著

责任编辑 吴晓月 王继伟

标准书号 ISBN 978-7-301-30909-4

出版发行 北京大学出版社

地 址 北京市海淀区成府路205号 100871

网 址 <http://www.pup.cn> 新浪微博: @北京大学出版社

电子信箱 pup7@pup.cn

电 话 邮购部 010-62752015 发行部 010-62750672 编辑部 010-62570390

印 刷 者 北京溢漾印刷有限公司

经 销 者 新华书店

787毫米×1092毫米 16开本 23.25印张 528千字

2019年12月第1版 2019年12月第1次印刷

印 数 1-4000册

定 价 79.00元

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

举报电话:010-62752024 电子信箱: fd@pup.pku.edu.cn

图书如有印装质量问题,请与出版部联系,电话:010-62756370

前言

Preface

为什么写这本书？

随着互联网进入大数据时代，尤其是人工智能浪潮兴起的时代，爬虫技术迎来了一波新的振兴浪潮。在大数据架构中，数据的收集存储与统计分析占据了极为重要的地位，而数据的收集很大程度上依赖于爬虫的爬取，所以网络爬虫也逐渐变得越来越火爆。

在众多的网络爬虫工具中，Python 以其使用简单、功能强大等优点成为网络爬虫开发的最常用工具。相比其他语言，Python 是一门非常适合开发网络爬虫的编程语言，内置大量的框架和库，可以轻松实现网络爬虫功能。Python 爬虫可以做的事情很多，如广告过滤、Ajax 数据爬取、动态渲染页面爬取、App 数据抓取、使用代理爬取、模拟登录爬取、数据存取等，Python 爬虫还可以用于数据分析，在数据的抓取方面可以说作用巨大！

这本书的特点是什么？

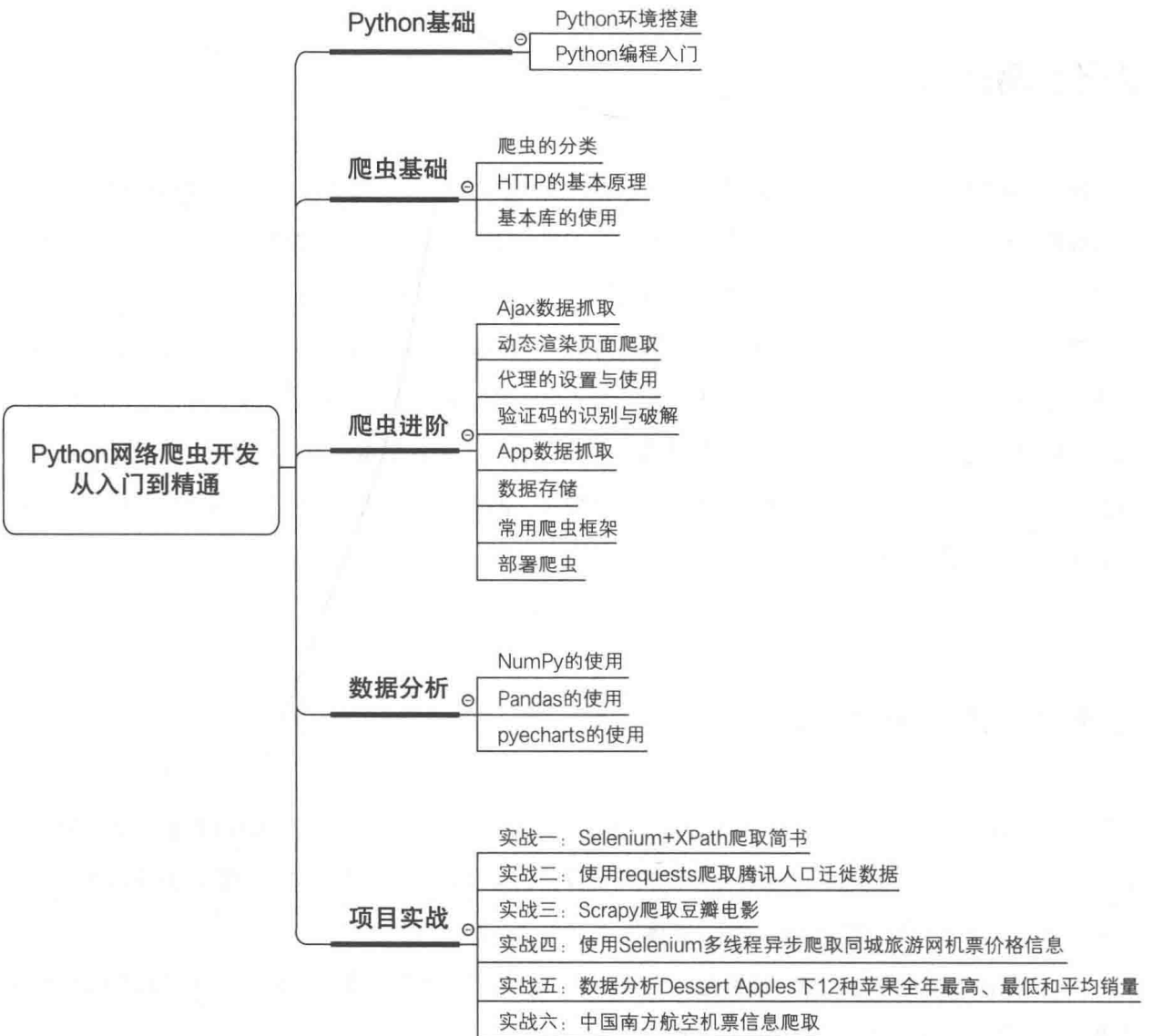
本书力求简单、实用。坚持以实例为主，理论为辅的路线。全书共 13 章，从 Python 基础、爬虫开发常用网络请求库，到爬虫框架使用和分布式爬虫设计，以及最后的数据存储、分析、实战训练等，覆盖了爬虫项目开发阶段的整个生命周期。整体上本书内容有以下几个特点。

(1) 没有高深的理论，每一章都是以实例为主，读者参考源码，修改实例，就能得到自己想要的结果。目的是让读者看得懂、学得会、做得出。

(2) 实训与问答，几乎每章都有配备。目的是让读者看完之后，尽快巩固知识，举一反三，学以致用。

(3) 内容系统全面，实战应用性强。本书内容在写作定位上，适合零基础读者学习，然后逐步掌握相关知识技能，从而达到从入门到精通的学习效果。另外，全书在知识讲解中，都安排了丰富的实训实战案例，目的是增强读者的实际动手能力。

在这本书里写了些什么？



写给读者的建议

读者在阅读本书时，如果是零基础，建议先从 Python 基础开始学习。因为学习爬虫需要读者对 Python 的基础语法和结构有深刻的理解和熟练应用，这样才能在后面的内容学习中达到事半功倍的效果。读者需要注意的是，本书在初稿之前所使用的 Python 版本为 3.6.x。至于原因会在第 1 章中阐述，这里不做过多的解释。

写爬虫的难点不是能否拿下数据，而是在于在实际工作中，整合各种需求业务场景，实现爬虫合理的任务调度、性能优化等。所以这里建议读者在阅读本书时，着重于爬取思路和逻辑方面的思考，不要太过于纠结代码。针对于同一个网站或 App 可以尝试采用不同的策略和解决办法去爬取，观察每一种方法的优缺点并进行总结和积累。当今的反爬技术每天都在更新迭代，将来的爬虫也会越来越难写。但是都万变不离其宗，写爬虫是个研究性的工作，需要每天不断地学习和研究各种案例。希望读者多思考，勤动手。

除了书，您还能得到什么？

(1) 赠送：案例源码。提供与书中相关案例的源代码，方便读者学习参考。

(2) 赠送：Python 常见面试题精选（50 道），旨在帮助读者在工作面试时提升过关率。习题见附录，具体答案参见下方的资源下载。

(3) 赠送：“微信高手技巧随身查”“QQ 高手技巧随身查”“手机办公 10 招就够”3 本电子书，教会读者移动办公诀窍。

(4) 赠送：“5 分钟学会番茄工作法”视频教程。教会读者在职场中高效地工作、轻松应对职场“那些事儿”，真正让读者“不加班，只加薪”！

(5) 赠送：“10 招精通超级时间整理术”视频教程。专家传授 10 招时间整理术，教会读者如何整理时间、有效利用时间。无论是职场还是生活，都要学会时间整理。这是因为时间是人类最宝贵的财富，只有合理整理时间，充分利用时间，才能让读者的人生价值最大化。

温馨提示：以上资源，请用微信扫一扫下方二维码关注公众号，输入代码 HyPc32B 获取学习资源的下载地址及密码。



官方微信公众号



资源下载

本书由凤凰高新教育策划，刘延林老师编写。在本书的编写过程中，我们竭尽所能地为您呈现最好、最全的实用内容，但仍难免有疏漏和不妥之处，敬请广大读者不吝指正。

读者信箱：2751801073@qq.com

目录

Contents

第 1 篇 快速入门篇

第 1 章 Python 基础	2
1.1 Python 环境搭建.....	3
1.1.1 Windows 系统下的 Python 环境安装与配置.....	3
1.1.2 Linux 系统下的 Python 环境安装	7
1.1.3 Mac OS X 系统搭建 Python 3.....	11
1.1.4 IDE 开发工具介绍	13
1.2 Python 编程入门.....	16
1.2.1 第一个 Python 程序.....	16
1.2.2 Python 注释.....	17
1.2.3 数据类型和变量.....	17
1.2.4 字符串和编码.....	19
1.2.5 列表.....	23
1.2.6 元组.....	24
1.2.7 字典.....	25
1.2.8 条件语句.....	25
1.2.9 循环语句.....	26
1.2.10 函数.....	29
1.2.11 类.....	30
1.3 新手实训	33
1.4 新手问答	35
本章小结	35

第 2 章 Python 爬虫入门	36
2.1 爬虫的分类.....	37
2.1.1 通用网络爬虫.....	37
2.1.2 聚焦网络爬虫.....	37
2.1.3 增量式网络爬虫.....	37
2.1.4 深层网络爬虫.....	38
2.2 爬虫的基本结构和 workflow.....	38
2.3 爬虫策略	39
2.3.1 深度优先遍历策略	39
2.3.2 宽度优先遍历策略	40
2.3.3 大站优先策略.....	40
2.3.4 最佳优先搜索策略	40
2.4 HTTP 的基本原理	40
2.4.1 URI 和 URL 介绍	40
2.4.2 超文本.....	41
2.4.3 HTTP 和 HTTPS.....	42
2.4.4 HTTP 的请求过程.....	43
2.5 网页基础	45
2.5.1 网页的组成.....	46
2.5.2 网页的结构.....	48
2.6 Session 和 Cookie	49
2.6.1 Session 和 Cookie 的基本原理.....	49
2.6.2 Session 和 Cookie 的区别.....	51
2.6.3 常见误区.....	51
2.7 新手实训	51
2.8 新手问答	54
本章小结	55
第 3 章 基本库的使用	56
3.1 urllib.....	57
3.1.1 urlopen().....	57
3.1.2 简单抓取网页.....	57
3.1.3 设置请求超时.....	58

3.1.4	使用 data 参数提交数据.....	58
3.1.5	Request.....	59
3.1.6	简单使用 Request.....	60
3.1.7	Request 高级用法.....	61
3.1.8	使用代理.....	62
3.1.9	认证登录.....	62
3.1.10	Cookie 设置.....	63
3.1.11	HTTPResponse.....	63
3.1.12	错误解析.....	64
3.2	requests.....	64
3.2.1	requests 模块的安装.....	65
3.2.2	requests 模块的使用方法介绍.....	65
3.2.3	requests.get().....	65
3.2.4	requests 库的异常.....	67
3.2.5	requests.head().....	68
3.2.6	requests.post().....	68
3.2.7	requests.put() 和 requests.patch().....	68
3.3	re 正则使用.....	69
3.3.1	re.match 函数.....	69
3.3.2	re.search 函数.....	70
3.3.3	re.match 与 re.search 的区别.....	71
3.3.4	检索和替换.....	72
3.3.5	re.compile 函数.....	72
3.3.6	findall 函数.....	74
3.4	XPath.....	75
3.4.1	XPath 的使用方法.....	75
3.4.2	利用实例讲解 XPath 的使用.....	76
3.4.3	获取所有节点.....	77
3.4.4	获取子节点.....	77
3.4.5	获取文本信息.....	77
3.5	新手实训.....	78
3.6	新手问答.....	81
	本章小结.....	82

第 4 章 Ajax 数据抓取	83
4.1 Ajax 简介	84
4.1.1 实例引入	84
4.1.2 Ajax 的基本原理	85
4.1.3 Ajax 方法分析	88
4.2 使用 Python 模拟 Ajax 请求数据	91
4.2.1 分析请求	91
4.2.2 分析响应结果	92
4.2.3 编写代码模拟抓取	92
4.3 新手实训	93
4.4 新手问答	96
本章小结	96
第 5 章 动态渲染页面爬取	97
5.1 Selenium 的使用	98
5.1.1 安装 Selenium 库	98
5.1.2 Selenium 定位方法	99
5.1.3 控制浏览器操作	101
5.1.4 WebDriver 常用方法	102
5.1.5 其他常用方法	104
5.1.6 鼠标键盘事件	104
5.1.7 获取断言信息	107
5.1.8 设置元素等待	109
5.1.9 多表单切换	110
5.1.10 下拉框选择	112
5.1.11 调用 JavaScript 代码	113
5.1.12 窗口截图	113
5.1.13 无头模式	114
5.2 Splash 的基本使用	115
5.2.1 Splash 的功能介绍	115
5.2.2 Docker 的安装	115
5.2.3 Splash 的安装	122
5.2.4 初次实例体验	123

5.2.5	Splash Scripts.....	125
5.3	新手实训	127
5.4	新手问答	131
	本章小结	132
第 6 章	代理的设置与使用	133
6.1	代理设置	134
6.1.1	urllib 代理设置	134
6.1.2	requests 代理设置.....	134
6.1.3	Selenium 代理设置.....	135
6.2	代理池构建	136
6.2.1	获取 IP	137
6.2.2	验证代理是否可用	138
6.2.3	使用代理池.....	139
6.3	付费代理的使用.....	140
6.3.1	讯代理的使用.....	140
6.3.2	阿布云代理的使用	142
6.4	ADSL 拨号代理的搭建.....	145
6.4.1	ADSL 简介	145
6.4.2	购买动态拨号 VPS 云主机	145
6.4.3	测试拨号.....	147
6.4.4	设置代理服务器.....	150
6.4.5	动态获取 IP	152
6.4.6	使用 Python 实现拨号.....	153
6.5	新手问答	155
	本章小结	156
第 7 章	验证码的识别与破解.....	157
7.1	普通图形验证码的识别	158
7.1.1	使用 OCR 进行简单识别.....	158
7.1.2	对验证码进行预处理	159
7.1.3	CNN 验证码识别.....	163
7.2	极验滑动验证码的破解.....	164
7.2.1	分析思路.....	164

7.2.2	使用 Selenium 实现模拟淘宝登录的拖动验证	165
7.2.3	验证修改代码	166
7.3	极验滑动拼图验证码破解	168
7.3.1	分析思路	168
7.3.2	代码实现拖动拼接	169
7.3.3	运行测试	174
7.4	新手问答	174
	本章小结	175
第 8 章	App 数据抓取	176
8.1	Fiddler 的基本使用	177
8.1.1	Fiddler 设置	177
8.1.2	手机设置	178
8.1.3	抓取猎聘网 App 请求包	180
8.2	Charles 的基本使用	182
8.2.1	Charles 安装	183
8.2.2	证书设置	184
8.2.3	手机端配置	186
8.2.4	抓包	188
8.2.5	分析	192
8.2.6	重发	195
8.3	Appium 的基本使用	196
8.3.1	安装 Appium	196
8.3.2	启动 App	200
8.3.3	appPackage 和 appActivity 参数的获取方法	209
8.3.4	Python 代码驱动 App	211
8.3.5	常用 API 方法	213
8.4	新手问答	217
	本章小结	217
第 9 章	数据存储	218
9.1	文件存储	219
9.1.1	TEXT 文件存储	219
9.1.2	JSON 文件存储	220

9.1.3	CSV 文件存储.....	221
9.1.4	Excel 文件存储.....	222
9.2	数据库存储.....	224
9.2.1	MySQL 存储.....	224
9.2.2	MongoDB	228
9.2.3	Redis 存储.....	231
9.2.4	PostgreSQL	233
9.3	新手实训	236
9.4	新手问答	239
	本章小结.....	240

第 2 篇 技能进阶篇

第 10 章	常用爬虫框架	242
10.1	PySpider 框架.....	243
10.1.1	安装 PySpider	243
10.1.2	PySpider 的基本功能	243
10.1.3	PySpider 架构	243
10.1.4	第一个 PySpider 爬虫	244
10.1.5	保存数据到 MySQL 数据库.....	250
10.2	Scrapy 框架.....	252
10.2.1	安装 Scrapy	253
10.2.2	创建项目	253
10.2.3	定义 Item.....	254
10.2.4	编写第一个爬虫 (Spider)	254
10.2.5	运行爬取.....	255
10.2.6	提取 Item.....	255
10.2.7	在 Shell 中尝试 Selector 选择器.....	256
10.2.8	提取数据.....	257
10.2.9	使用 Item.....	258
10.2.10	Item Pipeline.....	260
10.2.11	将 Item 写入 JSON 文件	260

10.2.12 保存到数据库.....	261
10.3 Scrapy-Splash 的使用.....	262
10.3.1 新建项目.....	263
10.3.2 配置.....	263
10.3.3 编写爬虫.....	264
10.3.4 运行爬虫.....	265
10.4 新手实训.....	266
10.5 新手问答.....	269
本章小结.....	269
第 11 章 部署爬虫.....	270
11.1 Linux 系统下安装 Python 3.....	271
11.1.1 安装 Python 3.....	271
11.1.2 安装 virtualenv.....	272
11.2 Docker 的使用.....	273
11.2.1 Docker Hello World.....	273
11.2.2 运行交互式的容器.....	273
11.2.3 启动容器（后台模式）.....	274
11.2.4 停止容器.....	274
11.3 Docker 安装 Python.....	274
11.3.1 docker pull python:3.5.....	275
11.3.2 通过 Dockerfile 构建.....	275
11.3.3 使用 python 镜像.....	277
11.4 Docker 安装 MySQL.....	277
本章小结.....	278
第 12 章 数据分析.....	279
12.1 NumPy 的使用.....	280
12.1.1 NumPy 安装.....	280
12.1.2 NumPy ndarray 对象.....	280
12.1.3 NumPy 数据类型.....	282
12.1.4 数组属性.....	285
12.1.5 NumPy 创建数组.....	288

12.1.6 NumPy 切片和索引	290
12.1.7 数组的运算	291
12.1.8 NumPy Matplotlib	292
12.2 Pandas 的使用	296
12.2.1 从 CSV 文件中读取数据	296
12.2.2 向 CSV 文件中写入数据	297
12.2.3 Pandas 数据帧 (DataFrame)	298
12.2.4 Pandas 函数应用	301
12.2.5 Pandas 排序	303
12.2.6 Pandas 聚合	306
12.2.7 Pandas 可视化	309
12.3 pyecharts 的使用	311
12.3.1 绘制第一个图表	311
12.3.2 使用主题	313
12.3.3 使用 pyecharts-snapshot 插件	313
12.3.4 图形绘制过程	313
12.3.5 多次显示图表	314
12.3.6 Pandas&NumPy 简单示例	314
12.4 新手实训	315
12.5 新手问答	316
本章小结	316

第 3 篇 项目实战篇

第 13 章 爬虫项目实战	318
13.1 实战一：Selenium+XPath 爬取简书	319
13.1.1 打开简书首页分析	319
13.1.2 爬取思路	321
13.1.3 编写爬虫代码	321
13.1.4 实例总结	325
13.2 实战二：使用 requests 爬取腾讯人口迁徙数据	326
13.2.1 分析网页结构	326

13.2.2	爬取思路.....	328
13.2.3	动手编码实现爬取	328
13.2.4	实例总结.....	330
13.3	实战三：Scrapy 爬取豆瓣电影	330
13.3.1	分析豆瓣电影网页结构	330
13.3.2	爬取的数据结构定义（items.py）	332
13.3.3	爬虫器（MovieSpider.py）	332
13.3.4	pipeline 管道保存数据.....	333
13.3.5	将数据存储到 MySQL 数据库	333
13.3.6	实例总结.....	334
13.4	实战四：使用 Selenium 多线程异步爬取同城旅游网机票价格信息	334
13.4.1	分析同城旅游网.....	334
13.4.2	编码实现抓取数据	336
13.4.3	实例总结.....	343
13.5	实战五：数据分析 Dessert Apples 下 12 种苹果全年最高、最低和平均销量	343
13.5.1	Pandas 读取数据	344
13.5.2	获取索引，drop_duplicates() 去重.....	344
13.5.3	实现分析数据.....	345
13.5.4	实例总结.....	346
13.6	实战六：中国南方航空机票信息爬取	346
13.6.1	分析中国南方航空网	347
13.6.2	编写代码进行爬取	349
13.6.3	实例总结.....	352
	本章小结	352

附录	Python 常见面试题精选	353
----	----------------------	-----