Pro Perl Parsing

Christopher M. Frenz

Apress®

Pro Perl Parsing

Copyright © 2005 by Christopher M. Frenz

Lead Editors: Jason Gilmore and Matthew Moodie
Technical Reviewer: Teodor Zlatanov
Editorial Board: Steve Anglin, Dan Appleman, Ewan Buckingham, Gary Cornell, Tony Davis, Jason Gilmore, Jonathan Hassell, Chris Mills, Dominic Shakeshaft, Jim Sumser
Associate Publisher: Grace Wong
Project Manager: Beth Christmas
Copy Edit Manager: Nicole LeClerc
Copy Editor: Kim Wimpsett
Assistant Production Director: Kari Brooks-Copony
Production Editor: Laura Cheu
Compositor: Linda Weidemann, Wolf Creek Press
Proofreader: Nancy Sixsmith
Indexer: Tim Tate
Artist: Wordstop Technologies Pvt. Ltd., Chennai
Cover Designer: Kurt Krames
Manufacturing Manager: Tom Debolski

Library of Congress Cataloging-in-Publication Data

Frenz, Christopher.
Pro Perl parsing / Christopher M. Frenz.
p. cm.
Includes index.
ISBN 1-59059-504-1 (hardcover : alk. paper)
1. Perl (Computer program language) 2. Natural language processing (Computer science) I. Title.
QA76.73.P22F72 2005
005.13'3--dc22
2005017530

All rights reserved. No part of this work may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage or retrieval system, without the prior written permission of the copyright owner and the publisher.

Printed and bound in the United States of America 9 8 7 6 5 4 3 2 1

Trademarked names may appear in this book. Rather than use a trademark symbol with every occurrence of a trademarked name, we use the names only in an editorial fashion and to the benefit of the trademark owner, with no intention of infringement of the trademark.

Distributed to the book trade worldwide by Springer-Verlag New York, Inc., 233 Spring Street, 6th Floor, New York, NY 10013. Phone 1-800-SPRINGER, fax 201-348-4505, e-mail orders-ny@springer-sbm.com, or visit http://www.springeronline.com.

For information on translations, please contact Apress directly at 2560 Ninth Street, Suite 219, Berkeley, CA 94710. Phone 510-549-5930, fax 510-549-5939, e-mail info@apress.com, or visit http://www.apress.com.

The information in this book is distributed on an “as is” basis, without warranty. Although every precaution has been taken in the preparation of this work, neither the author(s) nor Apress shall have any liability to any person or entity with respect to any loss or damage caused or alleged to be caused directly or indirectly by the information contained in this work.

The source code for this book is available to readers at http://www.apress.com in the Downloads section.

Contents

About the Author xiii
About the Technical Reviewer xv
Acknowledgments xvii
Introduction xix

■CHAPTER 1 **Parsing and Regular Expression Basics** 1

Parsing and Lexing 2
Parse::Lex 4
Using Regular Expressions 6
A State Machine 7
Pattern Matching 12
Quantifiers 14
Predefined Subpatterns 15
Posix Character Classes 16
Modifiers 17
Assertions 20
Capturing Substrings 24
Substitution 26
Troubleshooting Regexes 26
GraphViz::Regex 27
Using Regexp::Common 28
Regexp::Common::Balanced 29
Regexp::Common::Comments 30
Regexp::Common::Delimited 30
Regexp::Common::List 30
Regexp::Common::Net 31
Regexp::Common::Number 31
Universal Flags 32
Standard Usage 32
Subroutine-Based Usage 33
In-Line Matching and Substitution 34
Creating Your Own Expressions 35
Summary 36

CHAPTER 2 Grammars

CHAPTER 2 Grammars 37
Introducing Generative Grammars 38
Grammar Recipes 39
Sentence Construction 41
Introducing the Chomsky Method. 42
Type 1 Grammars (Context-Sensitive Grammars) 44
Type 2 Grammars (Context-Free Grammars) 48
Type 3 Grammars (Regular Grammars) 54
Using Perl to Generate Sentences 55
Perl-Based Sentence Generation 56
Avoiding Common Grammar Errors 59
Generation vs. Parsing 60
Summary 61
CHAPTER 3 Parsing Basics 63
Exploring Common Parser Characteristics 64
Introducing Bottom-Up Parsers. 65
Coding a Bottom-Up Parser in Perl 68
Introducing Top-Down Parsers 73
Coding a Top-Down Parser in Perl 74
Using Parser Applications 78
Programming a Math Parser 80
Summary 83
CHAPTER 4 Using Parse::Yapp. 85
Creating the Grammar File 85
The Header Section 86
The Rule Section 87
The Footer Section. 88
Using yapp 94
The -v Flag 99
The -m Flag 103
The -s Flag 103
Using the Generated Parser Module 104
Evaluating Dynamic Content 105
Summary 108

CHAPTER 5 Performing Recursive-Descent Parsing with Parse::RecDescent 109

Examining the Module's Basic Functionality 109
Constructing Rules 111
    Subrules 112
Introducing Actions 115
    @item and %item 116
    @arg and %arg 117
    $return 118
    $text 120
    $thisline and $prevline 120
    $thiscolumn and $prevcolumn 121
    $thisoffset and $prevoffset 121
    $thisparser 121
    $thisrule and $thisprod 122
    $score 122
Introducing Startup Actions 122
Introducing Autoactions 124
Introducing Autotrees 125
Introducing Autostubbing 127
Introducing Directives 128
    <commit> and <uncommit> 129
    <reject> 130
    <skip> 131
    <resync> 132
    <error> 132
    <defer> 132
    <perl ....> 133
    <score> and <autoscore> 134
Precompiling the Parser 135
Summary 135

CHAPTER 6 **Accessing Web Data with HTML::TreeBuilder** 137

Introducing HTML Basics 137
Specifying Titles 138
Specifying Headings 139
Specifying Paragraphs 140
Specifying Lists 141
Embedding Links 142
Understanding the Nested Nature of HTML. 143
Accessing Web Content with LWP 145
Using LWP::Simple 146
Using LWP 146
Using HTML::TreeBuilder 150
Controlling TreeBuilder Parser Attributes 152
Searching Through the Parse Tree 154
Understanding the Fair Use of Information Extraction Scripts 158
Summary 159

CHAPTER 7 **Parsing XML Documents with XML::LibXML and XML::SAX** 161

Understanding the Nature and Structure of XML Documents 163
The Document Prolog 164
Elements and the Document Body 166
Introducing Web Services 172
XML-RPC 173
RPC::XML 173
Simple Object Access Protocol (SOAP). 174
SOAP::Lite 175
Parsing with XML::LibXML 177
Using DOM to Parse XML 177
Parsing with XML::SAX::ParserFactory 179
Summary 182

CHAPTER 8 **Introducing Miscellaneous Parsing Modules** 185

Using Text::Balanced 185
Using extract_delimited 186
Using extract_bracketed 188
Using extract_codeblock 189

Using extract_quotelike 190
Using extract_variable 191
Using extract_multiple 192
Using Date::Parse 193
Using XML::RSS::Parser 194
Using Math::Expression 197
Summary 199

CHAPTER 9 Finding Solutions to Miscellaneous Parsing Problems 201

Parsing Command-Line Arguments 201
Parsing Configuration Files 204
Refining Searches 205
Formatting Output 212
Summary 214

CHAPTER 10 Performing Text and Data Mining 217

Introducing Data Mining Basics 218
Introducing Descriptive Modeling 219
Clustering 219
Summarization 220
Association Rules 221
Sequence Discovery 224
Introducing Predictive Modeling 224
Classification 225
Regression 225
Time Series Analysis 228
Prediction 229
Summary 241

INDEX 243