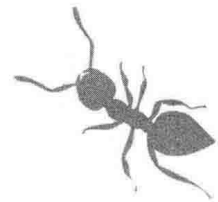
Python 网络爬虫开发从入门到精通

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

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?

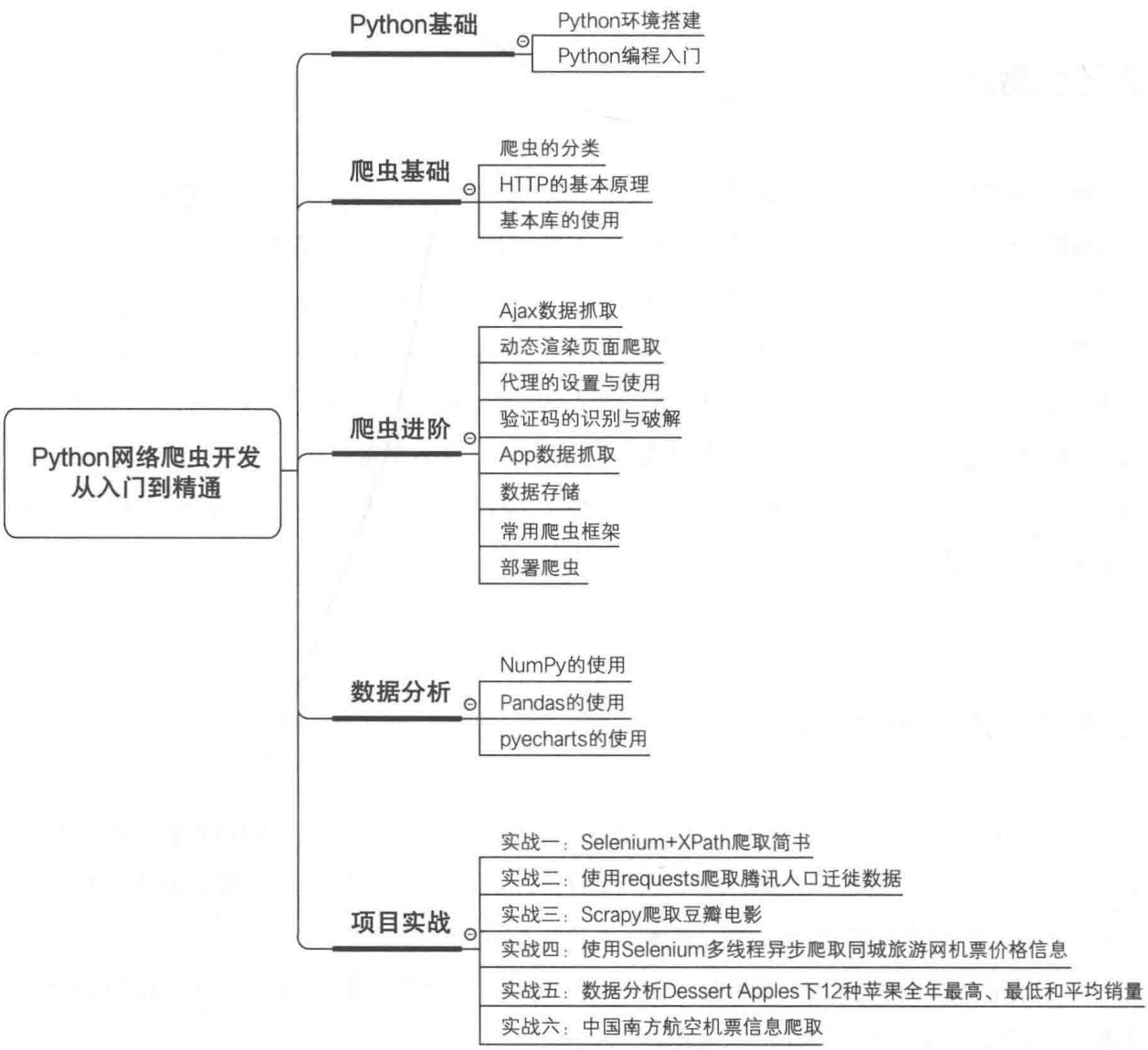
本书力求简单、实用。坚持以实例为主，理论为辅的路线。全书共 13 章，从 Python 基础、爬虫开发常用网络请求库，到爬虫框架使用和分布式爬虫设计，以及最后的数据存储、分析、实战训练等，覆盖了爬虫项目开发阶段的整个生命周期。整体上本书内容有以下几个特点。

（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 2 篇 技能进阶篇