Beginning Regular Expressions

Andrew Watt

wrox

Programmer to Programmer'

Contents

Introduction xxi

Who This Book Is For xxi
What This Book Covers xxii
How This Book Is Structured xxii
What You Need to Use This Book xxiii
Conventions xxiii
Source Code xxiv
Errata xxiv
p2p.wrox.com xxv

Chapter 1: **Introduction to Regular Expressions 1**

What Are Regular Expressions? 2
What Can Regular Expressions Be Used For? 5
Finding Doubled Words 5
Checking Input from Web Forms 5
Changing Date Formats 6
Finding Incorrect Case 6
Adding Links to URLs 6
Regular Expressions You Already Use 7
Search and Replace in Word Processors 7
Directory Listings 7
Online Searching 8
Why Regular Expressions Seem Intimidating 8
Compact, Cryptic Syntax 8
Whitespace Can Significantly Alter the Meaning 9
No Standards Body 12
Differences between Implementations 12
Characters Change Meaning in Different Contexts 13
Regular Expressions Can Be Case Sensitive 15
Case-Sensitive and Case-Insensitive Matching 15
Case and Metacharacters 16

Continual Evolution in Techniques Supported 16
Multiple Solutions for a Single Problem 16
What You Want "to Do with a Regular Expression 17
The Languages That Support Regular Expressions 17
Replacing Text in Quantity 17

Chapter 2: Regular Expression Tools and an Approach to Using Them 21

Regular Expression Tools 21
findstr 22
Microsoft Word 23
StarOffice Writer/OpenOffice.org Writer 27
Komodo Rx Package 28
PowerGrep 28
Microsoft Excel 28
Language- and Platform-Specific Tools 29
JavaScript and JScript 29
VBScript 29
Visual Basic.NET 29
C# 30
PHP 30
Java 30
Perl 30
MySQL 30
SQL Server 2000 31
W3C XML Schema 31
An Analytical Approach to Using Regular Expressions 31
Express and Document What You Want to Do in English 32
Consider the Data Source and Its Likely Contents 34
Consider the Regular Expression Options Available 34
Consider Sensitivity and Specificity 34
Create Appropriate Regular Expressions 35
Document All but Simple Regular Expressions 35
Document What You Expect the Regular Expression to Do 36
Document What You Want to Match 37
Document What You Don't Want to Select 37
Use Whitespace to Aid in Clear Documentation of the Regular Expression 37
Test the Results of a Regular Expression 38

Chapter 3: Simple Regular Expressions 41

Matching Single Characters 42
Matching Sequences of Characters That Each Occur Once 47
Introducing Metacharacters 49
Matching Sequences of Different Characters 54
Matching Optional Characters 56
Matching Multiple Optional Characters 59
Other Cardinality Operators 62
The * Quantifier 62
The + Quantifier 64
The Curly-Brace Syntax 66
The {n} Syntax 66
The {n,m} Syntax 67
{0,m} 67
{n,m} 69
{n,} 70
Exercises 71

Chapter 4: Metacharacters **and Modifiers** Z3

Regular Expression Metacharacters 74
Thinking about Characters and Positions 74
The Period (.) Metacharacter 75
Matching Variably Structured Part Numbers 78
Matching a Literal Period 80
The \w Metacharacter 81
The \W Metacharacter 82
Digits and Nondigits 83
The \d Metacharacter 84
Canadian Postal Code Example 85
The \D Metacharacter 89
Alternatives to \d and \D 90
Whitespace and Non-Whitespace Metacharacters 92
The \s Metacharacter 93
Handling Optional Whitespace 96
The \S Metacharacter 98
The \t Metacharacter 98
The \n Metacharacter 99
Escaped Characters 102
Finding the Backslash 102

Modifiers **103**
Global Search 103
Case-Insensitive Search 104
Exercises **104**

Chapter 5: Character Classes **105**

Introduction to Character Classes **105**
Choice between Two Characters 108
Using Quantifiers with Character Classes 111
Using the \b Metacharacter in Character Classes 112
Selecting Literal Square Brackets 113
Using Ranges in Character Classes **114**
Alphabetic Ranges 115
Use [A-z] With Care 116
Digit Ranges in Character Classes 117
Hexadecimal Numbers 119
IP Addresses 120
Reverse Ranges in Character Classes 128
A Potential Range Trap 129
Finding HTML Heading Elements 132
Metacharacter Meaning within Character Classes **133**
The ^{A} metacharacter 133
How to Use the - Metacharacter 135
Negated Character Classes **136**
Combining Positive and Negative Character Classes 137
POSIX Character Classes **139**
The [:alnum:] Character Class 139
Exercises **141**

Chapter 6: String, Line, and Word Boundaries **143**

String, Line, and Word Boundaries **144**
The ^{A} Metacharacter 144
The ^{A} Metacharacter and Multiline Mode 146
The $ Metacharacter 149
The $ Metacharacter in Multiline Mode 150
Using the ^{A} and $ Metacharacters Together 153
Matching Blank Lines 155
Working with Dollar Amounts 158
Revisiting the IP Address Example 161
What Is a Word? **164**

Identifying Word Boundaries **164**
The \< Syntax 164
The \>Syntax 166
The \b Syntax 168
The \B Metacharacter 168
Less-Common Word-Boundary Metacharacters 169
Exercises **169**

Chapter 7: Parentheses in Regular Expressions : HI

Grouping Using Parentheses **171**
Parentheses and Quantifiers 173
Matching Literal Parentheses 175
U.S. Telephone Number Example 175
Alternation **177**
Choosing among Multiple Options 180
Unexpected Alternation Behavior 182
Capturing Parentheses **185**
Numbering of Captured Groups 185
Numbering When Using Nested Parentheses 186
Named Groups 187
Non-Capturing Parentheses **188**
Back References **190**
Exercises **193**

Chapter 8: Lookahead and Lookbehind 195.

Why You Need Lookahead and Lookbehind **196**
The (? metacharacters 196
Lookahead **197**
Positive Lookahead 199
Positive Lookahead — Star Training Example 199
Positive Lookahead — Later in Same Sentence 200
Negative Lookahead 202
Positive Lookahead Examples **203**
Positive Lookahead in the Same Document 203
Inserting an Apostrophe 205
Lookbehind **209**
Positive Lookbehind 209
Negative Lookbehind 213
How to Match Positions **214**
Adding Commas to Large Numbers 216
Exercises **220**

Chapter 9: Sensitivity and Specificity of Regular Expressions 221

What Are Sensitivity and Specificity? **222**
Extreme Sensitivity, Awful Specificity 222
Email Addresses Example 224
Replacing Hyphens Example 228
The Sensitivity/Specificity Trade-Off 230
How Metacharacters Affect Sensitivity and Specificity 230
Sensitivity, Specificity, and Positional Characters 231
Sensitivity, Specificity, and Modes 232
Sensitivity, Specificity, and Lookahead and Lookbehind 232
How Much Should the Regular Expressions Do? 232
Knowing the Data, Sensitivity, and Specificity 233
Abbreviations 234
Characters from Other Languages 234
Names 235
Sensitivity and How to Achieve It 236
Specificity and How to Maximize It 236
Revisiting the Star Training Company Example 236
Exercises 240

Chapter 10; Documenting and Debugging Regular Expressions 241

Documenting Regular Expressions **242**
Document the Problem Definition 242
Add Comments to Your Code 243
Making Use of Extended Mode 243
Know Your Data 246
Abbreviations 246
Proper Names 246
Incorrect Spelling 247
Creating Test Cases **247**
Debugging Regular Expressions **248**
Treacherous Whitespace 248
Backslashes Causing Problems 251
Considering Other Causes 251

Chapter 11: **Regular** Expressions **In Microsoft Word 253.**

The User Interface **253**
Metacharacters Available **256**
Quantifiers 256
The @ Quantifier 258
The {n,m} Syntax 260

Modes 262
Character Classes 265
Back References 265
Lookahead and Lookbehind 265
Lazy Matching versus Greedy Matching 265
Examples 268
Character Class Examples, Including Ranges 268
Whole Word Searches 269
Search-and-Replace Examples 270
Changing Name Structure Using Back References 270
Manipulating Dates 273
The Star Training Company Example 275
Regular Expressions in Visual Basic for Applications 278
Exercises 280

Chapter 12: Regular Expressions In StarOfflce/OpenOfflce.org Writer 281

The User Interface 282
Metacharacters Available 284
Quantifiers 285
Modes 286
Character Classes 286
Alternation 289
Back References 292
Lookahead and Lookbehind 294
Search Example 294
Search-and-Replace Example 297
Online Chats 297
POSIX Character Classes 301
Matching Numeric Digits 302
Exercises 304

Chapter 13: Regular Expressions Using findstr 305

Introducing findstr 305
Finding Literal Text 306
Metacharacters Supported by findstr 308
Quantifiers 310
Character Classes 311
Word-Boundary Positions 313
Beginning- and End-of-Line Positions 315
Command-Line Switch Examples 316
The /v Switch 316
The /a Switch 318

Single File Examples **319**
Simple Character Class Example 320
Find Protocols Example 320
Multiple File Example **321**
A Filelist Example **322**
Exercises **323**

Chapter 14: PowerGREP **325**

The PowerGREP Interface **325**
A Simple Find Example 326
The Replace Tab 328
The File Finder Tab 329
Syntax Coloring 330
Other Tabs 331
Metacharacters Supported **331**
Numeric Digits and Alphabetic Characters 332
Quantifiers 333
Back References 335
Alternation 339
Line Position Metacharacters 339
Word-Boundary Metacharacters 340
Lookahead and Lookbehind 342
Longer Examples **343**
Finding HTML Horizontal Rule Elements 343
Matching Time Example 346
Exercises **349**

Chapter 15: Wildcards in Microsoft Excel **351**

The Excel Find Interface **351**
The Wildcards Excel Supports **355**
Escaping Wildcard Characters 359
Using Wildcards in Data Forms **360**
Using Wildcards in Filters **362**
Exercises **363**

Chapter 16: Regular Expression Functionality In SQL Server 2000 **365**

Metacharacters Supported **366**
Using LIKE with Regular Expressions **366**
The % Metacharacter 366
The _ Metacharacter 372
Character Classes 373

Negated Character Classes 376
Using Full-Text Search 379
Using The CONTAINS Predicate 386
Document Filters on Image Columns 391
Exercises 391

Chapter 17: Using Regular Expressions with MySQL 393

Getting Started with MySQL 393
The Metacharacters MySQL Supports 396
Using the _ and % Metacharacters 397
Testing Matching of Literals: _ and % Metacharacters 400
Using the REGEXP Keyword and Metacharacters 401
Using Positional Metacharacters 404
Using Character Classes 406
Quantifiers 408
Social Security Number Example 410
Exercises 411

Chapter 18: Regular Expressions and Microsoft Access 413

The Interface to Metacharacters in Microsoft Access 413
Creating a Hard-Wired Query 414
Creating a Parameter Query 419
The Metacharacters Supported in Access 422
Using the ? Metacharacter 422
Using the * Metacharacter 423
Using the # Metacharacter 424
Using the # Character with Date/Time Data 425
Using Character Classes in Access 426
Exercises 428

Chapter 19: Regular Expressions In JScript and JavaScript 429

Using Regular Expressions in JavaScript and JScript 430
The RegExp Object 432
Attributes of the RegExp Object 438
The Other Properties of the RegExp Object 438
The test() Method of the RegExp Object 441
The exec() Method of the RegExp Object 441
The String Object 448
Metacharacters in JavaScript and JScript 451

Documenting JavaScript Regular Expressions **452**
SSN Validation Example **452**
Exercises **454**

Chapter 20: Regular Expressions and VBScript **455**

The RegExp Object and How to Use It **455**
The RegExp Object's Pattern Property 456
The RegExp Object's Global Property 458
The RegExp Object's IgnoreCase Property 462
The RegExp Object's Test() Method 464
The RegExp Object's Replace() Method 465
The RegExp Object's Execute() Method 467
Using the Match Object and the Matches Collection **471**
Supported Metacharacters **473**
Quantifiers 474
Positional Metacharacters 475
Character Classes 478
Word Boundaries 479
Lookahead 479
Grouping and Nongrouping Parentheses 482
Exercises **483**

Chapter 21: Visual Basic .NET and Regular Expressions **485**

The System.Text.RegularExpressions namespace **486**
A Simple Visual Basic .NET Example 486
The Classes of System.Text.RegularExpressions 490
The Regex Object 490
Using the Match Object and Matches Collection 492
Using the Match.Success Property and Match.NextMatch Method 495
The GroupCollection and Group Classes 497
The CaptureCollection and Capture Class 499
The RegexOptions Enumeration 502
Case-Insensitive Matching: The IgnoreCase Option 502
Multiline Matching: The Effect on the A and $ Metacharacters 505
Inline Documentation Using the IgnorePatternWhitespace Option 505
Right to Left Matching: The RightToLeft Option 507
The Metacharacters Supported in Visual Basic .NET **508**
Lookahead and Lookbehind 510
Exercises **510**

Chapter 22: C# and Regular Expressions 511

The Classes of the System.Text.RegularExpressions namespace 512
An Introductory Example 512
The Classes of System.Text.RegularExpressions 517
The Regex Class 517
The Options Property of the Regex Class 518
The Regex Class's RightToLeft Property 518
Regex Class Methods 518
The CompileToAssemblyO Method 519
The GetGroupNamesO Method 519
The GetGroupNumbersO Method 519
GroupNumberFromNameO and GroupNameFromNumberO Methods 519
The IsMatch() Method 520
The MatchO Method 521
The Matches() Method 522
The Replace() Method 526
The Split() Method 528
Using the Static Methods of the Regex Class 531
The IsMatch() Method as a Static 531
The Match() Method as a Static 531
The MatchesQ Method as a Static 531
The Replace() Method as a Static 532
The Split() Method as a Static 532
The Match and Matches Classes 532
The Match Class 532
The GroupCollection and Group Classes 536
The RegexOptions Class 539
The IgnorePatternWhitespace Option 539
Metacharacters Supported in Visual C# .NET 542
Using Named Groups 544
Using Back References 545
Exercise 547

Chapter 23: PHP and Regular Expressions 549

Getting Started with PHP 5.0 549
How PHP Structures Support for Regular Expressions 553
The ereg() Set of Functions 553
The ereg() Function 554
The ereg() Function with Three Arguments 556

The eregiQ Function 559

The ereg_replace() Function 561

The eregi_replace() Function 563

The split() Function 564

The splitiO Function 566

The sql_regcase() Function 567

Perl Compatible Regular Expressions 568

Pattern Delimiters in PCRE 568

Escaping Pattern Delimiters 570

Matching Modifiers in PCRE 570

Using the preg_match() Function 571

Using the preg_match_all() Function 574

Using the preg_grep() Function 576

Using the preg_quote() Function 579

Using the preg_replace() Function 579

Using the preg_replace_callback() Function 580

Using the preg_split() Function 580

The Metacharacters Supported in **PHP** 581

Supported Metacharacters with ereg() 582

Using POSIX Character Classes with **PHP** 582

Supported Metacharacters with PCRE 585

Positional Metacharacters 586

Character Classes in PHP 587

Documenting PHP Regular Expressions 589

Exercises **590**

Chapter 24: Regular Expressions in W3C XML Schema **591**

W3C XML Schema Basics 592

Tools for Using W3C XML Schema 592

Comparing XML Schema and DTDs 593

How Constraints Are Expressed in W3C **XML** Schema 598

W3C XML Schema Datatypes 599

Derivation by Restriction 602

Unicode and W3C XML Schema 604

Unicode Overview 604

Using Unicode Character Classes 605

Matching Decimal Numbers 606

Mixing Unicode Character Classes with Other Metacharacters 607

Unicode Character Blocks 608

Using Unicode Character Blocks 609

Metacharacters Supported in W3C **XML** Schema 612

Positional Metacharacters 613

Matching Numeric Digits 614
Alternation 615
Using the \w and \s Metacharacters 615
Escaping Metacharacters 616
Exercises 616

Chapter 25: Regular Expressions In Java 619

Introduction to the java.utii.regex Package 620
Obtaining and Installing Java 620
The Pattern Class 620
Using the matches() Method Statically 621
Two Simple Java Examples 621
The Properties (Fields) of the Pattern Class 629
The CASEJNSENSITIVE Flag 629
Using the COMMENTS Flag 630
The DOTALL Flag 632
The MULTILINE Flag 632
The UNICODE_CASE Flag 632
The UNIXJ.INES Flag 632
The Methods of the Pattern Class 633
The compileO Method 633
The flags() Method 633
The matcher() Method 633
The matches() Method 633
The pattern() Method 634
The split() Method 634
The Matcher Class 634
The appendReplacement() Method 635
The appendTail() Method 638
The end() Method 638
The find() Method 638
The group() Method 638
The groupCount() Method 639
The lookingAt() Method 639
The matches() Method 642
The pattern() Method 642
The replaceAII() Method 642
The replaceFirst() Method 644
The reset() Method 644
The start() Method 644
The PatternSyntaxException Class 644

Metacharacters Supported in the java.utii.regex Package 645
Using the \d Metacharacter 645
Character Classes 647
The POSIX Character Classes in the java.utii.regex Package 651
Unicode Character Classes and Character Blocks 652
Using Escaped Characters 653
Using Methods of the String Class 654
Using the matches() Method 654
Using the replaceFirst() Method 656
Using the replaceAIIQ Method 658
Using the split() Method 658
Exercises 658

Chapter 26: Regular Expressions In Perl 659

Obtaining and Installing Perl 659
Creating a Simple Perl Program 663
Basics of Perl Regular Expression Usage 667
Using the Perl Regular Expression Operators 667
Using the m// Operator 667
Using Other Regular Expression Delimiters 675
Matching Using Variable Substitution 676
Using the s/// Operator 678
Using s/// with the Global Modifier 679
Using s/// with the Default Variable 681
Using the split Operator 682
The Metacharacters Supported in Perl 684
Using Quantifiers in Perl 685
Using Positional Metacharacters 686
Captured Groups in Perl 687
Using Back References in Perl 689
Using Alternation 690
Using Character Classes in Perl 692
Using Lookahead 696
Using Lookbehind 698
Using the Regular Expression Matching Modes in Perl 700
Escaping Metacharacters 701
A Simple Perl Regex Tester 703
Exercises 705

Appendix A: Exercise Answers 707

Index 727