

Krishna Bhavsar, Naresh Kumar,
Pratap Dangeti

Natural Language Processing with Python Cookbook

Over 60 recipes to implement text analytics solutions
using deep learning principles



Packt>

Natural Language Processing with Python Cookbook

Natural language processing (NLP) is a field of computer science, artificial intelligence, and computational linguistics concerned with the interaction between computers and human (natural) languages; in particular, it's about programming computers to process large natural language corpora.

This book includes unique recipes that will teach you various aspects of performing NLP with NLTK—the leading Python platform for the task. You will come across various recipes in the book, covering (among other topics) natural language understanding, NLP, and syntactic analysis. You will learn how to understand language, plan sentences, and work around various ambiguities. You will learn how to efficiently use NLTK and implement text classification, identify parts of speech, tag words, and more. You will also learn how to analyze sentence structures and master lexical analysis, syntactic and semantic analysis, pragmatic analysis, and the application of deep learning techniques.

By the end of this book, you will have all the knowledge you need to implement NLP with Python.

Things you will learn:

- Explore various corpora that is available with NLTK and understand how to use WordNet corpus
- Understand how to manage and process raw text like HTML, RSS, PDF, word documents, and so on
- Learn how to pre-process raw text using techniques like tokenization, stemming, spell checker, and so on and also implement them on your own using regular expressions
- Understand the basics of pattern matching in textual analytics with regular expressions
- Learn to use and write your own POS taggers and grammars
- Learn how to perform named entity (NE) extraction and also learn parsers like RD, shift reduce, chart parsers
- Generate text from Shakespeare's writing using LSTM
- Utilize the BABI dataset and LSTM to model episodes
- Develop chat bot in generative way with deep learning

Natural Language Processing with Python Cookbook

Krishna Bhavsar, Nareesh Kumar, Pratap Dangeti



Natural Language Processing with Python Cookbook

Over 60 recipes to implement text analytics solutions using deep learning principles

Krishna Bhavsar
Naresh Kumar
Pratap Dangeti

Packt>

BIRMINGHAM - MUMBAI

Natural Language Processing with Python Cookbook

Copyright © 2017 Packt Publishing

All rights reserved. No part of this book may be reproduced, stored in a retrieval system, or transmitted in any form or by any means, without the prior written permission of the publisher, except in the case of brief quotations embedded in critical articles or reviews.

Every effort has been made in the preparation of this book to ensure the accuracy of the information presented. However, the information contained in this book is sold without warranty, either express or implied. Neither the authors, nor Packt Publishing, and its dealers and distributors will be held liable for any damages caused or alleged to be caused directly or indirectly by this book.

Packt Publishing has endeavored to provide trademark information about all of the companies and products mentioned in this book by the appropriate use of capitals. However, Packt Publishing cannot guarantee the accuracy of this information.

First published: November 2017

Production reference: 1221117

Published by Packt Publishing Ltd.

Livery Place

35 Livery Street

Birmingham

B3 2PB, UK.

ISBN 978-1-78728-932-1

www.packtpub.com

Credits

Authors

Krishna Bhavsar
Naresh Kumar
Pratap Dangeti

Copy Editor

Vikrant Phadkay

Reviewer

Juan Tomas Oliva Ramos

Project Coordinator

Nidhi Joshi

Commissioning Editor

Veena Pagare

Proofreader

Safis Editing

Acquisition Editor

Aman Singh

Indexer

Tejal Daruwale Soni

Content Development Editor

Aishwarya Pandere

Graphics

Tania Dutta

Technical Editors

Dinesh Pawar
Suwarna Rajput

Production Coordinator

Shraddha Falebhai

About the Authors

Krishna Bhavsar has spent around 10 years working on natural language processing, social media analytics, and text mining in various industry domains such as hospitality, banking, healthcare, and more. He has worked on many different NLP libraries such as Stanford CoreNLP, IBM's SystemText and BigInsights, GATE, and NLTK to solve industry problems related to textual analysis. He has also worked on analyzing social media responses for popular television shows and popular retail brands and products. He has also published a paper on sentiment analysis augmentation techniques in 2010 NAACL. He recently created an NLP pipeline/toolset and open sourced it for public use. Apart from academics and technology, Krishna has a passion for motorcycles and football. In his free time, he likes to travel and explore. He has gone on pan-India road trips on his motorcycle and backpacking trips across most of the countries in South East Asia and Europe.

First and foremost, I would like to thank my mother for being the biggest motivating force and a rock-solid support system behind all my endeavors in life. I would like to thank the management team at Synerzip and all my friends for being supportive of me on this journey. Last but not least, special thanks to Ram and Dorothy for keeping me on track during this professionally difficult year.

Naresh Kumar has more than a decade of professional experience in designing, implementing, and running very-large-scale Internet applications in Fortune Top 500 companies. He is a full-stack architect with hands-on experience in domains such as e-commerce, web hosting, healthcare, big data and analytics, data streaming, advertising, and databases. He believes in open source and contributes to it actively. Naresh keeps himself up-to-date with emerging technologies, from Linux systems internals to frontend technologies. He studied in BITS-Pilani, Rajasthan with dual degree in computer science and economics.

Pratap Dangeti develops machine learning and deep learning solutions for structured, image, and text data at TCS, in its research and innovation lab in Bangalore. He has acquired a lot of experience in both analytics and data science. He received his master's degree from IIT Bombay in its industrial engineering and operations research program. Pratap is an artificial intelligence enthusiast. When not working, he likes to read about next-gen technologies and innovative methodologies. He is also the author of the book *Statistics for Machine Learning* by Packt.

I would like to thank my mom, Lakshmi, for her support throughout my career and in writing this book. I dedicate this book to her. I also thank my family and friends for their encouragement, without which it would not have been possible to write this book.

About the Reviewer

Juan Tomas Oliva Ramos is an environmental engineer from the University of Guanajuato, Mexico, with a master's degree in administrative engineering and quality. He has more than 5 years of experience in the management and development of patents, technological innovation projects, and the development of technological solutions through the statistical control of processes.

He has been a teacher of statistics, entrepreneurship, and the technological development of projects since 2011. He became an entrepreneur mentor and started a new department of technology management and entrepreneurship at Instituto Tecnológico Superior de Purísima del Rincón Guanajuato, Mexico.

Juan is an Alfaomega reviewer and has worked on the book *Wearable Designs for Smart Watches, Smart TVs and Android Mobile Devices*.

Juan has also developed prototypes through programming and automation technologies for the improvement of operations, which have been registered for patents.

I want to thank God for giving me wisdom and humility to review this book.

I thank Packt for giving me the opportunity to review this amazing book and to collaborate with a group of committed people

I want to thank my beautiful wife, Brenda, our two magic princesses (Maria Regina and Maria Renata) and our next member (Angel Tadeo), all of you, give me the strength, happiness, and joy to start a new day. Thanks for being my family.

www.PacktPub.com

For support files and downloads related to your book, please visit www.PacktPub.com. Did you know that Packt offers eBook versions of every book published, with PDF and ePub files available? You can upgrade to the eBook version at www.PacktPub.com and as a print book customer, you are entitled to a discount on the eBook copy. Get in touch with us at service@packtpub.com for more details. At www.PacktPub.com, you can also read a collection of free technical articles, sign up for a range of free newsletters and receive exclusive discounts and offers on Packt books and eBooks.



<https://www.packtpub.com/mapt>

Get the most in-demand software skills with Mapt. Mapt gives you full access to all Packt books and video courses, as well as industry-leading tools to help you plan your personal development and advance your career.

Why subscribe?

- Fully searchable across every book published by Packt
- Copy and paste, print, and bookmark content
- On demand and accessible via a web browser

Customer Feedback

Thanks for purchasing this Packt book. At Packt, quality is at the heart of our editorial process. To help us improve, please leave us an honest review on this book's Amazon page at <https://www.amazon.com/dp/178728932X>. If you'd like to join our team of regular reviewers, you can email us at customerreviews@packtpub.com. We award our regular reviewers with free eBooks and videos in exchange for their valuable feedback. Help us be relentless in improving our products!

Table of Contents

Preface	1
<hr/>	
Chapter 1: Corpus and WordNet	9
<hr/>	
Introduction	9
Accessing in-built corpora	10
How to do it...	10
Download an external corpus, load it, and access it	13
Getting ready	13
How to do it...	13
How it works...	15
Counting all the wh words in three different genres in the Brown corpus	16
Getting ready	16
How to do it...	16
How it works...	18
Explore frequency distribution operations on one of the web and chat text corpus files	18
Getting ready	19
How to do it...	19
How it works...	21
Take an ambiguous word and explore all its senses using WordNet	21
Getting ready	22
How to do it...	22
How it works...	25
Pick two distinct synsets and explore the concepts of hyponyms and hypernyms using WordNet	26
Getting ready	26
How to do it...	26
How it works...	29
Compute the average polysemy of nouns, verbs, adjectives, and adverbs according to WordNet	29
Getting ready	30
How to do it...	30
How it works...	31
<hr/>	
Chapter 2: Raw Text, Sourcing, and Normalization	33
<hr/>	
Introduction	33

The importance of string operations	34
Getting ready...	34
How to do it...	34
How it works...	36
Getting deeper with string operations	36
How to do it...	36
How it works...	39
Reading a PDF file in Python	39
Getting ready	39
How to do it...	40
How it works...	41
Reading Word documents in Python	42
Getting ready...	42
How to do it...	42
How it works...	45
Taking PDF, DOCX, and plain text files and creating a user-defined corpus from them	46
Getting ready	46
How to do it...	47
How it works...	49
Read contents from an RSS feed	50
Getting ready	50
How to do it...	50
How it works...	52
HTML parsing using BeautifulSoup	52
Getting ready	53
How to do it...	53
How it works...	55
Chapter 3: Pre-Processing	57
Introduction	57
Tokenization – learning to use the inbuilt tokenizers of NLTK	58
Getting ready	58
How to do it...	58
How it works...	60
Stemming – learning to use the inbuilt stemmers of NLTK	61
Getting ready	61
How to do it...	61
How it works...	63
Lemmatization – learning to use the WordnetLemmatizer of NLTK	63

Getting ready	63
How to do it...	64
How it works...	66
Stopwords – learning to use the stopwords corpus and seeing the difference it can make	66
Getting ready	66
How to do it...	66
How it works...	69
Edit distance – writing your own algorithm to find edit distance between two strings	69
Getting ready	69
How to do it...	70
How it works...	72
Processing two short stories and extracting the common vocabulary between two of them	72
Getting ready	72
How to do it...	73
How it works...	78
Chapter 4: Regular Expressions	79
Introduction	79
Regular expression – learning to use *, +, and ?	80
Getting ready	80
How to do it...	80
How it works...	82
Regular expression – learning to use \$ and ^, and the non-start and non-end of a word	82
Getting ready	83
How to do it...	83
How it works...	85
Searching multiple literal strings and substring occurrences	86
Getting ready	86
How to do it...	86
How it works...	88
Learning to create date regex and a set of characters or ranges of character	88
How to do it...	88
How it works...	90
Find all five-character words and make abbreviations in some sentences	91

How to do it...	91
How it works...	92
Learning to write your own regex tokenizer	92
Getting ready	92
How to do it...	93
How it works...	94
Learning to write your own regex stemmer	94
Getting ready	94
How to do it...	95
How it works...	96
Chapter 5: POS Tagging and Grammars	97
<hr/>	
Introduction	97
Exploring the in-built tagger	98
Getting ready	98
How to do it...	98
How it works...	99
Writing your own tagger	100
Getting ready	100
How to do it...	101
How it works...	102
Training your own tagger	107
Getting ready	107
How to do it...	107
How it works...	109
Learning to write your own grammar	111
Getting ready	112
How to do it...	112
How it works...	113
Writing a probabilistic CFG	115
Getting ready	115
How to do it...	116
How it works...	117
Writing a recursive CFG	119
Getting ready	120
How to do it...	120
How it works...	122
Chapter 6: Chunking, Sentence Parse, and Dependencies	125
<hr/>	
Introduction	125

Using the built-in chunker	125
Getting ready	126
How to do it...	126
How it works...	127
Writing your own simple chunker	128
Getting ready	130
How to do it...	130
How it works...	131
Training a chunker	133
Getting ready	134
How to do it...	134
How it works...	135
Parsing recursive descent	136
Getting ready	137
How to do it...	137
How it works...	139
Parsing shift-reduce	140
Getting ready	140
How to do it...	140
How it works...	142
Parsing dependency grammar and projective dependency	143
Getting ready	144
How to do it...	144
How it works...	145
Parsing a chart	146
Getting ready	147
How to do it...	147
How it works...	148
Chapter 7: Information Extraction and Text Classification	151
Introduction	151
Understanding named entities	152
Using inbuilt NERs	153
Getting ready	154
How to do it...	154
How it works...	156
Creating, inversing, and using dictionaries	157
Getting ready	157
How to do it...	157
How it works...	159

Choosing the feature set	163
Getting ready	163
How to do it...	163
How it works...	165
Segmenting sentences using classification	168
Getting ready	168
How to do it...	168
How it works...	170
Classifying documents	172
Getting ready	172
How to do it...	172
How it works...	174
Writing a POS tagger with context	177
Getting ready	177
How to do it...	177
How it works...	179
Chapter 8: Advanced NLP Recipes	183
Introduction	183
Creating an NLP pipeline	184
Getting ready	185
How to do it...	185
How it works...	187
Solving the text similarity problem	192
Getting ready	193
How to do it...	193
How it works...	195
Identifying topics	199
Getting ready	199
How to do it...	199
How it works...	201
Summarizing text	205
Getting ready	205
How to do it...	205
How it works...	207
Resolving anaphora	209
Getting ready	210
How to do it...	210
How it works...	212
Disambiguating word sense	215