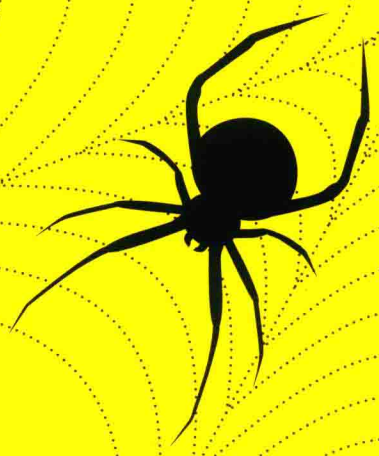
科大讯飞人工智能教育培训与研究专家多年实战经验的总结

手把手带领“小白”从零开始全面掌握Scrapy网络爬虫的核心技术

以案例为导向，通过对案例的不断迭代和优化，加深读者对知识的理解



从零开始学 Scrapy网络爬虫

（视频教学版）

张涛◎编著

- **超值配书资料：17小时配套教学视频、案例源代码、教学PPT**
- 全面涵盖Python基础、爬虫原理、Scrapy框架、数据库存储、动态页面爬取、模拟登录、反爬虫技术、文件和图片下载、分布式爬虫等内容
- 选用多个知名且有代表性的网站作为爬取目标，有很强的实用性和可操作性
- 详解14个爬虫综合案例，并重点剖析抢票软件项目的实现原理及实现过程，提高读者解决实际问题的能力

机械工业出版社
China Machine Press

从零开始学 Scrapy网络爬虫

（视频教学版）

张涛◎编著

机械工业出版社
China Machine Press

图书在版编目（CIP）数据

从零开始学Scrapy网络爬虫：视频教学版/张涛编著. —北京：机械工业出版社，2019.9

ISBN 978-7-111-63474-4

Ⅰ. 从… Ⅱ. 张… Ⅲ. 软件工具－程序设计 Ⅳ. TP311.561

中国版本图书馆CIP数据核字（2019）第181393号

本书从零开始，循序渐进地介绍了目前最流行的网络爬虫框架 Scrapy。即使你没有任何编程基础，阅读本书也不会有压力，因为书中有针对性地介绍了 Python 编程技术。另外，本书在讲解过程中以案例为导向，通过对案例的不断迭代、优化，让读者加深对知识的理解，并通过 14 个项目案例，提高读者解决实际问题的能力。

本书共 13 章。其中，第 1～4 章为基础篇，介绍了 Python 基础、网络爬虫基础、Scrapy 框架及基本的爬虫功能。第 5～10 章为进阶篇，介绍了如何将爬虫数据存储于 MySQL、MongoDB 和 Redis 数据库中；如何实现异步 AJAX 数据的爬取；如何使用 Selenium 和 Splash 实现动态网站的爬取；如何实现模拟登录功能；如何突破反爬虫技术，以及如何实现文件和图片的下载。第 11~13 章为高级篇，介绍了使用 Scrapy-Redis 实现分布式爬虫；使用 Scrapyd 和 Docker 部署分布式爬虫；使用 Gerapy 管理分布式爬虫，并实现了一个抢票软件的综合项目。

本书适合爬虫初学者、爱好者及高校相关专业的学生阅读，也适合数据爬虫工程师作为参考读物，同时还适合各大院校和培训机构作为教材使用。

从零开始学 Scrapy 网络爬虫（视频教学版）

出版发行：机械工业出版社（北京市西城区百万庄大街 22 号 邮政编码：100037）

责任编辑：欧振旭 李华君 责任校对：姚志娟

印 刷：中国电影出版社印刷厂 版 次：2019 年 9 月第 1 版第 1 次印刷

开 本：186mm×240mm 1/16 印 张：18.75

书 号：ISBN 978-7-111-63474-4 定 价：99.00 元

客服电话：（010）88361066 88379833 68326294 投稿热线：（010）88379604

华章网站：www.hzbook.com 读者信箱：hzit@hzbook.com

作者简介

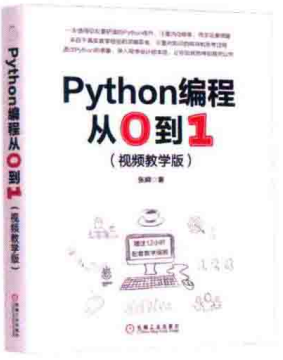
张涛 毕业于中国科学技术大学，获硕士学位。目前在科大讯飞公司从事人工智能教育培训与研究工作。加入科大讯飞公司之前，曾经在知名的日资企业任研发经理，负责日本大型证券系统的设计与开发。有7年大学课程改革与教学经验，主要研究方向为Python网络爬虫、数据分析和机器学习。

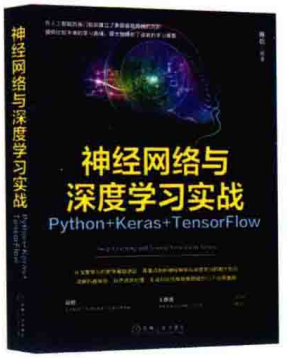
超值配书资源

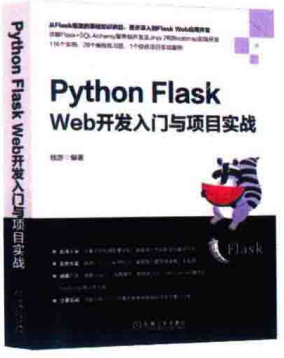
- 17小时配套教学视频
- 案例源代码文件
- 教学PPT

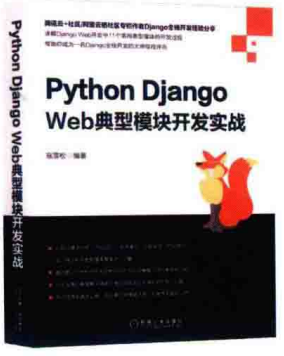
本书配套资源需要读者自行下载，请在www.hzbook.com网站上搜索到本书，然后单击“资料下载”按钮，即可在本书页面上找到下载链接。

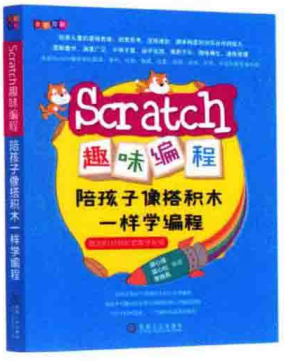
推荐阅读

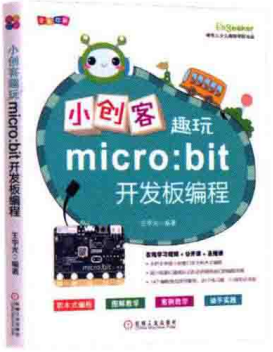












欢迎IT领域的各位技术专家洽谈出版事宜。如果有写书或投稿意向，请加QQ或者微信具体商谈。

QQ：627173439　　微信：oyzx_xp

前言

随着人工智能浪潮的到来，笔者身边有越来越多的人投入到人工智能和大数据的学习与研究中。他们来自不同的行业，有高校老师和学生，有 AI 研究专家，有物理或数学专业人才。他们都迫切希望能够获取大量相关领域的数据，用于学习和研究。而互联网中源源不断的海量数据为他们提供了一个既经济又可靠的来源。如何简单、高效、快捷地获取这些数据呢？笔者试图为他们推荐几本能快速入手的书籍。经过一番了解，发现目前市场上关于网络爬虫的图书主要分为两类：一类是翻译成中文的外版图书，其定位相对高端，且翻译质量参差不齐，阅读难度较大，不易上手，故不适合初学者学习；另一类是国内原创的一些关于网络爬虫的图书，这些书大多要求读者具备一定的 Python 编程基础，虽然书中对各种网络爬虫框架都有介绍，但是不深入也不成体系，对于零基础或非计算机专业的人员来说，显然也不太适合。

于是，他们就“怂恿”我，希望我能编写一本从零基础开始学起的网络爬虫书籍。虽然我从事网络爬虫教学工作多年，但我深知教学跟写书是两码事。教学注重临场发挥，思维比较发散；而写书要求文笔流畅、逻辑严谨缜密。我实在没有信心接受这个挑战。直到有一天，机械工业出版社的编辑联系到了我，认为我从事教育和研究工作，能讲、会说、有技术，对写书来说正是最大的优势。于是在编辑的鼓励和指导下，我开始构思和梳理文章脉络：首先，本书受众要广，即使是零基础或非计算机专业的“小白”也能上手；其次，本书内容不追求多和杂，只选用最流行、最好用、最强大的网络爬虫框架介绍即可；最后，本书的可操作性和实用性要强，通过迭代案例加深读者对知识的理解与应用，以典型的、知名的网站为爬取目标，提高读者解决实际问题的能力。本书正是遵循这样的思路逐步推进，不断优化，最后顺利地完成了写作。

本书有何特色

1. 由浅入深，循序渐进

本书从零开始，先介绍 Python 语言、网络爬虫基础、Scrapy 框架结构等基础内容；再介绍 Scrapy 的数据库存储、动态页面爬取、突破反爬虫技术等核心技术；接着介绍分布式爬虫的实现、部署和管理等高级技术；最后介绍了一个完整的综合项目的开发过程。

2. 视频教学，讲解详尽

为了便于读者高效、直观地学习，书中每一章的重点内容都专门录制了配套教学视频。

读者可以将图书内容和教学视频结合起来，深入、系统地学习，相信一定会取得更好的学习效果。

3．注释详细，一目了然

无论是在 Python 程序设计，还是在 Scrapy 爬虫实现部分，本书均对代码做了详细的注释，读者理解起来会更加顺畅。另外，对于多步骤的操作过程，本书在图例中使用数字做了标注，便于读者准确操作。

4．案例丰富，实用易学

本书提供了 14 个实用性很强的项目案例，这些案例爬取的目标均是知名的、具有代表性的、应用价值较高的网站。读者通过实际操练这些项目案例，可以更加透彻地理解 Scrapy 网络爬虫的相关知识。

5．提供课件，方便教学

笔者专门为本书制作了专业的教学 PPT，以方便相关院校或培训机构的教学人员讲课时使用。

本书内容

第1篇　基础篇

第 1 章　Python 基础

本章介绍了 Python 环境搭建，并详细介绍了 Python 基本语法、Python 内置数据结构及 Python 模块化设计，为 Scrapy 网络爬虫开发打下坚实的编程基础。

第 2 章　网络爬虫基础

本章介绍了与网络爬虫技术相关的 HTTP 基本原理、网页基础，以及使用 XPath 提取网页信息的方法，为 Scrapy 网络爬虫开发打下坚实的理论基础。

第 3 章　Scrapy 框架介绍

本章首先介绍了网络爬虫的原理；然后介绍了 Scrapy 框架的结构及执行流程，并实现了 Scrapy 的安装；最后结合案例，实现了第一个 Scrapy 网络爬虫功能。

第 4 章　Scrapy 网络爬虫基础

本章深入 Scrapy 框架内部，介绍了使用 Spider 提取数据、使用 Item 封装数据、使用 Pipeline 处理数据的方法，并通过一个项目案例，演示了一个功能完备的 Scrapy 项目的实现过程。

第2篇　进阶篇

第 5 章　数据库存储

本章介绍了关系型数据库 MySQL、非关系型数据库 MongoDB 和 Redis 的下载、安装及基本操作，并通过 3 个项目案例，实现了将爬取来的数据分别存储于这 3 个数据库中的方法。

第 6 章　JavaScript 与 AJAX 数据爬取

本章通过两个项目案例，介绍了使用 Scrapy 爬取通过 JavaScript 或 AJAX 加载的数据的方法和技巧。

第 7 章　动态渲染页面的爬取

本章介绍了使用 Selenium 和 Splash 这两个工具来模拟浏览器进行数据爬取的方法，并通过两个项目案例，进一步巩固使用 Selenium 和 Splash 的方法与技巧。

第 8 章　模拟登录

本章介绍了某些需要登录才能访问的页面爬取方法，并介绍了模拟登录、验证码识别和 Cookie 自动登录等知识，还通过一个项目案例，进一步巩固了实现模拟登录的方法和技巧。

第 9 章　突破反爬虫技术

本章介绍了突破反爬虫的几种技术，主要有降低请求频率、修改请求头、禁用 Cookie、伪装成随机浏览器及更换 IP 地址等，通过这些举措，可以有效避免目标网站的侦测，提高爬虫成功率。

第 10 章　文件和图片下载

本章介绍了使用 Scrapy 的中间件批量下载文件和图片的方法，并通过两个项目案例，进一步巩固了文件和图片下载的方法与技巧。

第3篇　高级篇

第 11 章　Scrapy-Redis 实现分布式爬虫

本章介绍了使用 Scrapy-Redis 实现分布式爬虫的方法。首先介绍了分布式爬虫的原理，然后介绍了实现分布式爬虫的思路和核心代码，最后通过一个图片下载的项目案例，构造了一个分布式爬虫系统。

第 12 章　Scrapyd 部署分布式爬虫

本章介绍了分布式系统的部署和管理。首先介绍了使用 Scrapyd 和 Scrapyd-Client 部署分布式爬虫，然后介绍了使用 Docker 批量部署分布式爬虫，最后介绍了如何使用 Gerapy 管理分布式爬虫。

第 13 章　综合项目：抢票软件的实现

本章通过全面分析 12306 购票网站的特点，结合 Scrapy 网络爬虫框架和 Selenium 浏

览器工具，使用 Python 面向对象的设计模式，完成了一个综合性和实用性都较强的项目：抢票软件。

本书配套资源获取方式

本书涉及以下配套资源：

- 配套教学视频；
- 实例源代码文件；
- 教学 PPT。

这些配套资源需要读者自行下载。请登录华章公司网站 www.hzbook.com，在该网站上搜索到本书，然后单击“资料下载”按钮，在本书页面上找到下载链接即可下载。

适合阅读本书的读者

- 网络爬虫初学者；
- 网络爬虫爱好者；
- 网络爬虫从业人员；
- 数据工程师；
- 高等院校的老师和学生；
- 相关培训机构的学员。

本书作者

笔者毕业于中国科学技术大学软件工程专业，获硕士学位。现就职于知名的智能语音技术公司，有 10 余年软件项目管理经验。在高等院校担任网络爬虫及机器学习方面的授课工作。

本书能够顺利出版，首先要感谢本书编辑欧振旭！他花费了大量时间和精力对本书提出了有价值的修改意见和建议；还要感谢其他为本书的出版提供过帮助的编辑和朋友！没有他们的大力支持，本书也很难与读者见面。

由于笔者水平所限，加之成书时间有限，书中可能还存在一些疏漏和不当之处，敬请各位读者斧正。联系邮箱：hzbook2017@163.com。

张涛

目录

第 2 篇　进阶篇

第 3 篇 高级篇

第1篇 基础篇