

**Jalaj Thanaki**

Foreword by:

**Sarita Arora**

Chief Analytics Officer, SMECorner, Mumbai, India

# Python Natural Language Processing

Explore NLP with machine learning and deep learning techniques



**Packt**>

# Python Natural Language Processing

This book starts off by laying the foundation of Natural Language Processing (NLP) and why Python is one of the best options for building an NLP-based expert system with advantages such as community support and the availability of frameworks. It also gives you a better understanding of the available free forms of corpus and different types of dataset. After this, you will know how to choose a dataset for NLP applications and find the right NLP techniques to process sentences in datasets and understand their structure. You will also learn how to tokenize different parts of sentences and look at the ways to analyze them.

During the course of the book, you will explore the semantic as well as syntactic analysis of text. You will understand how to solve various ambiguities in processing human language and will come across various scenarios while performing text analysis.

You will learn the very basics of getting the environment ready for NLP, move on to the initial setup, and then quickly understand sentences and language parts. You will learn the power of machine learning and deep learning to extract information from text data.

By the end of the book, you will have a clear understanding of NLP and will have worked on multiple examples that implement NLP in the real world.

## Things you will learn:

- Focus on Python programming paradigms, which are used to develop NLP applications
- Understand corpus analysis and different types of data attribute
- Learn NLP using Python libraries such as NLTK, Polyglot, SpaCy, Stanford CoreNLP, and so on
- Learn about features extraction and feature selection as part of features engineering
- Explore the advantages of vectorization in deep learning
- Get a better understanding of the architecture of a rule-based system
- Optimize and fine-tune supervised and unsupervised machine learning algorithms for NLP problems
- Identify deep learning techniques for natural language processing and natural language generation problems

**Packt**  
www.packtpub.com

\$ 49.99 US  
£ 41.99 UK

Prices do not include local sales  
Tax or VAT where applicable



# Python Natural Language Processing

Jalaj Thanaki



# Python Natural Language Processing

Explore NLP with machine learning and deep learning techniques

**Jalaj Thanaki**

**Packt**>  

---

**BIRMINGHAM - MUMBAI**

# Python Natural Language Processing

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 author, 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: July 2017

Production reference: 1280717

Published by Packt Publishing Ltd.

Livery Place

35 Livery Street

Birmingham

B3 2PB, UK.

ISBN 978-1-78712-142-3

[www.packtpub.com](http://www.packtpub.com)

# Credits

**Author**

Jalaj Thanaki

**Copy Editor**

Safis Editing

**Reviewers**

Devesh Raj

Gayetri Thakur

Prabhanjan Tattar

Chirag Mahapatra

**Project Coordinator**

Manthan Patel

**Commissioning Editor**

Veena Pagare

**Proofreader**

Safis Editing

**Acquisition Editor**

Aman Singh

**Indexer**

Tejal Daruwale Soni

**Content Development Editor**

Jagruti Babaria

**Production Coordinator**

Deepika Naik

**Technical Editor**

Sayli Nikalje



# Foreword

Data science is rapidly changing the world and the way we do business --be it retail, banking and financial services, publishing, pharmaceutical, manufacturing, and so on. Data of all forms is growing exponentially--quantitative, qualitative, structured, unstructured, speech, video, and so on. It is imperative to make use of this data to leverage all functions--avoid risk and fraud, enhance customer experience, increase revenues, and streamline operations.

Organizations are moving fast to embrace data science and investing a lot into high-end data science teams. Having spent more than 12 years in the BFSI domain, I get overwhelmed with the transition that the BFSI industry has seen in embracing analytics as a business and no longer a support function. This holds especially true for the fin-tech and digital lending world of which Jalaj and myself are a part of.

I have known Jalaj since her college days and am impressed with her exuberance and self-motivation. Her research skills, perseverance, commitment, discipline, and quickness to grasp even the most difficult concepts have made her achieve success in a short span of 4 years on her corporate journey.

Jalaj is a gifted intellectual with a strong mathematical and statistical understanding and demonstrates a continuous passion for learning the new and complex analytical and statistical techniques that are emerging in the industry. She brings experience to the data science domain and I have seen her deliver impressive projects around NLP, machine learning, basic linguistic analysis, neural networks, and deep learning. The blistering pace of the work schedule that she sets for herself, coupled with the passion she puts into her work, leads to definite and measurable results for her organization.

One of her most special qualities is her readiness to solve the most basic to the most complex problem in the interest of the business. She is an excellent team player and ensures that the organization gains the maximum benefit of her exceptional talent.

In this book, Jalaj takes us on an exciting and insightful journey through the natural language processing domain. She starts with the basic concepts and moves on to the most advanced concepts, such as how machine learning and deep learning are used in NLP.

I wish Jalaj all the best in all her future endeavors.

Sarita Arora  
Chief Analytics Officer, SMECorner  
Mumbai, India



# About the Author

**Jalaj Thanaki** is a data scientist by profession and data science researcher by practice. She likes to deal with data science related problems. She wants to make the world a better place using data science and artificial intelligence related technologies. Her research interest lies in natural language processing, machine learning, deep learning, and big data analytics. Besides being a data scientist, Jalaj is also a social activist, traveler, and nature-lover.

# Acknowledgement

I would like to dedicate this book to my husband, Shetul Thanaki, for his constant support, encouragement, and creative suggestions.

I give deep thanks and gratitude to my parents, my in-laws, my family, and my friends, who have helped me at every stage of my life. I would also like to thank all the mentors that I've had over the years. I really appreciate the efforts by technical reviewers for reviewing this book. I would also like to thank my current organization, SMECorner, for its support. I am a big fan of open source communities and education communities, so I really want to thank communities such as Kaggel, Udacity, and Coursera who have helped me, in a direct or indirect manner, to understand the various concepts of data science. Without learning from these communities, there is not a chance I could be doing what I do today.

I would like to thank Packt Publishing and Aman Singh, who approached me to write this book. I really appreciate the effort put in by the entire Packt editorial team to make this book as good as possible. Special thanks to Aman Singh, Jagruti Babaria, Menka Bohra, Manthan Patel, Nidhi Joshi, Sayli Nikalje, Manisha Sinha, Safis, and Tania Dutta.

I would like to recognize the efforts of technical editing team, strategy and management team, marketing team, sales team, graphics designer team, pre-production team, post production team, layout coordinators team, and indexer team for making my authoring journey so smooth.

I feel really compelled to pass my knowledge on to those willing to learn.

Thank you God for being kind to me!

Cheers and Happy Reading!

# About the Reviewers

**Devesh Raj** is a data scientist with 10 years of experience in developing algorithms and solving problems in various domains--healthcare, manufacturing, automotive, production, and so on, applying machine learning (supervised and unsupervised machine learning techniques) and deep learning on structured and unstructured data (computer vision and NLP).

**Gayetri Thakur** is a linguist working in the area of natural language processing. She has worked on co-developing NLP tools such as automatic grammar checker, named entity recognizer, and text-to-speech and speech-to-text systems. She currently works for Google India Pvt.Ltd. India.

She is pursuing a PhD in linguistics and has completed her masters in linguistics from Banaras Hindu University.

**Prabhanjan Tattar** has over 9 years of experience as a statistical analyst. Survival analysis and statistical inference are his main areas of research/interest, and he has published several research papers in peer-reviewed journals and authored three books on R: *R Statistical Application Development by Example*, Packt Publishing, *A Course in Statistics with R*, Wiley, and *Practical Data Science Cookbook*, Packt Publishing. He also maintains the R packages `gpk`, `RSADBE`, and `ACSWR`.

**Chirag Mahapatra** is a software engineer who works on applying machine learning and natural language processing to problems in trust and safety. He currently works at Trooly (acquired by Airbnb). In the past, he has worked at `A9.com` on the ads data platform.

# www.PacktPub.com

For support files and downloads related to your book, please visit [www.PacktPub.com](http://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](http://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](mailto:service@packtpub.com) for more details.

At [www.PacktPub.com](http://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/1787121429>.

If you'd like to join our team of regular reviewers, you can e-mail us at [customerreviews@packtpub.com](mailto: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

<b>Preface</b>	1
<b>Chapter 1: Introduction</b>	9
<b>Understanding natural language processing</b>	9
<b>Understanding basic applications</b>	15
Understanding advanced applications	16
<b>Advantages of togetherness - NLP and Python</b>	17
<b>Environment setup for NLTK</b>	18
<b>Tips for readers</b>	20
<b>Summary</b>	20
<b>Chapter 2: Practical Understanding of a Corpus and Dataset</b>	21
<b>What is a corpus?</b>	21
<b>Why do we need a corpus?</b>	23
<b>Understanding corpus analysis</b>	26
Exercise	29
<b>Understanding types of data attributes</b>	29
Categorical or qualitative data attributes	30
Numeric or quantitative data attributes	31
<b>Exploring different file formats for corpora</b>	32
<b>Resources for accessing free corpora</b>	34
<b>Preparing a dataset for NLP applications</b>	35
Selecting data	35
Preprocessing the dataset	36
Formatting	36
Cleaning	36
Sampling	37
Transforming data	37
<b>Web scraping</b>	37
<b>Summary</b>	41
<b>Chapter 3: Understanding the Structure of a Sentences</b>	43
<b>Understanding components of NLP</b>	43
<b>Natural language understanding</b>	44
Natural language generation	44
Differences between NLU and NLG	45
Branches of NLP	45

<b>Defining context-free grammar</b>	46
Exercise	49
<b>Morphological analysis</b>	49
What is morphology?	49
What are morphemes?	49
What is a stem?	49
What is morphological analysis?	49
What is a word?	51
Classification of morphemes	52
Free morphemes	52
Bound morphemes	52
Derivational morphemes	53
Inflectional morphemes	53
What is the difference between a stem and a root?	57
Exercise	57
Lexical analysis	57
What is a token?	57
What are part of speech tags?	58
Process of deriving tokens	58
Difference between stemming and lemmatization	60
Applications	60
<b>Syntactic analysis</b>	60
What is syntactic analysis?	60
<b>Semantic analysis</b>	63
What is semantic analysis?	63
Lexical semantics	63
Hyponymy and hyponyms	64
Homonymy	64
Polysemy	64
What is the difference between polysemy and homonymy?	65
Application of semantic analysis	65
<b>Handling ambiguity</b>	65
Lexical ambiguity	66
Syntactic ambiguity	67
Approach to handle syntactic ambiguity	67
Semantic ambiguity	68
Pragmatic ambiguity	68
<b>Discourse integration</b>	68
Applications	69
<b>Pragmatic analysis</b>	69
<b>Summary</b>	69

<b>Chapter 4: Preprocessing</b>	71
<b>Handling corpus-raw text</b>	71
Getting raw text	72
Lowercase conversion	73
Sentence tokenization	74
Challenges of sentence tokenization	76
Stemming for raw text	77
Challenges of stemming for raw text	78
Lemmatization of raw text	78
Challenges of lemmatization of raw text	81
Stop word removal	81
Exercise	83
<b>Handling corpus-raw sentences</b>	84
Word tokenization	84
Challenges for word tokenization	85
Word lemmatization	85
Challenges for word lemmatization	86
<b>Basic preprocessing</b>	86
Regular expressions	87
Basic level regular expression	87
Basic flags	87
Advanced level regular expression	92
Positive lookahead	92
Positive lookbehind	93
Negative lookahead	93
Negative lookbehind	93
<b>Practical and customized preprocessing</b>	95
Decide by yourself	95
Is preprocessing required?	96
What kind of preprocessing is required?	97
Understanding case studies of preprocessing	97
Grammar correction system	97
Sentiment analysis	98
Machine translation	98
Spelling correction	98
Approach	99
<b>Summary</b>	103
<b>Chapter 5: Feature Engineering and NLP Algorithms</b>	105
<b>Understanding feature engineering</b>	107
What is feature engineering?	107
What is the purpose of feature engineering?	108
Challenges	108



<b>Basic feature of NLP</b>	109
<b>Parsers and parsing</b>	109
Understanding the basics of parsers	109
Understanding the concept of parsing	112
Developing a parser from scratch	113
Types of grammar	113
Context-free grammar	114
Probabilistic context-free grammar	117
Calculating the probability of a tree	118
Calculating the probability of a string	120
Grammar transformation	121
Developing a parser with the Cocke-Kasami-Younger Algorithm	123
Developing parsers step-by-step	127
Existing parser tools	128
The Stanford parser	128
The spaCy parser	131
Extracting and understanding the features	132
Customizing parser tools	134
Challenges	134
<b>POS tagging and POS taggers</b>	135
Understanding the concept of POS tagging and POS taggers	135
Developing POS taggers step-by-step	136
Plug and play with existing POS taggers	139
A Stanford POS tagger example	139
Using polyglot to generate POS tagging	140
<b>Exercise</b>	140
Using POS tags as features	141
Challenges	141
<b>Name entity recognition</b>	141
Classes of NER	142
Plug and play with existing NER tools	143
A Stanford NER example	143
A Spacy NER example	144
Extracting and understanding the features	144
Challenges	145
<b>n-grams</b>	145
Understanding n-gram using a practice example	147
Application	148
<b>Bag of words</b>	149
Understanding BOW	149
Understanding BOW using a practical example	150
Comparing n-grams and BOW	151
Applications	151
<b>Semantic tools and resources</b>	151
<b>Basic statistical features for NLP</b>	152
Basic mathematics	152