



Python 网络爬虫实战

吕云翔 张扬 © 编著

- 完整的项目案例
- 视频讲解
- 课件与源码



內容簡介

本書在編寫時，深入研讀了最新的Python 3.5，以及相關的企業級應用，包括如何構建高性能的Web服務器，以及如何利用Python進行數據分析、機器學習等。本書還介紹了Python在網絡安全、網絡爬虫、網絡釣魚等方面的應用。本書適合Python初學者閱讀，也適合對Python感興趣的讀者閱讀。



Python 网络爬虫实战

吕云翔 张扬 © 编著

清華大學出版社
北京

内 容 简 介

本书介绍如何利用 Python 进行网络爬虫程序的开发,从 Python 语言的基本特性入手,详细介绍了 Python 爬虫开发的相关知识,涉及 HTTP、HTML、JavaScript、正则表达式、自然语言处理、数据科学等内容。全书共分为 14 章,包括 Python 基础知识、网站分析、网页解析、Python 文件的读写、Python 与数据库、AJAX 技术、模拟登录、文本与数据分析、网站测试、Scrapy 爬虫框架、爬虫性能等多个主题,内容覆盖网络抓取与爬虫编程中的主要知识和技术,在重视理论基础的前提下从实用性和丰富度出发,结合实例演示了编写爬虫程序的核心流程。

本书适合 Python 语言初学者、网络爬虫技术爱好者、数据分析从业人员以及高等院校计算机科学、软件工程等相关专业的师生阅读。

本书封面贴有清华大学出版社防伪标签,无标签者不得销售。

版权所有,侵权必究。侵权举报电话:010-62782989 13701121933

图书在版编目(CIP)数据

Python 网络爬虫实战/吕云翔,张扬编著. —北京:清华大学出版社,2019
(清华科技大讲堂)
ISBN 978-7-302-51592-0

I. ①P… II. ①吕… ②张… 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(2018)第 257255 号

策划编辑:魏江江
责任编辑:王冰飞
封面设计:刘 键
责任校对:时翠兰
责任印制:丛怀宇

出版发行:清华大学出版社

网 址: <http://www.tup.com.cn>, <http://www.wqbook.com>

地 址:北京清华大学学研大厦 A 座 邮 编:100084

社 总 机:010-62770175 邮 购:010-62786544

投稿与读者服务:010-62776969, c-service@tup.tsinghua.edu.cn

质量反馈:010-62772015, zhiliang@tup.tsinghua.edu.cn

课件下载: <http://www.tup.com.cn>, 010-62795954

印 装 者:清华大学印刷厂

经 销:全国新华书店

开 本:185mm×260mm 印 张:25.5

字 数:433 千字

版 次:2019 年 5 月第 1 版

印 次:2019 年 5 月第 1 次印刷

印 数:1~2000

定 价:79.80 元

产品编号:075108-01

前言

网络爬虫(Web Crawler)是指一类能够自动化访问网络并抓取某些信息的程序,有时候也被称为“网络机器人”。它们被广泛用于互联网搜索引擎及各种网站的开发中,同时也是大数据和数据分析领域中的重要角色。爬虫可以按一定的逻辑大批量采集目标页面内容,并对数据做进一步处理,人们借此能够更好、更快地获得并使用他们感兴趣的信息,从而方便地完成很多有价值的工作。

Python 是一种解释型、面向对象的、动态数据类型的高级程序设计语言,Python 语法简洁、功能强大,在众多高级语言中拥有十分出色的编写效率,同时还拥有活跃的开源社区和海量程序库,十分适合进行网络内容的抓取和处理。本书将以 Python 语言为基础,由浅入深地探讨网络爬虫技术,同时通过具体的程序编写和实践来帮助读者了解和学习 Python 爬虫。

本书共分为 14 章,其中第 1~3 章为基础篇,第 4~6 章为进阶篇,第 7~9 章为高级篇,第 10~14 章为实践篇,最后为附录。第 1 章、第 2 章介绍了 Python 语言和编写爬虫程序的基础知识;第 3 章讨论了 Python 中对文件和数据的存储,涉及数据库的相关知识;第 4 章、第 5 章的内容针对相对复杂一些的爬虫抓取任务,主要着眼于动态内容和表单登录等方面;第 6 章涉及对抓取到的原始数据的深入处理和分析;第 7~9 章旨在从不同视角讨论爬虫程序,基于爬虫介绍了多个不同主题的内容;第 10~14 章通过一些实际的例子深入讨论爬虫编程的理论知识;最后在附录中介绍了 Python 语言和爬虫编程中常用的知识和工具。

本书的主要特点如下。

- 内容全面,结构清晰。本书详细介绍了网络爬虫技术的方方面面,讨论了数据抓取、数据处理和数据分析的整个流程。全书结构清晰,坚持理论知识与实践操作相结合。
- 循序渐进,生动简洁。本书从最简单的 Python 程序示例开始,在网络爬虫的核心主题之下一步步深入,兼顾内容的广度与深度,在内容编写上使用生动

简洁的阐述方式,力争详略得当。

- 示例丰富,实战性强。网络爬虫是实践性、操作性非常强的技术,本书将提供丰富的代码作为读者的参考,同时对必要的术语和代码进行解释。本书从生活实际出发,选取实用性、趣味性兼具的主题进行网络爬虫实践。
- 内容新颖,不落窠臼。本书中的程序代码均采用最新的 Python 3 版本,并使用了目前主流的各种 Python 框架和库来编写程序,注重内容的先进性。学习网络爬虫需要动手实践才能真正理解,本书最大限度地保证了代码与程序示例的易用性和易读性。

本书在第 10~14 章,针对 5 个爬虫实践,配有微课视频讲解,以方便读者更好地理解 Python 爬虫相关的理论和实践知识。

本书的编者为吕云翔、张扬,曾洪立参与了部分内容的编写及资料整理工作。

由于编者的水平有限,书中的不足在所难免,恳请广大读者批评指正。

编 者

2019 年 1 月

目 录



本书源码下载

基 础 篇

第 1 章 Python 与网络爬虫	3
1.1 Python 语言	4
1.1.1 什么是 Python	4
1.1.2 Python 的应用现状	5
1.2 Python 的安装与开发环境配置	6
1.2.1 在 Windows 上安装	6
1.2.2 在 Ubuntu 和 Mac OS 上安装	8
1.2.3 PyCharm 的使用	8
1.2.4 Jupyter Notebook	14
1.3 Python 的基本语法	16
1.3.1 数据类型	17
1.3.2 逻辑语句	24
1.3.3 Python 中的函数与类	28
1.3.4 如何学习 Python	31
1.4 互联网、HTTP 与 HTML	31
1.4.1 互联网与 HTTP 协议	31
1.4.2 HTML	33
1.5 HelloSpider	36
1.5.1 第一个爬虫程序	36
1.5.2 对爬虫程序的思考	39
1.6 调研网站	41
1.6.1 网站的 robots.txt 与 Sitemap	41

1.6.2	查看网站所用的技术	44
1.6.3	查看网站所有者的信息	46
1.6.4	使用开发者工具检查网页	47
1.7	本章小结	51
第2章	数据的采集	52
2.1	从抓取开始	52
2.2	正则表达式	53
2.2.1	初识正则表达式	53
2.2.2	正则表达式的简单使用	56
2.3	BeautifulSoup	59
2.3.1	BeautifulSoup 的安装与特点	60
2.3.2	BeautifulSoup 的基本使用	63
2.4	XPath 与 lxml	67
2.4.1	XPath	67
2.4.2	lxml 与 XPath 的使用	69
2.5	遍历页面	71
2.5.1	抓取下一个页面	71
2.5.2	完成爬虫程序	72
2.6	使用 API	76
2.6.1	API 简介	76
2.6.2	API 使用示例	78
2.7	本章小结	82
第3章	文件与数据的存储	83
3.1	Python 中的文件	83
3.1.1	基本的文件读写	83
3.1.2	序列化	86
3.2	字符串	86
3.3	Python 与图片	88
3.3.1	PIL 与 Pillow	88
3.3.2	Python 与 OpenCV 简介	90

3.4	CSV 文件	92
3.4.1	CSV 简介	92
3.4.2	CSV 的读写	92
3.5	使用数据库	95
3.5.1	使用 MySQL	95
3.5.2	使用 SQLite3	97
3.5.3	使用 SQLAlchemy	99
3.5.4	使用 Redis	101
3.6	其他类型的文档	102
3.7	本章小结	108

进 阶 篇

第 4 章	JavaScript 与动态内容	111
4.1	JavaScript 与 AJAX 技术	112
4.1.1	JavaScript 语言	112
4.1.2	AJAX	116
4.2	抓取 AJAX 数据	117
4.2.1	分析数据	117
4.2.2	提取数据	123
4.3	抓取动态内容	129
4.3.1	动态渲染页面	129
4.3.2	使用 Selenium	130
4.3.3	PyV8 与 Splash	138
4.4	本章小结	142
第 5 章	表单与模拟登录	143
5.1	表单	143
5.1.1	表单与 POST	143
5.1.2	发送表单数据	145
5.2	Cookie	149
5.2.1	什么是 Cookie	149

5.2.2	在 Python 中使用 Cookie	151
5.3	模拟登录网站	153
5.3.1	分析网站	153
5.3.2	通过 Cookie 模拟登录	155
5.4	验证码	159
5.4.1	图片验证码	159
5.4.2	滑动验证	161
5.5	本章小结	166
第 6 章	数据的进一步处理	167
6.1	Python 与文本分析	167
6.1.1	什么是文本分析	167
6.1.2	jieba 与 SnowNLP	169
6.1.3	NLTK	173
6.1.4	文本的分类与聚类	177
6.2	数据处理与科学计算	179
6.2.1	从 MATLAB 到 Python	179
6.2.2	NumPy	180
6.2.3	Pandas	186
6.2.4	Matplotlib	193
6.2.5	SciPy 与 SymPy	197
6.3	本章小结	197

高级篇

第 7 章	更灵活和更多样的爬虫	201
7.1	更灵活的爬虫——以微信数据的抓取为例	201
7.1.1	用 Selenium 抓取 Web 微信信息	201
7.1.2	基于 Python 的微信 API 工具	206
7.2	更多样的爬虫	210
7.2.1	PyQuery	210
7.2.2	在线爬虫应用平台	214

7.2.3	使用 urllib	215
7.3	对爬虫的部署和管理	226
7.3.1	配置远程主机	226
7.3.2	编写本地爬虫	229
7.3.3	部署爬虫	235
7.3.4	查看运行结果	236
7.3.5	使用爬虫管理框架	236
7.4	本章小结	241
第 8 章	浏览器模拟与网站测试	242
8.1	关于测试	242
8.1.1	什么是测试	242
8.1.2	什么是 TDD	243
8.2	Python 的单元测试	244
8.2.1	使用 unittest	244
8.2.2	其他方法	247
8.3	使用 Python 爬虫测试网站	248
8.4	使用 Selenium 测试	251
8.4.1	Selenium 测试常用的网站交互	251
8.4.2	结合 Selenium 进行单元测试	253
8.5	本章小结	255
第 9 章	更强大的爬虫	256
9.1	爬虫框架	256
9.1.1	Scrapy 是什么	256
9.1.2	Scrapy 的安装与入门	258
9.1.3	编写 Scrapy 爬虫	261
9.1.4	其他爬虫框架	264
9.2	网站反爬虫	265
9.2.1	反爬虫的策略	265
9.2.2	伪装 headers	267
9.2.3	使用代理	271

9.2.4 访问频率	275
9.3 多进程与分布式	276
9.3.1 多进程编程与爬虫抓取	276
9.3.2 分布式爬虫	278
9.4 本章小结	279

实 践 篇

第 10 章 爬虫实践：下载网页中的小说和购物评论	283
10.1 下载网络小说	283
10.1.1 分析网页	283
10.1.2 编写爬虫	285
10.1.3 运行并查看 TXT 文件	290
10.2 下载购物评论	291
10.2.1 查看网络数据	292
10.2.2 编写爬虫	295
10.2.3 数据下载结果与爬虫分析	302
10.3 本章小结	304
第 11 章 爬虫实践：保存感兴趣的图片	305
11.1 豆瓣网站分析与爬虫设计	305
11.1.1 从需求出发	305
11.1.2 处理登录问题	307
11.2 编写爬虫程序	309
11.2.1 爬虫脚本	309
11.2.2 程序分析	313
11.3 运行并查看结果	317
11.4 本章小结	318
第 12 章 爬虫实践：网上影评分析	319
12.1 需求分析与爬虫设计	319
12.1.1 网页分析	319
12.1.2 函数设计	320

12.2 编写爬虫	321
12.2.1 编写程序	321
12.2.2 可能的改进	327
12.3 本章小结	329
第 13 章 爬虫实践：使用爬虫下载网页	330
13.1 设计抓取程序	330
13.2 运行程序	335
13.3 展示网页	336
第 14 章 爬虫实践：使用爬虫框架	342
14.1 Gain 框架	342
14.2 使用 Gain 做简单抓取	343
14.3 PySpider 框架	348
14.4 使用 PySpider 进行抓取	351
附录 A	359
A.1 Python 中的一些重要概念	359
A.1.1 * args 与 ** kwargs 的使用	359
A.1.2 global 关键词	361
A.1.3 enumerate 枚举	362
A.1.4 迭代器与生成器	362
A.2 Python 中的常用模块	364
A.2.1 collections	364
A.2.2 arrow	369
A.2.3 timeit	370
A.2.4 pickle	371
A.2.5 os	372
A.2.6 sys	372
A.2.7 itertools	373
A.2.8 functools	374
A.2.9 threading、queue 与 multiprocessing	376
A.3 requests 库	383

A. 3.1	requests 基础	383
A. 3.2	更多用法	386
A. 4	正则表达式	387
A. 4.1	什么是正则表达式	387
A. 4.2	正则表达式的基础语法	388
参考文献	392



第 1 章



Python与网络爬虫

网络爬虫(Web Crawler)有时候也叫网络蜘蛛(Web Spider),是指这样一类程序——它们可以自动连接到互联网站点,读取网页中的内容或者存放在网络上的各种信息,并按照某种策略对目标信息进行采集(例如对某个网站的全部页面进行读取)。实际上,世界上最大的搜索网站——Google 搜索本身就建构在爬虫技术之上,像 Google、百度这样的搜索引擎会通过爬虫程序来不断更新自身的网站内容和对其他网站的网络索引。从某种意义上说,用户每次通过搜索引擎查询一个关键词,就是在搜索引擎服务者的爬虫程序所“爬”到的信息中进行查询。当然,搜索引擎背后所使用的技术十分复杂,其爬虫技术通常也不是一般个人开发的小型程序所能比拟的。其实,爬虫程序本身并不复杂,用户只要懂一点编程知识,了解一点 HTTP 和 HTML,就可以写出属于自己的爬虫,实现很多有意思的功能。

在众多编程语言中,本书选择 Python 来编写爬虫程序,因为 Python 不仅语法简洁、便于上手,而且拥有庞大的开发者社区和浩如烟海的模块库,对于普通的程序编写而言有极大的便利。虽然 Python 和 C/C++ 等语言相比可能在性能上有所欠缺,但毕竟瑕不掩瑜,是目前最好的选择。

1.1 Python 语言

Python 是目前最流行的编程语言之一,本书对它的历史和发展作一些简单介绍,然后看看 Python 的基本语法,对于没有 Python 编程经验的读者而言,可以借此对 Python 有一个初步的了解。

1.1.1 什么是 Python

Guido van Rossum 在 1989 年发明了 Python,而 Python 的第一个公开发行人版发行于 1991 年。因为 Guido 是电视剧 *Monty Python's Flying Circus* 的爱好者,所以将这种新的脚本语言命名为 Python。

从最根本的角度来说,Python 是一种解释型、面向对象的、动态数据类型的高级程序设计语言。值得注意的是,Python 是开源的,源代码遵循 GPL(GNU General Public License)协议,这就意味着它对所有个人开发者是完全开放的,这也使得 Python 在开发者中迅速流行开来,来自全球各地的 Python 使用者为这门语言的发展贡献了很多力量。Python 的哲学是优雅、明确和简单。著名的 *the Zen of Python* (Python 之禅)^①这样说道:

“

优美胜于丑陋,

明了胜于晦涩,

简洁胜于复杂,

复杂胜于凌乱,

扁平胜于嵌套,

间隔胜于紧凑,

可读性很重要。

即便假借特例的实用性之名,也不可违背这些规则,

不要包容所有错误,除非你确定需要这样做,

^① 作者为 Tim Peters,英文原文可见“<https://www.python.org/dev/peps/pep-0020/>”。