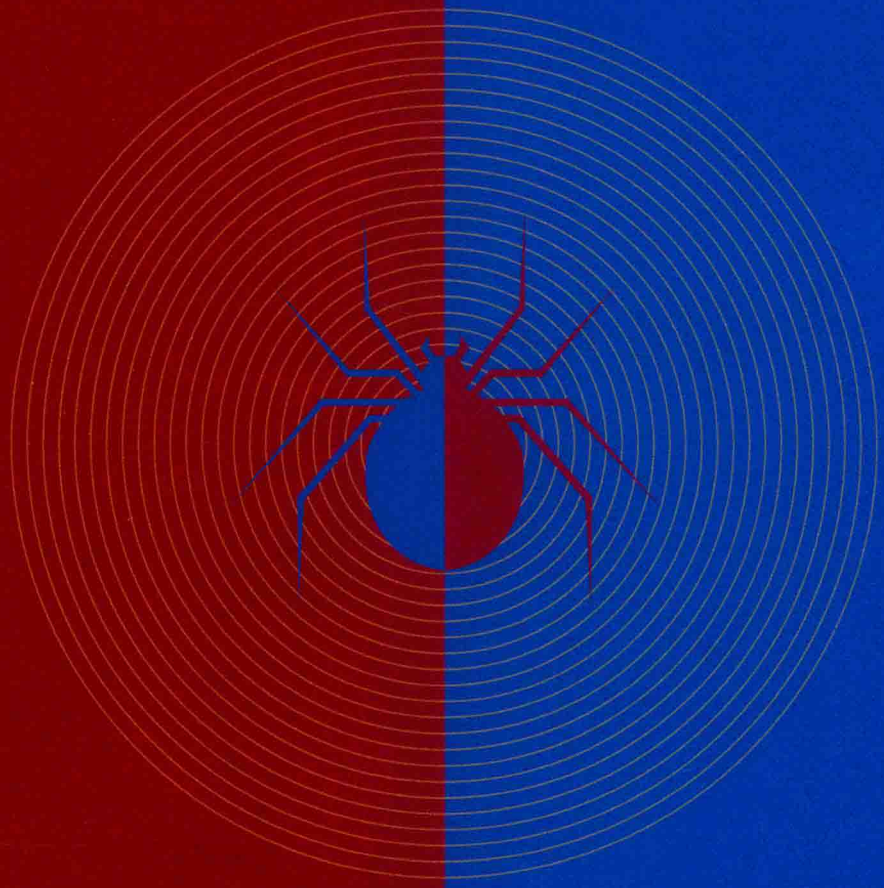
TURING 图灵原创

Python 3

反爬虫原理与绕过实战

韦世东◎著



中国工信出版集团

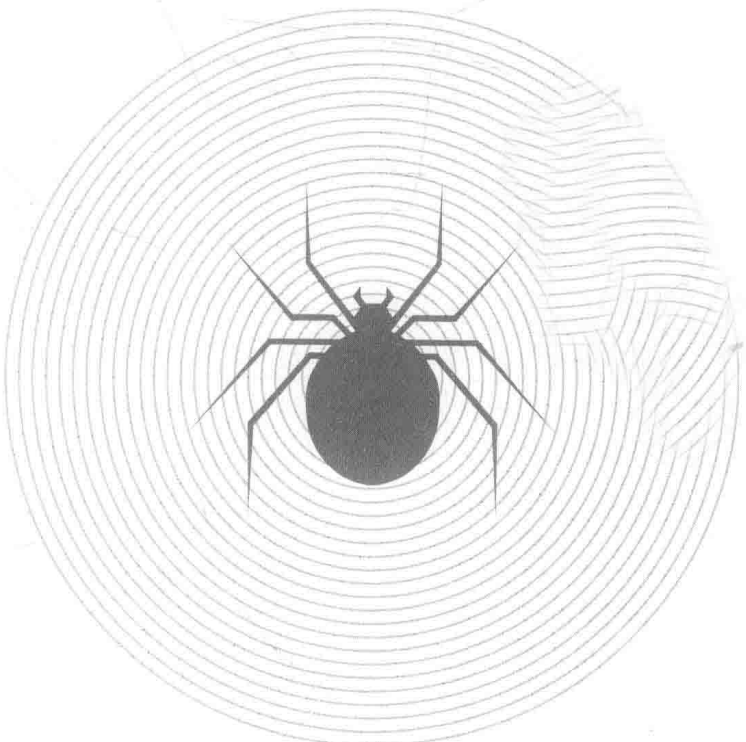
人民邮电出版社
POSTS & TELECOM PRESS

TURING 图灵原创

Python 3

反爬虫原理与绕过实战

韦世东◎著



人民邮电出版社
北 京

图书在版编目（CIP）数据

Python 3反爬虫原理与绕过实战 / 韦世东著. — 北京 : 人民邮电出版社, 2020.1
（图灵原创）
ISBN 978-7-115-52873-5

Ⅰ. ①P… Ⅱ. ①韦… 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268678号

内 容 提 要

本书描述了爬虫技术与反爬虫技术的对抗过程，并详细介绍了这其中的原理和具体实现方法。首先讲解开发环境的配置、Web 网站的构成、页面渲染以及动态网页和静态网页对爬虫造成的影响。然后介绍了不同类型的反爬虫原理、具体实现和绕过方法，还涉及常见验证码的实现过程，并使用深度学习技术完成了验证。最后介绍了常见的编码和加密原理、JavaScript 代码混淆知识、前端禁止事件以及与爬虫相关的法律知识和风险点。

本书既适合需要储备反爬虫知识的前端工程师和后端工程师阅读，也适合需要储备绕过知识的爬虫工程师、爬虫爱好者以及 Python 程序员阅读。

◆ 著　　韦世东
责任编辑　王军花
责任印制　周昇亮
◆ 人民邮电出版社出版发行　　北京市丰台区成寿寺路11号
邮编　100164　　电子邮件　315@ptpress.com.cn
网址　http://www.ptpress.com.cn
北京市艺辉印刷有限公司印刷
◆ 开本：800×1000　1/16
印张：24.5　　彩插：2
字数：565千字　　2020年 1 月第 1 版
印数：1－4 000册　　2020年 1 月北京第 1 次印刷

定价：89.00元

读者服务热线：(010)51095183转600　印装质量热线：(010)81055316
反盗版热线：(010)81055315
广告经营许可证：京东工商广登字 20170147 号

站在巨人的肩膀上

Standing on Shoulders of Giants

iTuring.cn

站在巨人的肩膀上

Standing on Shoulders of Giants

iTuring.cn

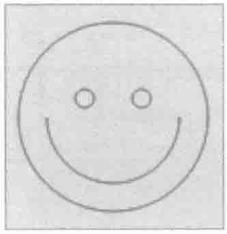


图 7-15　Canvas 绘制的笑脸图案

图 8-6　Charles 工具栏

图 9-1　示例 15 的页面

图 9-2　灰度处理后的图片

图 9-3　二值化处理后的验证码图片

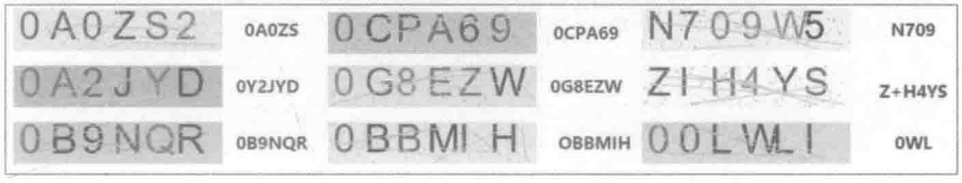


图 9-4　腾讯 OCR 识别结果

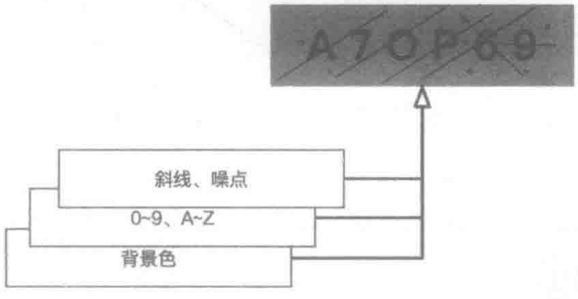


图 9-5　字符验证码的组成

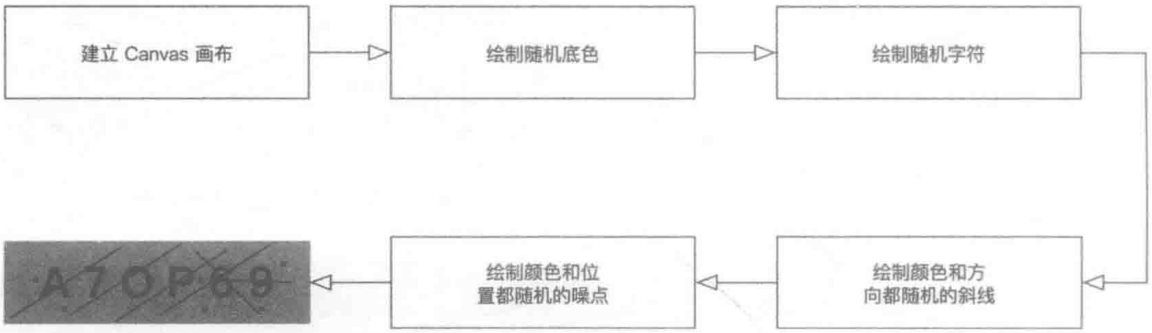


图 9-6　验证码的绘制流程

图 9-7　验证码

图 9-8　验证码

图 9-28　增加了干扰信息的计算型验证码

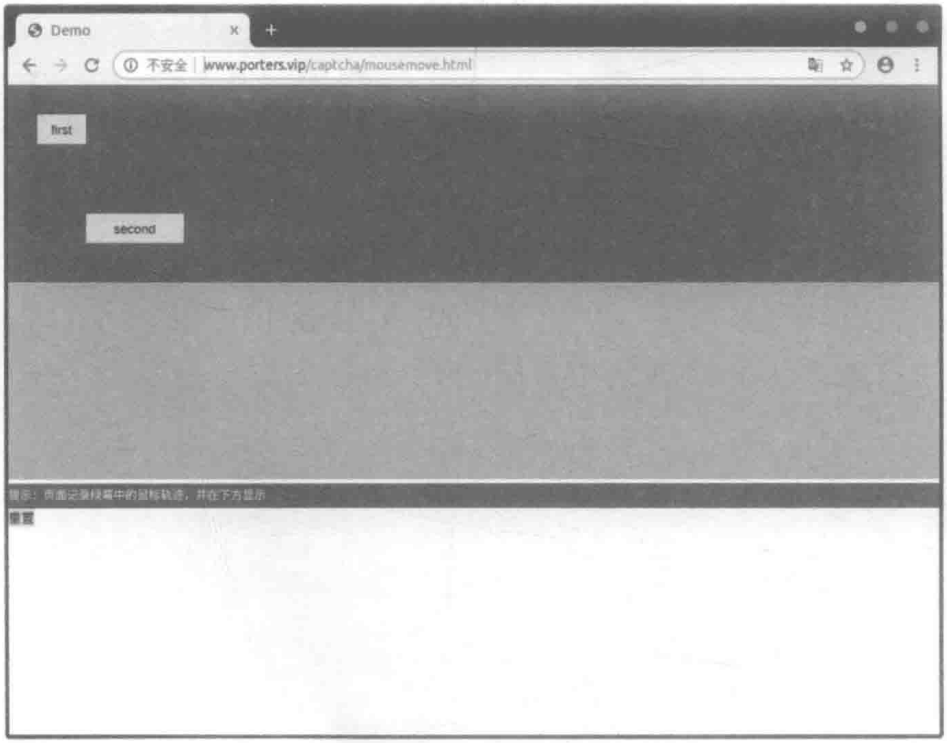


图 9-63　示例 21 页面

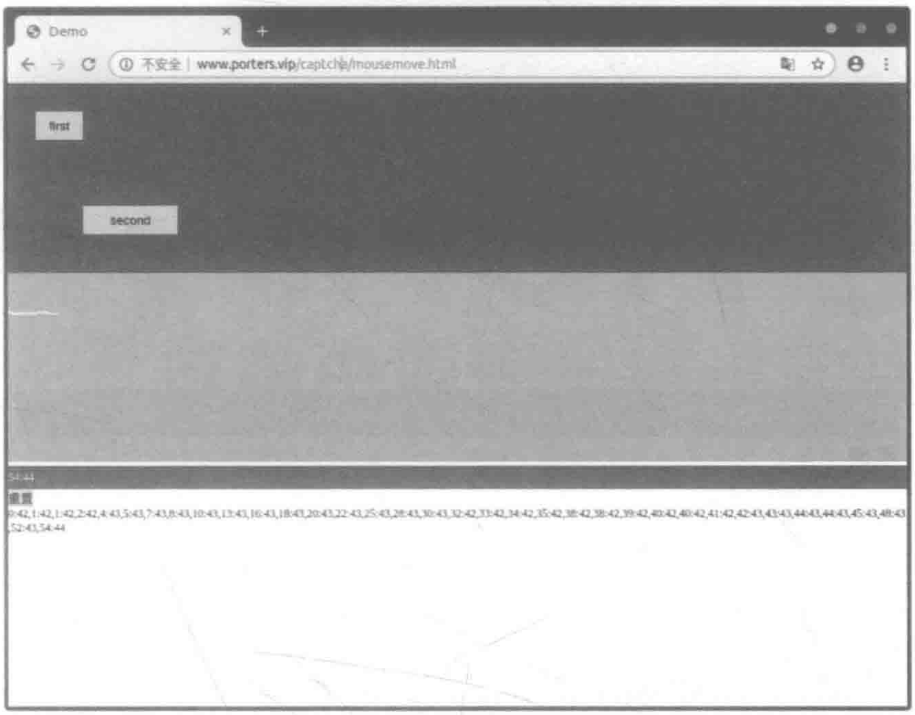


图 9-64　鼠标移动轨迹和坐标信息

图 9-65　Selenium 套件点击按钮时的轨迹和鼠标坐标信息

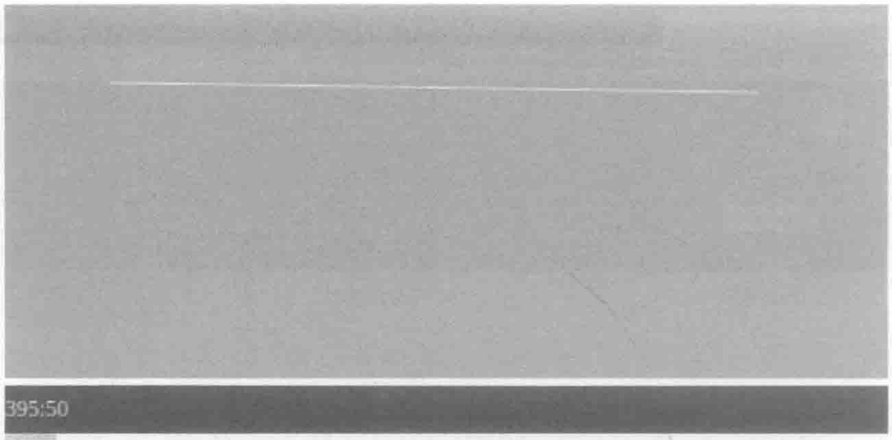


图 9-66　Selenium 套件执行滑动操作产生的鼠标轨迹和坐标信息

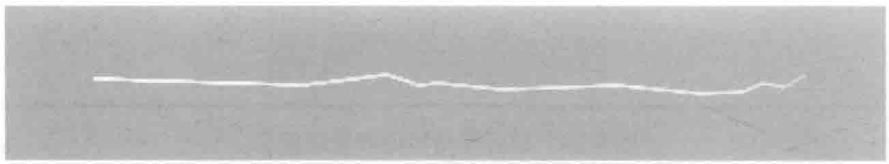


图 9-67　Selenium 套件模拟手臂晃动产生的鼠标轨迹

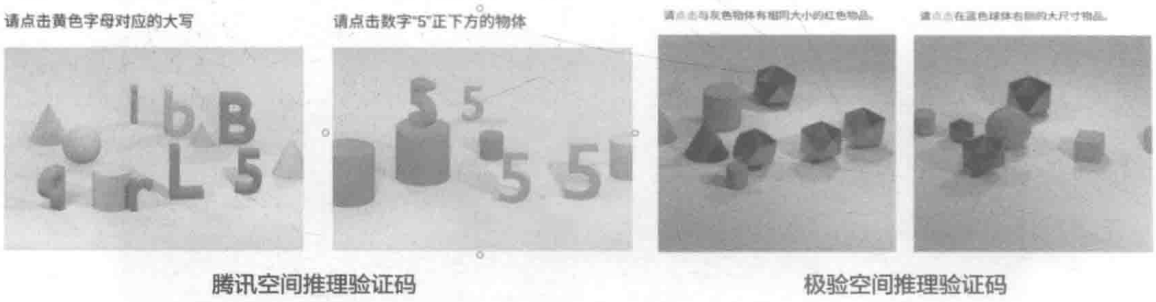


图 9-72　空间推理验证码产品及所属公司

序

我们正处于信息爆炸的大数据时代，数据在互联网上的传播和呈现方式多种多样，如何获取这些杂乱的数据呢？爬虫就是其中的一种方式。与此同时，在这茫茫的数据大海中，高质量的、整洁的数据变得越来越重要，这些数据甚至就是一个公司生存的支柱。要保护这些数据，不让它们被轻易爬走，反爬虫技术应运而生。

近几年，爬虫技术和反爬虫技术在不断斗争的过程中变得越来越高深和复杂。从简单的 User-Agent 识别到混淆验证码加密，“花样”越来越多，破解难度也越来越大，爬虫工程师和反爬虫工程师每天斗智斗勇，殚精竭虑。

知己知彼方能百战不殆。不论是爬虫工程师还是反爬虫工程师，如果想要把自己的方向做好，就需要对这两个方向的技术都有一定程度的研究。譬如拿爬虫工程师来说，如果对反爬虫的知识知其然而不知其所以然，势必会对反爬虫的绕过力有不逮。所以，双方都需要对爬虫和反爬虫技术有一定程度的了解。然而从目前来看，市面上还没有一本系统讲解爬虫和反爬虫技术的书。

我的好友韦世东是一名高级爬虫工程师，对各种爬虫和反爬虫的技巧进行过深入的研究。书中首先对各类反爬虫技术进行了合理的归类，然后通过剖析多个案例帮助大家理解各类反爬虫技术的原理。内容包括但不限于 Cookie 反爬虫、WebSocket 反爬虫、字体反爬虫、WebDriver 反爬虫、App 反爬虫、验证码反爬虫，几乎涵盖了市面上所有的反爬虫技术类型，内容十分详尽，另外他还针对各类反爬虫给出了对应的绕过和破解方案。通过本书，大家可以全面了解到爬虫和反爬虫的各类技术。本书干货满满，强烈推荐给大家。

崔庆才

微软小冰工程师

《Python 3 网络爬虫开发实战》作者

前言

爬虫是当今互联网使用非常广泛的技术之一，现已应用于金融、房产、贸易与科技等诸多领域。无论是大数据计算、数据分析还是机器学习，都离不开爬虫。爬虫工作在很多时候是企业业务开展的基础与主线，将爬取内容进行清洗和处理，得到的就是极具价值的数据。

爬虫技术的门槛低，易于学习，因此成为初学者用来上手的学习对象。大数据和深度学习都需要大量的数据来支撑，而爬虫是目前较好的数据来源手段。随着这几年大数据和深度学习的火热，爬虫技术的发展进入了高峰期，由此给服务器带来的压力也成倍增长。

企业为了保证服务器的正常运转，或者为了降低服务器的运转压力与成本，不得不使出各种各样的技术手段来阻止爬虫工程师们毫无节制地向服务器索取资源，我们将这种行为称为反爬虫。反爬虫技术是互联网技术中为了限制爬虫而产生的技术总称。无论是在技术学习还是实际工作中，反爬虫技术都是所有爬虫工程师要面对的问题。常见的反爬虫原理和绕过技巧也是中高级爬虫工程师，尤其是在大型互联网企业的中高级爬虫工程师面试中关注的焦点。同样，作为一名开发者，了解反爬虫原理和绕过技巧有助于设计出更合理的反爬虫策略，会使你从同行中脱颖而出，大放异彩。

在平时的交流中，我发现很多朋友对于验证码识别、JavaScript 混淆、WebSocket 和字体反爬虫有一种莫名的恐惧感，觉得这些是很难解决的问题。实际上，只要我们了解其工作原理，就能够找到突破口。爬虫与反爬虫都是综合知识的应用，单纯了解某个反爬虫的实现方法或绕过技巧是不够的，我们应该深入了解其实现原理，这样才能够在爬虫工程师的职业道路上走得更远。

我希望通过梳理并总结以往工作中的经验，结合练习平台 Steamboat，帮助更多的爬虫工程师和开发者了解并掌握反爬虫技术与反爬虫绕过的技能。

本书案例均来自于实际的项目，大部分是国内知名互联网企业在用的反爬虫手段。由于爬虫技术的更新速度非常快，为了保证大家的学习质量，本书为读者准备了一个练习平台。书中介绍到的所有反爬虫示例均收录在练习平台中。大家只需要跟着书本指引操作，就可以在个人计算机或云服务器上搭建练习平台，这部分内容会在第 1 章中介绍。

反爬虫和绕过技巧涉及的知识点非常多，且跨度较大，本书主要讲解其中的原理和实践应用，让大家在学习之后可以快速将所学知识应用到实际工作当中。除此之外，还会讲解一些网络传输相关的

知识以及一些工具的用法等。

以剑养剑，攻守兼备才能够在技术的江湖路上任逍遥。

阅读建议

这是一本围绕着反爬虫原理展开的书，书中提到了浏览器的基本结构、网页渲染原理、加密和混淆规范，还有很多 RFC 文档（Request For Comments，一系列以编号排定的互联网协议和标准文件）。RFC 文档分为提议性的、部分在用的和正式标准。无论是开发者还是爬虫工程师，熟读常见的 RFC 文档对工作会有很大的帮助。

动手实践很重要，这不仅能让你掌握书本知识，而且还有可能在练习中有新的发现。为此，本书为读者准备了一个练习平台，其中包含 21 个示例。练习平台上的示例均为本书作者编写，且与本书示例一一对应。因此，示例内容不会改动，并且无须担心相关的法律问题，这保证了大家的学习能顺利进行。

本书共 10 章，从开发环境配置到原理，再到实际的反爬虫案例剖析，内容循序渐进。建议读者按照章节顺序阅读，并在阅读过程中亲自动手练习，巩固所学知识。

本书内容

本书共 10 章，章节内容归纳如下。

- 第 1 章介绍了本书所涉及的大部分开发环境配置。本章无须完整阅读，在需要时查阅即可。
- 第 2 章介绍了 Web 网站的构成和页面渲染方面的知识。了解服务器端、客户端的组成，工作形式和通信协议，这会为我们后面的学习打下坚实的基础。
- 第 3 章简单讲述了动态网页和静态网页对爬虫造成的影响。回顾了一些爬虫方面的基本概念和知识，并对反爬虫这一概念进行了介绍和约定。
- 第 4 章以信息校验型反爬虫为主线，讲解了基于 HTTP 协议和 WebSocket 协议对客户端请求进行校验的反爬虫原理和具体实现方法，并以爬虫工程师的角度演示了绕过过程。
- 第 5 章介绍了常见的动态渲染反爬虫，深入了解其原理，并介绍了几种应对方法和多种渲染工具的基本用法。这一章通过场景假设的方式来讲解不同需求的应对方法。
- 第 6 章介绍了目前被广泛使用的文本混淆反爬虫知识，包括图片伪装、CSS 偏移、SVG 映射和字体反爬虫等。每个案例均以爬虫工程师的角度演示绕过过程，再剖析其原理。最后讨论了文本混淆反爬虫的通用解决方法。

- 第 7 章介绍了特征识别反爬虫，包括绕过过程和实现原理。相对其他反爬虫手段来说，特征识别反爬虫具有一定的隐蔽性。它在爬虫程序发起时对其进行识别和过滤，这能够有效地减轻服务器的压力。
- 第 8 章介绍了 App 数据爬取的关键和常用的反爬虫手段，包括代码混淆、参数加密和安全加固等，同时还介绍了抓包和 App 逆向方面的知识。
- 第 9 章是验证码相关的内容，包含市面上常见的验证码类型，例如字符验证码、计算型验证码和行为验证码。每个验证码案例均以爬虫工程师的角度演示绕过过程，再以开发者的角度演示验证码的实现过程。部分验证码的绕过用到了深度学习中的卷积神经网络和用于目标检测的 YOLO 算法。在最后一节中，我们对商用验证码厂商的产品进行了基本介绍和难度分析。
- 第 10 章是综合知识的介绍。首先介绍了常见的编码和加密原理，并以对应的 RFC 文档为基础，讲解编码、解码、加密和解密的过程。然后介绍了常见的 JavaScript 代码混淆知识，讲解了混淆原理和还原技巧，并动手实现了一个简单的混淆器。接着学习了前端禁止事件方面的知识，如禁止鼠标右键、禁止键盘按键等。最后通过几个案例了解了与爬虫相关的法律知识和风险点，并列出了《数据安全管理办法（征求意见稿）》中与爬虫相关的条例。

致谢

本书的顺利编写，得益于家人和朋友的帮助。首先感谢我的家人，我的爸爸妈妈、岳父岳母、夫人、妹妹和我的女儿。有了他们的支持，我才能用心写作。

特别感谢崔庆才（静觅）在我学习路上和写作期间给予的帮助。没有他的支持和帮助，我的进步也不可能这么快。他是我奋力追赶的目标，也是我前进的方向。

感谢唐轶飞（大鱼）为我解决学习路上遇到的问题。当我还是一个“萌新”的时候，是他给我解疑答惑，使我少走弯路，顺利成长。

感谢陈祥安（cxa）与我共同学习、共同进步。他是一个乐于奉献的人，常把新的学习材料和知识分享给我，让我保持对新技术的研究热情。

感谢我的前同事李宏强、我的师弟盘启强和我的妹妹韦东慧。他们参与了书中部分案例的编写，并在写作过程中提供了很多帮助。也正是有了他们的帮助，这本书的内容才变得如此精彩。

感谢在我学习过程中与我探讨技术的各位朋友，QQ 群群友和微信群群友，他们对技术的研究和原理探究的精神带动着我，使我学到不少知识。

感谢掘金社区为本书提供的支持。

感谢王军花编辑，她在书稿立项和写作过程中给我提供了很多建议，这正是本书内容如此流畅的原因。

感谢在我学习之路和写作过程中提供帮助的每一个人！

免责声明

爬虫技术是一把双刃剑。本书的写作初衷是希望读者将本书学到的技术用于防护，提高应用防护等级。本书中的所有内容仅供技术学习与研究，请勿将本书讲解的反爬虫绕过方法和技巧用于非法用途。

相关资源

书中用到的部分代码存放在 GitHub（详见 https://github.com/asyncins/antispider）[①]，代码与章节内容的对应关系可查阅仓库中的 README.md 文件。

我是一个爬虫工程师，同时也是 Python 开发者和 Rust 开发者。我会在微信公众号和技术博客中更新相关的技术文章，欢迎读者访问交流。当然，大家也可以添加我的微信，期待和你共同进步，一起变强！


夜幕团队


进击的 Coder


算法和反爬虫


韦世东微信

韦世东

2019 年 6 月

① 本书代码也可从图灵社区（iTuring.cn）本书主页免费注册下载。

目录