
Regular Expressions Cookbook

Jan Goyvaerts and Steven Levithan

O'REILLY®

Beijing • Cambridge • Farnham • Köln • Sebastopol • Taipei • Tokyo

Table of Contents

Preface	ix
1. Introduction to Regular Expressions	1
Regular Expressions Defined	1
Searching and Replacing with Regular Expressions	5
Tools for Working with Regular Expressions	7
2. Basic Regular Expression Skills	25
2.1 Match Literal Text	26
2.2 Match Nonprintable Characters	28
2.3 Match One of Many Characters	30
2.4 Match Any Character	34
2.5 Match Something at the Start and/or the End of a Line	36
2.6 Match Whole Words	41
2.7 Unicode Code Points, Properties, Blocks, and Scripts	43
2.8 Match One of Several Alternatives	55
2.9 Group and Capture Parts of the Match	57
2.10 Match Previously Matched Text Again	60
2.11 Capture and Name Parts of the Match	62
2.12 Repeat Part of the Regex a Certain Number of Times	64
2.13 Choose Minimal or Maximal Repetition	67
2.14 Eliminate Needless Backtracking	70
2.15 Prevent Runaway Repetition	72
2.16 Test for a Match Without Adding It to the Overall Match	75
2.17 Match One of Two Alternatives Based on a Condition	81
2.18 Add Comments to a Regular Expression	83
2.19 Insert Literal Text into the Replacement Text	85
2.20 Insert the Regex Match into the Replacement Text	87
2.21 Insert Part of the Regex Match into the Replacement Text	88
2.22 Insert Match Context into the Replacement Text	92

3. Programming with Regular Expressions	95
Programming Languages and Regex Flavors	95
3.1 Literal Regular Expressions in Source Code	100
3.2 Import the Regular Expression Library	106
3.3 Creating Regular Expression Objects	108
3.4 Setting Regular Expression Options	114
3.5 Test Whether a Match Can Be Found Within a Subject String	121
3.6 Test Whether a Regex Matches the Subject String Entirely	127
3.7 Retrieve the Matched Text	132
3.8 Determine the Position and Length of the Match	138
3.9 Retrieve Part of the Matched Text	143
3.10 Retrieve a List of All Matches	150
3.11 Iterate over All Matches	155
3.12 Validate Matches in Procedural Code	161
3.13 Find a Match Within Another Match	165
3.14 Replace All Matches	169
3.15 Replace Matches Reusing Parts of the Match	176
3.16 Replace Matches with Replacements Generated in Code	181
3.17 Replace All Matches Within the Matches of Another Regex	187
3.18 Replace All Matches Between the Matches of Another Regex	189
3.19 Split a String	195
3.20 Split a String, Keeping the Regex Matches	203
3.21 Search Line by Line	208
4. Validation and Formatting	213
4.1 Validate Email Addresses	213
4.2 Validate and Format North American Phone Numbers	219
4.3 Validate International Phone Numbers	224
4.4 Validate Traditional Date Formats	226
4.5 Accurately Validate Traditional Date Formats	229
4.6 Validate Traditional Time Formats	234
4.7 Validate ISO 8601 Dates and Times	237
4.8 Limit Input to Alphanumeric Characters	241
4.9 Limit the Length of Text	244
4.10 Limit the Number of Lines in Text	248
4.11 Validate Affirmative Responses	253
4.12 Validate Social Security Numbers	254
4.13 Validate ISBNs	257
4.14 Validate ZIP Codes	264
4.15 Validate Canadian Postal Codes	265
4.16 Validate U.K. Postcodes	266
4.17 Find Addresses with Post Office Boxes	266

4.18 Reformat Names From "FirstName LastName" to "LastName, FirstName"	268
4.19 Validate Credit Card Numbers	271
4.20 European VAT Numbers	278
5. Words, Lines, and Special Characters	285
5.1 Find a Specific Word	285
5.2 Find Any of Multiple Words	288
5.3 Find Similar Words	290
5.4 Find All Except a Specific Word	294
5.5 Find Any Word Not Followed by a Specific Word	295
5.6 Find Any Word Not Preceded by a Specific Word	297
5.7 Find Words Near Each Other	300
5.8 Find Repeated Words	306
5.9 Remove Duplicate Lines	308
5.10 Match Complete Lines That Contain a Word	312
5.11 Match Complete Lines That Do Not Contain a Word	313
5.12 Trim Leading and Trailing Whitespace	314
5.13 Replace Repeated Whitespace with a Single Space	317
5.14 Escape Regular Expression Metacharacters	319
6. Numbers	323
6.1 Integer Numbers	323
6.2 Hexadecimal Numbers	326
6.3 Binary Numbers	329
6.4 Strip Leading Zeros	330
6.5 Numbers Within a Certain Range	331
6.6 Hexadecimal Numbers Within a Certain Range	337
6.7 Floating Point Numbers	340
6.8 Numbers with Thousand Separators	343
6.9 Roman Numerals	344
7. URLs, Paths, and Internet Addresses	347
7.1 Validating URLs	347
7.2 Finding URLs Within Full Text	350
7.3 Finding Quoted URLs in Full Text	352
7.4 Finding URLs with Parentheses in Full Text	353
7.5 Turn URLs into Links	356
7.6 Validating URNs	356
7.7 Validating Generic URLs	358
7.8 Extracting the Scheme from a URL	364
7.9 Extracting the User from a URL	366
7.10 Extracting the Host from a URL	367

7.11 Extracting the Port from a URL	369
7.12 Extracting the Path from a URL	371
7.13 Extracting the Query from a URL	374
7.14 Extracting the Fragment from a URL	376
7.15 Validating Domain Names	376
7.16 Matching IPv4 Addresses	379
7.17 Matching IPv6 Addresses	381
7.18 Validate Windows Paths	395
7.19 Split Windows Paths into Their Parts	397
7.20 Extract the Drive Letter from a Windows Path	402
7.21 Extract the Server and Share from a UNC Path	403
7.22 Extract the Folder from a Windows Path	404
7.23 Extract the Filename from a Windows Path	406
7.24 Extract the File Extension from a Windows Path	407
7.25 Strip Invalid Characters from Filenames	408
8. Markup and Data Interchange	411
8.1 Find XML-Style Tags	417
8.2 Replace Tags with 	434
8.3 Remove All XML-Style Tags Except and 	438
8.4 Match XML Names	441
8.5 Convert Plain Text to HTML by Adding <p> and Tags	447
8.6 Find a Specific Attribute in XML-Style Tags	450
8.7 Add a cellspacing Attribute to <table> Tags That Do Not Already Include It	455
8.8 Remove XML-Style Comments	458
8.9 Find Words Within XML-Style Comments	462
8.10 Change the Delimiter Used in CSV Files	466
8.11 Extract CSV Fields from a Specific Column	469
8.12 Match INI Section Headers	473
8.13 Match INI Section Blocks	475
8.14 Match INI Name-Value Pairs	476
Index	479