

---

# Natural Language Processing with Python

*Steven Bird, Ewan Klein, and Edward Loper*

**O'REILLY®**

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

---

# Table of Contents

<b>Preface .....</b>	<b>ix</b>
<b>1. Language Processing and Python .....</b>	<b>1</b>
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	34
1.8 Exercises	35
<b>2. Accessing Text Corpora and Lexical Resources .....</b>	<b>39</b>
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
<b>3. Processing Raw Text .....</b>	<b>79</b>
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
<b>4.</b>	<b>Writing Structured Programs</b>	<b>129</b>
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
<b>5.</b>	<b>Categorizing and Tagging Words</b>	<b>179</b>
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
<b>6.</b>	<b>Learning to Classify Text</b>	<b>221</b>
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
<b>7.</b>	<b>Extracting Information from Text</b>	<b>261</b>
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
<b>8.</b>	<b>Analyzing Sentence Structure</b>	<b>291</b>
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
<b>9.</b>	<b>Building Feature-Based Grammars</b>	<b>327</b>
9.1	Grammatical Features	327
9.2	Processing Feature Structures	337
9.3	Extending a Feature-Based Grammar	344
9.4	Summary	356
9.5	Further Reading	357
9.6	Exercises	358
<b>10.</b>	<b>Analyzing the Meaning of Sentences</b>	<b>361</b>
10.1	Natural Language Understanding	361
10.2	Propositional Logic	368
10.3	First-Order Logic	372
10.4	The Semantics of English Sentences	385
10.5	Discourse Semantics	397
10.6	Summary	402
10.7	Further Reading	403
10.8	Exercises	404
<b>11.</b>	<b>Managing Linguistic Data</b>	<b>407</b>
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
<b>Afterword: The Language Challenge</b> .....	<b>441</b>
<b>Bibliography</b> .....	<b>449</b>
<b>NLTK Index</b> .....	<b>459</b>
<b>General Index</b> .....	<b>463</b>