

Analyzing Text with the Natural Language Toolkit

Natural Language Processing with Python



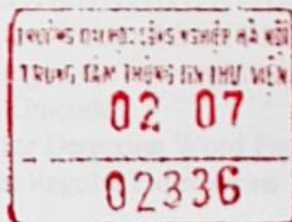
O'REILLY®

Steven Bird, Ewan Klein & Edward Loper

Natural Language Processing with Python

1. Language Processing and Python	1
1.1 Computing with Language: Text and Statistics	1
1.2 A Closer Look at Python: Text as Lists of Words	10
1.3 Computing with Language: Simple Statistics	16
1.4 Back to Python: Making Decisions and Taking Control	22
1.5 Automatic Natural Language Understanding	27
1.6 Summary	33
1.7 Further Reading	33
1.8 Exercises	34

Steven Bird, Ewan Klein, and Edward Loper



O'REILLY®

Beijing • Boston • Farnham • Sebastopol • Tokyo

Natural Language Processing with Python

by Steven Bird, Ewan Klein, and Edward Loper

Copyright © 2009 Steven Bird, Ewan Klein, and Edward Loper. All rights reserved.

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: Julie Steele

Production Editor: Loranah Dimant

Copyeditor: Genevieve d'Entremont

Proofreader: Loranah Dimant

Indexer: Ellen Troutman Zaig

Cover Designer: Karen Montgomery

Interior Designer: David Futato

Illustrator: Robert Romano

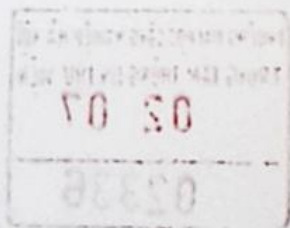
Printing History:

June 2009: First Edition.

Nutshell Handbook, the Nutshell Handbook logo, and the O'Reilly logo are registered trademarks of O'Reilly Media, Inc. *Natural Language Processing with Python*, the image of a right whale, 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: 978-0-596-51649-9

[LSI]

1290104768

[3/19]

Table of Contents

Preface	ix
1. Language Processing and Python	1
1.1 Computing with Language: Texts and Words	1
1.2 A Closer Look at Python: Texts as Lists of Words	10
1.3 Computing with Language: Simple Statistics	16
1.4 Back to Python: Making Decisions and Taking Control	22
1.5 Automatic Natural Language Understanding	27
1.6 Summary	33
1.7 Further Reading	33
1.8 Exercises	34
2. Accessing Text Corpora and Lexical Resources	39
2.1 Accessing Text Corpora	39
2.2 Conditional Frequency Distributions	52
2.3 More Python: Reusing Code	56
2.4 Lexical Resources	59
2.5 WordNet	67
2.6 Summary	73
2.7 Further Reading	73
2.8 Exercises	74
3. Processing Raw Text	79
3.1 Accessing Text from the Web and from Disk	80
3.2 Strings: Text Processing at the Lowest Level	87
3.3 Text Processing with Unicode	93
3.4 Regular Expressions for Detecting Word Patterns	97
3.5 Useful Applications of Regular Expressions	102
3.6 Normalizing Text	107
3.7 Regular Expressions for Tokenizing Text	109
3.8 Segmentation	112
3.9 Formatting: From Lists to Strings	116

3.10	Summary	121
3.11	Further Reading	122
3.12	Exercises	123
4.	Writing Structured Programs	129
4.1	Back to the Basics	130
4.2	Sequences	133
4.3	Questions of Style	138
4.4	Functions: The Foundation of Structured Programming	142
4.5	Doing More with Functions	149
4.6	Program Development	154
4.7	Algorithm Design	160
4.8	A Sample of Python Libraries	167
4.9	Summary	172
4.10	Further Reading	173
4.11	Exercises	173
5.	Categorizing and Tagging Words	179
5.1	Using a Tagger	179
5.2	Tagged Corpora	181
5.3	Mapping Words to Properties Using Python Dictionaries	189
5.4	Automatic Tagging	198
5.5	N-Gram Tagging	202
5.6	Transformation-Based Tagging	208
5.7	How to Determine the Category of a Word	210
5.8	Summary	213
5.9	Further Reading	214
5.10	Exercises	215
6.	Learning to Classify Text	221
6.1	Supervised Classification	221
6.2	Further Examples of Supervised Classification	233
6.3	Evaluation	237
6.4	Decision Trees	242
6.5	Naive Bayes Classifiers	245
6.6	Maximum Entropy Classifiers	250
6.7	Modeling Linguistic Patterns	254
6.8	Summary	256
6.9	Further Reading	256
6.10	Exercises	257
7.	Extracting Information from Text	261
7.1	Information Extraction	261

7.2	Chunking	264
7.3	Developing and Evaluating Chunkers	270
7.4	Recursion in Linguistic Structure	277
7.5	Named Entity Recognition	281
7.6	Relation Extraction	284
7.7	Summary	285
7.8	Further Reading	286
7.9	Exercises	286
8.	Analyzing Sentence Structure	291
8.1	Some Grammatical Dilemmas	292
8.2	What's the Use of Syntax?	295
8.3	Context-Free Grammar	298
8.4	Parsing with Context-Free Grammar	302
8.5	Dependencies and Dependency Grammar	310
8.6	Grammar Development	315
8.7	Summary	321
8.8	Further Reading	322
8.9	Exercises	322
9.	Building Feature-Based Grammars	327
9.1	Grammatical Features	327
9.2	Processing Feature Structures	337
9.3	Extending a Feature-Based Grammar	343
9.4	Summary	355
9.5	Further Reading	356
9.6	Exercises	357
10.	Analyzing the Meaning of Sentences	361
10.1	Natural Language Understanding	361
10.2	Propositional Logic	368
10.3	First-Order Logic	372
10.4	The Semantics of English Sentences	384
10.5	Discourse Semantics	397
10.6	Summary	402
10.7	Further Reading	403
10.8	Exercises	404
11.	Managing Linguistic Data	407
11.1	Corpus Structure: A Case Study	407
11.2	The Life Cycle of a Corpus	412
11.3	Acquiring Data	416
11.4	Working with XML	425

11.5	Working with Toolbox Data	431
11.6	Describing Language Resources Using OLAC Metadata	435
11.7	Summary	437
11.8	Further Reading	437
11.9	Exercises	438
Afterword: The Language Challenge		441
Bibliography		449
NLTK Index		459
General Index		463

Preface

This is a book about Natural Language Processing. By “natural language” we mean a language that is used for everyday communication by humans; languages such as English, Hindi, or Portuguese. In contrast to artificial languages such as programming languages and mathematical notations, natural languages have evolved as they pass from generation to generation, and are hard to pin down with explicit rules. We will take Natural Language Processing—or NLP for short—in a wide sense to cover any kind of computer manipulation of natural language. At one extreme, it could be as simple as counting word frequencies to compare different writing styles. At the other extreme, NLP involves “understanding” complete human utterances, at least to the extent of being able to give useful responses to them.

Technologies based on NLP are becoming increasingly widespread. For example, phones and handheld computers support predictive text and handwriting recognition; web search engines give access to information locked up in unstructured text; machine translation allows us to retrieve texts written in Chinese and read them in Spanish. By providing more natural human-machine interfaces, and more sophisticated access to stored information, language processing has come to play a central role in the multilingual information society.

This book provides a highly accessible introduction to the field of NLP. It can be used for individual study or as the textbook for a course on natural language processing or computational linguistics, or as a supplement to courses in artificial intelligence, text mining, or corpus linguistics. The book is intensely practical, containing hundreds of fully worked examples and graded exercises.

The book is based on the Python programming language together with an open source library called the *Natural Language Toolkit* (NLTK). NLTK includes extensive software, data, and documentation, all freely downloadable from <http://www.nltk.org/>. Distributions are provided for Windows, Macintosh, and Unix platforms. We strongly encourage you to download Python and NLTK, and try out the examples and exercises along the way.