

Expressões Regulares Cookbook

**Jan Goyvaerts
Steven Levithan**

Authorized Portuguese translation of the English edition of *Regular Expressions Cookbook* ISBN 9780596520687
© 2009, Jan Goyvaerts and Steve Levithan. 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 *Regular Expressions Cookbook* ISBN
9780596520687 © 2009, Jan Goyvaerts e Steve Levithan. 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. 2011.

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: Rafael Contatori e Edgard Damiani
Revisão gramatical: Jeferson Ferreira
Editoração eletrônica: Camila Kuwabata e Carolina Kuwabata

ISBN: 978-85-7522-279-9

Histórico de impressões:

Abril/2011 Primeira edição

Novatec Editora Ltda.
Rua Luís Antônio dos Santos 110
02460-000 – São Paulo, SP – Brasil
Tel.: +55 11 2959-6529
Fax: +55 11 2950-8869
E-mail: novatec@novatec.com.br
Site: www.novatec.com.br
Twitter: twitter.com/novateceditora
Facebook: facebook.com/novatec
LinkedIn: linkedin.com/in/novatec

Dados Internacionais de Catalogação na Publicação (CIP)
(Câmara Brasileira do Livro, SP, Brasil)

Goyvaerts, Jan
Expressões regulares Cookbook / Jan Goyvaerts,
Steven Levithan ; [tradução Rafael Contatori
/ Edgard Damiani]. -- São Paulo : Novatec Editora ;
Cambridge : O'Reilly Media, 2011.

Título original: Regular expressions Cookbook.
ISBN 978-85-7522-279-9

1. Cookbook (Linguagem de programação)
2. Expressões regulares I. Levithan, Steven.
II. Título.

10-08695

CDD-005.115

Índices para catálogo sistemático:

1. Expressões regulares Cookbook : Ciência da
computação 005.115
OGF20110406

Sumário

Prefácio	9
Capítulo 1 ■ Introdução às expressões regulares	15
Definição de expressões regulares	15
Pesquisa e substituição com expressões regulares	20
Ferramentas para se trabalhar com expressões regulares	22
grep	35
Capítulo 2 ■ Habilidades básicas de expressões regulares	40
2.1 Corresponder a um texto literal	41
2.2 Corresponder a caracteres não-imprimíveis	43
2.3 Corresponder a um dentre vários caracteres	46
2.4 Corresponder a qualquer caractere	50
2.5 Corresponder a alguma coisa no começo e/ou final de uma linha	53
2.6 Corresponder a palavras inteiras	58
2.7 Pontos de código, propriedades, blocos e alfabetos Unicode	61
2.8 Corresponder a uma dentre várias alternativas	74
2.9 Agrupar e capturar partes da correspondência	76
2.10 Corresponder novamente a textos previamente correspondidos	79
2.11 Capturar e nomear partes da correspondência	81
2.12 Repetir parte da Regex um certo número de vezes	84
2.13 Escolher entre repetição mínima ou máxima	87
2.14 Eliminar os retrocessos desnecessários	91
2.15 Prevenir repetições descontroladas	94
2.16 Testar uma correspondência sem acrescentá-la à correspondência global	96
2.17 Corresponder a uma de duas alternativas com base em uma condição	103
2.18 Adicionar comentários à expressão regular	105
2.19 Inserir texto literal no texto de substituição	107
2.20 Inserir a correspondência da expressão regular no texto de substituição	111
2.21 Inserir parte da correspondência da expressão regular no texto de substituição	112
2.22 Inserir o contexto de correspondência no texto de substituição	116
Capítulo 3 ■ Programando com expressões regulares	118
Linguagens de programação e sabores Regex	118
3.1 Expressões regulares literais no código-fonte	124

3.2 Importar a biblioteca de expressões regulares.....	131
3.3 Criar objetos de expressão regular	133
3.4 Definir opções das expressões regulares	140
3.5 Testar se uma correspondência pode ser encontrada dentro de uma string de assunto	148
3.6 Testar se uma regex corresponde totalmente à string de assunto	155
3.7 Recuperar o texto correspondido.....	160
3.8 Determinar a posição e o comprimento da correspondência.....	166
3.9 Recuperar parte do texto correspondido	172
3.10 Recuperar uma lista de todas as correspondências	180
3.11 Iterar todas as correspondências	186
3.12 Validar correspondências no código procedural	192
3.13 Encontrar uma correspondência dentro de outra	196
3.14 Substituir todas as correspondências	200
3.15 Substituir correspondências, reutilizando partes da correspondência.....	209
3.16 Substituir correspondências com substitutos gerados em código	214
3.17 Substituir todas as correspondências dentro das correspondências de outra expressão regular	220
3.18 Substituir todas as correspondências entre as correspondências de outra expressão regular ..	223
3.19 Dividir uma string.....	228
3.20 Dividir uma string, mantendo as correspondências da expressão regular	238
3.21 Pesquisar linha por linha	243
Capítulo 4 ■ Validação e formatação	247
4.1 Validar endereços de e-mail	247
4.2 Validar e formatar números de telefone norte-americanos	254
4.3 Validar números de telefone internacionais	259
4.4 Validar formatos de data tradicionais.....	262
4.5 Validar formatos tradicionais de data com exatidão.....	266
4.6 Validar formatos de horário tradicionais.....	271
4.7 Validando datas e horários ISO 8601	274
4.8 Limitar a entrada a caracteres alfanuméricos	279
4.9 Limitar o comprimento do texto	282
4.10 Limitar o número de linhas no texto	287
4.11 Validar respostas afirmativas	292
4.12 Validar números de Previdência Social.....	294
4.13 Validando ISBNs	297
4.14 Validar códigos postais	305
4.15 Validar códigos postais canadenses	306
4.16 Validar códigos postais do Reino Unido	307
4.17 Encontrar endereços com caixas postais.....	307
4.18 Reformatar nomes no formato “Nome Sobrenome” para “Sobrenome, Nome”	309
4.19 Validar números de cartão de crédito	313
4.20 Números VAT europeus	320
Capítulo 5 ■ Palavras, linhas e caracteres especiais	326
5.1 Encontrar uma palavra específica.....	326
5.2 Encontrar uma palavra entre várias	329

5.3 Pesquisar palavras similares	331
5.4 Encontrar todas, exceto uma palavra específica	335
5.5 Localizar qualquer palavra não seguida por uma palavra específica	337
5.6 Localizar qualquer palavra que não seja precedida por uma palavra específica	339
5.7 Encontrar palavras próximas umas das outras	343
5.8 Encontrar palavras repetidas	350
5.9 Remover linhas duplicadas.....	352
5.10 Corresponder a linhas inteiras que contenham uma determinada palavra.....	357
5.11 Corresponder a linhas completas que não contenham determinada palavra	359
5.12 Remover espaços em branco iniciais e finais.....	360
5.13 Substituir espaços em branco repetidos por um único espaço.....	363
5.14 Escapar metacaracteres de expressão regular	365
Capítulo 6 ■ Números	369
6.1 Números inteiros	369
6.2 Números hexadecimais.....	373
6.3 Números binários.....	376
6.4 Remover zeros à esquerda	377
6.5 Números dentro de um certo intervalo.....	378
6.6 Números hexadecimais dentro de um certo intervalo	385
6.7 Números de ponto flutuante.....	388
6.8 Números com separadores de milhar.....	391
6.9 Numerais romanos.....	393
Capítulo 7 ■ URLs, paths e endereços de Internet.....	396
7.1 Validar URLs.....	396
7.2 Encontrar URLs dentro de um texto completo	400
7.3 Encontrar URLs entre aspas no texto completo	402
7.4 Encontrar URLs entre parênteses no texto completo	403
7.5 Transformar URLs em links.....	405
7.6 Validar URNs	406
7.7 Validar URLs genéricas	409
7.8 Extrair o protocolo de uma URL	415
7.9 Extrair o usuário de uma URL	417
7.10 Extrair o host de uma URL.....	419
7.11 Extrair a porta de uma URL.....	421
7.12 Extrair o caminho de uma URL	423
7.13 Extrair a consulta de uma URL.....	427
7.14 Extrair o fragmento de uma URL	428
7.15 Validar nomes de domínio	429
7.16 Corresponder a endereços IPv4	432
7.17 Corresponder a endereços IPv6	434
7.18 Validar caminhos do Windows.....	450
7.19 Dividir caminhos do Windows em suas partes constituintes.....	453
7.20 Extrair a letra da unidade de um caminho Windows	458
7.21 Extrair o servidor e o compartilhamento de um caminho UNC.....	459

7.22 Extrair a pasta de um caminho Windows.....	461
7.23 Extrair o nome do arquivo de um caminho do Windows	463
7.24 Extrair a extensão de arquivo de um caminho Windows	464
7.25 Retirar caracteres inválidos de nomes de arquivos.....	465
Capítulo 8 ■ Marcação e intercâmbio de dados	467
8.1 Encontrar tags no estilo XML.....	474
8.2 Substituir tags por	494
8.3 Remover todas as tags de estilo XML, exceto e 	497
8.4 Corresponder a nomes XML	501
8.5 Converter texto simples em HTML adicionando tags <p> e 	508
8.6 Encontrar um atributo específico em tags no estilo XML.....	512
8.7 Adicionar um atributo cellpadding em tags <table> que ainda não o incluíam.....	517
8.8 Remover comentários no estilo XML.....	520
8.9 Encontrar palavras dentro de comentários no estilo XML.....	525
8.10 Mudar o delimitador usado em arquivos CSV	530
8.11 Extrair campos CSV de uma coluna específica	534
8.12 Corresponder a cabeçalhos de seção INI	538
8.13 Corresponder a blocos de seção INI.....	539
8.14 Corresponder a pares nome-valor INI	541
Índice remissivo	545