

# Web Scraping com Python

**2ª Edição**

**Ryan Mitchell**

**O'REILLY®**  
Novatec

Authorized Portuguese translation of the English edition of titled Web Scraping with Python, 2E, ISBN 9781491985571 © 2018 Ryan Mitchell. This translation is published and sold by permission of O'Reilly Media, Inc., the owner of all rights to publish and sell the same.

Tradução em português autorizada da edição em inglês da obra Web Scraping with Python, 2E, ISBN 9781491985571 © 2018 Ryan Mitchell. Esta tradução é publicada e vendida com a permissão da O'Reilly Media, Inc., detentora de todos os direitos para publicação e venda desta obra.

© Novatec Editora Ltda. [2019].

Todos os direitos reservados e protegidos pela Lei 9.610 de 19/02/1998. É proibida a reprodução desta obra, mesmo parcial, por qualquer processo, sem prévia autorização, por escrito, do autor e da Editora.

Editor: Rubens Prates

Tradução: Lúcia A. Kinoshita

Revisão gramatical: Tássia Carvalho

Editoração eletrônica: Carolina Kuwabata

ISBN: 978-85-7522-730-5

Histórico de impressões:

Março/2019	Segunda edição
Agosto/2016	Primeira reimpressão
Agosto/2015	Primeira edição (ISBN: 978-85-7522-447-2)

Novatec Editora Ltda.

Rua Luís Antônio dos Santos 110  
02460-000 – São Paulo, SP – Brasil

Tel.: +55 11 2959-6529

Email: [novatec@novatec.com.br](mailto:novatec@novatec.com.br)

Site: [www.novatec.com.br](http://www.novatec.com.br)

Twitter: [twitter.com/novateceditora](https://twitter.com/novateceditora)

Facebook: [facebook.com/novatec](https://facebook.com/novatec)

LinkedIn: [linkedin.com/in/novatec](https://linkedin.com/in/novatec)

# Sumário

Prefácio .....	11
<b>Parte I = Construindo scrapers.....</b>	<b>19</b>
<b>Capítulo 1 = Seu primeiro web scraper .....</b>	<b>20</b>
Conectando.....	20
Introdução ao BeautifulSoup .....	23
Instalando o BeautifulSoup .....	23
Executando o BeautifulSoup .....	26
Conectando-se de forma confiável e tratando exceções .....	28
<b>Capítulo 2 = Parsing de HTML avançado .....</b>	<b>32</b>
Nem sempre um martelo é necessário .....	32
Outras utilidades do BeautifulSoup.....	34
find() e find_all() com o BeautifulSoup.....	35
Outros objetos do BeautifulSoup .....	38
Navegando em árvores .....	38
Expressões regulares .....	43
Expressões regulares e o BeautifulSoup.....	47
Acessando atributos.....	49
Expressões lambda.....	49
<b>Capítulo 3 = Escrevendo web crawlers.....</b>	<b>51</b>
Percorrendo um único domínio.....	51
Rastreamento de um site completo .....	56
Coletando dados de um site completo.....	59
Rastreamento pela internet.....	61
<b>Capítulo 4 = Modelos de web crawling .....</b>	<b>67</b>
Planejando e definindo objetos.....	68
Lidando com diferentes layouts de sites .....	72

Estruturando os crawlers .....	77
Rastreando sites por meio de pesquisa .....	78
Rastreando sites por meio de links .....	82
Rastreando vários tipos de página .....	84
Pensando nos modelos de web crawlers .....	86
<b>Capítulo 5 = Scrapy.....</b>	<b>88</b>
Instalando o Scrapy .....	88
Escrevendo um scraper simples .....	90
Spidering com regras .....	92
Criando itens.....	96
Apresentando itens .....	98
Pipeline de itens.....	99
Fazendo log com o Scrapy .....	103
Outros recursos.....	104
<b>Capítulo 6 = Armazenando dados.....</b>	<b>105</b>
Arquivos de mídia .....	105
Armazenando dados no formato CSV .....	109
MySQL .....	111
Instalando o MySQL.....	112
Alguns comandos básicos.....	114
Integração com Python .....	117
Técnicas de banco de dados e boas práticas .....	121
“Six Degrees” no MySQL .....	124
Email .....	128
<b>Parte II = Coleta de dados avançada.....</b>	<b>130</b>
<b>Capítulo 7 = Lendo documentos.....</b>	<b>131</b>
Codificação de documentos .....	131
Texto.....	132
Codificação de texto e a internet global .....	133
CSV .....	138
Lendo arquivos CSV .....	138
PDF.....	140
Microsoft Word e .docx .....	142
<b>Capítulo 8 = Limpando dados sujos .....</b>	<b>147</b>
Código para limpeza de dados .....	147
Normalização de dados .....	151

Limpeza dos dados após a coleta.....	153
OpenRefine .....	154
<b>Capítulo 9 = Lendo e escrevendo em idiomas naturais .....</b>	<b>159</b>
Resumindo dados.....	160
Modelos de Markov.....	164
Six Degrees of Wikipedia: conclusão .....	169
Natural Language Toolkit.....	172
Instalação e configuração .....	172
Análise estatística com o NLTK.....	173
Análise lexicográfica com o NLTK .....	176
Recursos adicionais .....	179
<b>Capítulo 10 = Rastreamento de formulários e logins.....</b>	<b>181</b>
Biblioteca Python Requests.....	181
Submetendo um formulário básico.....	182
Botões de rádio, caixas de seleção e outras entradas .....	184
Submetendo arquivos e imagens.....	186
Lidando com logins e cookies .....	187
Autenticação de acesso básica do HTTP .....	188
Outros problemas de formulário .....	189
<b>Capítulo 11 = Scraping de JavaScript .....</b>	<b>191</b>
Introdução rápida ao JavaScript .....	192
Bibliotecas JavaScript comuns .....	193
Ajax e HTML dinâmico.....	195
Executando JavaScript em Python com o Selenium .....	197
Webdrivers adicionais do Selenium .....	203
Lidando com redirecionamentos.....	203
Última observação sobre o JavaScript.....	205
<b>Capítulo 12 = Rastreamento por meio de APIs.....</b>	<b>207</b>
Introdução rápida às APIs.....	207
Métodos HTTP e APIs .....	209
Mais sobre respostas de APIs° .....	210
Parsing de JSON.....	212
APIs não documentadas .....	213
Encontrando APIs não documentadas .....	215
Documentando APIs não documentadas .....	216
Encontrando e documentando APIs de modo automático .....	217
Combinando APIs com outras fontes de dados.....	220
Mais sobre APIs.....	225

<b>Capítulo 13 = Processamento de imagens e reconhecimento de texto .....</b>	<b>226</b>
Visão geral das bibliotecas .....	227
Pillow .....	227
Tesseract.....	228
NumPy .....	231
Processando textos bem formatados.....	231
Ajustes automáticos nas imagens .....	234
Coletando texto de imagens em sites .....	238
Lendo CAPTCHAs e treinando o Tesseract.....	241
Treinando o Tesseract .....	243
Lendo CAPTCHAs e enviando soluções .....	247
<b>Capítulo 14 = Evitando armadilhas no scraping.....</b>	<b>251</b>
Uma observação sobre ética .....	251
Parecendo um ser humano .....	253
Ajuste seus cabeçalhos .....	253
Lidando com cookies em JavaScript.....	255
Tempo é tudo.....	257
Recursos de segurança comuns em formulários .....	258
Valores de campos de entrada ocultos.....	259
Evitando honeypots .....	260
Lista de verificação para parecer um ser humano .....	262
<b>Capítulo 15 = Testando seu site com scrapers.....</b>	<b>264</b>
Introdução aos testes .....	265
O que são testes de unidade? .....	265
Módulo unittest de Python .....	266
Testando a Wikipédia .....	268
Testando com o Selenium .....	271
Interagindo com o site.....	271
unittest ou Selenium?.....	275
<b>Capítulo 16 = Web Crawling em paralelo .....</b>	<b>277</b>
Processos versus threads.....	277
Crawling com várias threads.....	278
Condições de concorrência e filas .....	281
Módulo threading .....	284
Rastreamento com multiprocessamento .....	287
Rastreamento da Wikipédia com multiprocessamento .....	289
Comunicação entre processos.....	291
Rastreamento com multiprocessamento – outra abordagem .....	294

<b>Capítulo 17 = Fazendo scraping remotamente .....</b>	<b>296</b>
Por que usar servidores remotos?.....	296
Evitando o bloqueio de endereços IP .....	297
Portabilidade e extensibilidade .....	298
Tor .....	299
PySocks .....	300
Hospedagem remota.....	301
Executando de uma conta que hospeda sites.....	301
Executando a partir da nuvem.....	302
Recursos adicionais .....	304
<b>Capítulo 18 = Aspectos legais e éticos do web scraping .....</b>	<b>306</b>
Marcas registradas, direitos autorais, patentes, oh, céus! .....	306
Lei de direitos autorais .....	308
Invasão de bens móveis.....	309
Lei de Fraude e Abuso de Computadores.....	312
robots.txt e Termos de Serviço .....	313
Três web scrapers .....	317
eBay versus Bidder’s Edge e transgressão a bens móveis.....	317
Estados Unidos versus Auernheimer e a Lei de Fraude e Abuso de Computadores .....	319
Field versus Google: direitos autorais e robots.txt .....	321
Seguindo em frente .....	322