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, Standford 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

ISBN 978-1-78712-142-3
54999
9 781787 121423

Python Natural Language Processing

Jalaj Thanaki

Python Natural Language Processing

Explore NLP with machine learning and deep learning techniques

Jalaj Thanaki

BIRMINGHAM - MUMBAI

Python Natural Language Processing

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

Credits

Author
Jalaj Thanaki

Reviewers
Devesh Raj
Gayetri Thakur
Prabhanjan Tattar
Chirag Mahapatra

Commissioning Editor
Veena Pagare

Acquisition Editor
Aman Singh

Content Development Editor
Jagruti Babaria

Technical Editor
Sayli Nikalje

Copy Editor
Safis Editing

Project Coordinator
Manthan Patel

Proofreader
Safis Editing

Indexer
Tejal Daruwale Soni

Production Coordinator
Deepika Naik

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`.

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/1787121429`.

If you'd like to join our team of regular reviewers, you can e-mail 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