Introducing Regular Expressions

Michael Fitzgerald

O'REILLY®
Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

Introducing Regular Expressions
by Michael Fitzgerald

Copyright © 2012 Michael Fitzgerald. All rights reserved.
Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (*http://my.safaribooksonline.com*). For more information, contact our corporate/institutional sales department: 800-998-9938 or *corporate@oreilly.com*.

Editor: Simon St. Laurent
Production Editor: Holly Bauer
Proofreader: Julie Van Keuren

Indexer: Lucie Haskins
Cover Designer: Karen Montgomery
Interior Designer: David Futato
Illustrator: Rebecca Demarest

July 2012: First Edition.

Revision History for the First Edition:
2012-07-10 First release
See *http://oreilly.com/catalog/errata.csp?isbn=9781449392680* for release details.

Nutshell Handbook, the Nutshell Handbook logo, and the O'Reilly logo are registered trademarks of O'Reilly Media, Inc. *Introducing Regular Expressions*, the image of a fruit bat, and related trade dress are trademarks of O'Reilly Media, Inc.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and O'Reilly Media, Inc., was aware of a trademark claim, the designations have been printed in caps or initial caps.

While every precaution has been taken in the preparation of this book, the publisher and authors assume no responsibility for errors or omissions, or for damages resulting from the use of the information contained herein.

ISBN: 978-1-449-39268-0

[LSI]

1341860829

Table of Contents

Preface ..... **vii**

1. What Is a Regular Expression? **1**
Getting Started with Regexpal 2
Matching a North American Phone Number 2
Matching Digits with a Character Class 4
Using a Character Shorthand 5
Matching Any Character 5
Capturing Groups and Back References 6
Using Quantifiers 6
Quoting Literals 8
A Sample of Applications 9
What You Learned in Chapter 1 11
Technical Notes 11

2. Simple Pattern Matching **13**
Matching String Literals 15
Matching Digits 15
Matching Non-Digits 17
Matching Word and Non-Word Characters 18
Matching Whitespace 20
Matching Any Character, Once Again 22
Marking Up the Text 24
Using *sed* to Mark Up Text 24
Using Perl to Mark Up Text 25
What You Learned in Chapter 2 27
Technical Notes 27

3. Boundaries **29**
The Beginning and End of a Line 29
Word and Non-word Boundaries 31

Other Anchors 33
Quoting a Group of Characters as Literals 34
Adding Tags 34
Adding Tags with *sed* 36
Adding Tags with Perl 37
What You Learned in Chapter 3 38
Technical Notes 38

4. Alternation, Groups, and Backreferences 41
Alternation 41
Subpatterns 45
Capturing Groups and Backreferences 46
Named Groups 48
Non-Capturing Groups 49
Atomic Groups 50
What You Learned in Chapter 4 50
Technical Notes 51

5. Character Classes 53
Negated Character Classes 55
Union and Difference 56
POSIX Character Classes 56
What You Learned in Chapter 5 59
Technical Notes 60

6. Matching Unicode and Other Characters 61
Matching a Unicode Character 62
Using *vim* 63
Matching Characters with Octal Numbers 64
Matching Unicode Character Properties 65
Matching Control Characters 68
What You Learned in Chapter 6 70
Technical Notes 71

7. Quantifiers 73
Greedy, Lazy, and Possessive 74
Matching with *, +, and ? 74
Matching a Specific Number of Times 75
Lazy Quantifiers 76
Possessive Quantifiers 77
What You Learned in Chapter 7 78
Technical Notes 79

8. Lookarounds 81
Positive Lookaheads 81
Negative Lookaheads 84
Positive Lookbehinds 85
Negative Lookbehinds 85
What You Learned in Chapter 8 86
Technical Notes 86

9. Marking Up a Document with HTML 87
Matching Tags 87
Transforming Plain Text with *sed* 88
Substitution with *sed* 89
Handling Roman Numerals with *sed* 90
Handling a Specific Paragraph with *sed* 91
Handling the Lines of the Poem with *sed* 91
Appending Tags 92
Using a Command File with *sed* 92
Transforming Plain Text with Perl 94
Handling Roman Numerals with Perl 95
Handling a Specific Paragraph with Perl 96
Handling the Lines of the Poem with Perl 96
Using a File of Commands with Perl 97
What You Learned in Chapter 9 98
Technical Notes 98

10. The End of the Beginning 101
Learning More 102
Notable Tools, Implementations, and Libraries 103
Perl 103
PCRE 103
Ruby (Oniguruma) 104
Python 104
RE2 105
Matching a North American Phone Number 105
Matching an Email Address 105
What You Learned in Chapter 10 106

Appendix: Regular Expression Reference 107

Regular Expression Glossary 123

Index 129