Jalaj Thanaki 著

Python自然语言处理

（影印版）

Python Natural Langu

东南大学出版社
SOUTHEAST UNIVERSITY PRESS

Packt

Python自然语言处理（影印版）

Python Natural Language Processing

本书首先阐述了自然语言处理（Natural Language Processing，NLP）的基础，以及为什么Python是构建基于NLP的专家系统的最佳选择之一，其中包括社区支持和可用框架等优势。它还能够使你更好地理解可用的免费语料库以及不同类型的数据集。随后，你会学到如何为NLP应用选择数据集，找到正确的NLP技术来处理数据集中的句子并理解其结构。另外还将学习如何标记句子的不同部分并查看其分析方法。

在阅读本书的过程中，你将探索文本的语义和句法分析。了解如何解决处理人类语言时出现的各种歧义，并且将碰到在执行文本分析时出现的各种情况。

你会学到设置NLP环境的基础知识，初始化设置，然后快速理解句子和语言。你将领会到利用机器学习和深度学习从文本数据中提取信息的威力。

在本书的结尾，你会对NLP有一个清晰的理解并在现实中实现多个NLP示例。

从本书中你将学到：

- 关注用于开发NLP应用程序的Python编程范式
- 理解语料库分析以及不同类型的数据属性
- 使用Python库（例如NLTK、Polyglot、SpaCy、Standford CoreNLP等）学习NLP
- 了解作为特征工程组成部分的特征提取和特征选择
- 探索矢量化在深度学习中的优势
- 更好地理解基于规则的系统体系结构
- 针对NLP问题优化和调校有监督和无监督的机器学习算法
- 识别自然语言处理和自然语言生成问题的深度学习技术

ISBN 978-7-5641-7865-9
9 787564 178659
定价：106.00 元

Packt
www.packtpub.com
责任编辑　张烨 / 责任印制　周荣虎

Python 自然语言处理(影印版)

Python Natural Language Processing

Jalaj Thanaki 著

南京　东南大学出版社

图书在版编目(CIP)数据

Python 自然语言处理:英文/(印)贾拉·萨拉基(Jalaj Thanaki)著. —影印本. —南京:东南大学出版社,2018.10

书名原文:Python Natural Language Processing

ISBN 978-7-5641-7865-9

Ⅰ. ①P… Ⅱ. ①贾… Ⅲ. ①软件工具—程序设计—英文 Ⅳ. ①TP311.561

中国版本图书馆 CIP 数据核字(2018)第 153451 号

图字:10-2018-104 号

Python 自然语言处理(影印版)

出版发行:东南大学出版社
地　　址:南京四牌楼 2 号　　邮编:210096
出 版 人:江建中
网　　址:http://www.seupress.com
电子邮件:press@seupress.com
印　　刷:常州市武进第三印刷有限公司
开　　本:787 毫米×980 毫米　　16 开本
印　　张:30.5
字　　数:597 千字
版　　次:2018 年 10 月第 1 版
印　　次:2018 年 10 月第 1 次印刷
书　　号:ISBN 978-7-5641-7865-9
定　　价:106.00 元

本社图书若有印装质量问题,请直接与营销部联系。电话(传真):025-83791830

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