

Jalaj Thanaki 著

Python自然语言处理

(影印版)

Python Natural Language

 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、Stanford 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

Packt

www.packtpub.com

责任编辑 张焯 / 责任印制 周荣虎

ISBN 978-7-5641-7865-9



9 787564 178659 >

定价: 106.00 元

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

© 2017 by PACKT Publishing Ltd.

Reprint of the English Edition, jointly published by PACKT Publishing Ltd and Southeast University Press, 2018.
Authorized reprint of the original English edition, 2018 PACKT Publishing Ltd, the owner of all rights to
publish and sell the same.

All rights reserved including the rights of reproduction in whole or in part in any form.

英文原版由 PACKT Publishing Ltd 出版 2017。

英文影印版由东南大学出版社出版 2018。此影印版的出版和销售得到出版权和销售权的所有者
——PACKT Publishing Ltd 的许可。

版权所有,未得书面许可,本书的任何部分和全部不得以任何形式重制。

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

Fo 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.

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 financial 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 experiential 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 of 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. Her hands-on 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.

Sanku Avasth
Chief Analytics Officer, PNB
Mumbai, India

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!



Chirag Mahapatra is a PhD in linguistics and has worked in the field of natural language processing for over 10 years. He is currently a senior research scientist at Microsoft Research, where he works on the development of natural language understanding systems. He has published several papers in the field of natural language processing and is also a frequent speaker at conferences in the area.

Get the most in-demand software skills with Packt. Packt 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.

Prashant Talwar has over 10 years of experience in the field of software development and is currently a senior software engineer at Microsoft. He has published several papers in the field of software development and is also a frequent speaker at conferences in the area.

Why subscribe?

• Full access to every book published by Packt

• Copy and paste, print and bookmark content

• On demand and accessible via a web browser

Chirag Mahapatra is a PhD in linguistics and has worked in the field of natural language processing for over 10 years. He is currently a senior research scientist at Microsoft Research, where he works on the development of natural language understanding systems. He has published several papers in the field of natural language processing and is also a frequent speaker at conferences in the area.

Table of Contents

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

Defining context-free grammar	46
Exercise	49
Morphological analysis	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
Syntactic analysis	60
What is syntactic analysis?	60
Semantic analysis	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
Handling ambiguity	65
Lexical ambiguity	66
Syntactic ambiguity	67
Approach to handle syntactic ambiguity	67
Semantic ambiguity	68
Pragmatic ambiguity	68
Discourse integration	68
Applications	69
Pragmatic analysis	69
Summary	69

Chapter 4: Preprocessing	71
Handling corpus-raw text	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
Handling corpus-raw sentences	84
Word tokenization	84
Challenges for word tokenization	85
Word lemmatization	85
Challenges for word lemmatization	86
Basic preprocessing	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
Practical and customized preprocessing	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
Summary	103
Chapter 5: Feature Engineering and NLP Algorithms	105
Understanding feature engineering	107
What is feature engineering?	107
What is the purpose of feature engineering?	108
Challenges	108

Basic feature of NLP	109
Parsers and parsing	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
POS tagging and POS taggers	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
Exercise	140
Using POS tags as features	141
Challenges	141
Name entity recognition	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
n-grams	145
Understanding n-gram using a practice example	147
Application	148
Bag of words	149
Understanding BOW	149
Understanding BOW using a practical example	150
Comparing n-grams and BOW	151
Applications	151
Semantic tools and resources	151
Basic statistical features for NLP	152
Basic mathematics	152

Basic concepts of linear algebra for NLP	153
Basic concepts of the probabilistic theory for NLP	154
Probability	154
Independent event and dependent event	155
Conditional probability	158
TF-IDF	159
Understanding TF-IDF	159
Understanding TF-IDF with a practical example	162
Using textblob	162
Using scikit-learn	163
Application	164
Vectorization	164
Encoders and decoders	165
One-hot encoding	165
Understanding a practical example for one-hot encoding	165
Application	167
Normalization	167
The linguistics aspect of normalization	167
The statistical aspect of normalization	168
Probabilistic models	168
Understanding probabilistic language modeling	169
Application of LM	171
Indexing	171
Application	172
Ranking	172
Advantages of features engineering	173
Challenges of features engineering	174
Summary	174
Chapter 6: Advanced Feature Engineering and NLP Algorithms	177
Recall word embedding	177
Understanding the basics of word2vec	178
Distributional semantics	178
Defining word2vec	180
Necessity of unsupervised distribution semantic model - word2vec	181
Challenges	181
Converting the word2vec model from black box to white box	183
Distributional similarity based representation	184
Understanding the components of the word2vec model	185
Input of the word2vec	185
Output of word2vec	186
Construction components of the word2vec model	187
Architectural component	188

Understanding the logic of the word2vec model	189
Vocabulary builder	190
Context builder	191
Neural network with two layers	193
Structural details of a word2vec neural network	194
Word2vec neural network layer's details	194
Softmax function	197
Main processing algorithms	199
Continuous bag of words	199
Skip-gram	200
Understanding algorithmic techniques and the mathematics behind the word2vec model	202
Understanding the basic mathematics for the word2vec algorithm	202
Techniques used at the vocabulary building stage	204
Lossy counting	204
Using it at the stage of vocabulary building	204
Applications	204
Techniques used at the context building stage	204
Dynamic window scaling	205
Understanding dynamic context window techniques	205
Subsampling	205
Pruning	206
Algorithms used by neural networks	206
Structure of the neurons	207
Basic neuron structure	207
Training a simple neuron	209
Define error function	210
Understanding gradient descent in word2vec	211
Single neuron application	212
Multi-layer neural networks	213
Backpropagation	215
Mathematics behind the word2vec model	217
Techniques used to generate final vectors and probability prediction stage	220
Hierarchical softmax	220
Negative sampling	221
Some of the facts related to word2vec	221
Applications of word2vec	222
Implementation of simple examples	223
Famous example (king - man + woman)	223
Advantages of word2vec	224
Challenges of word2vec	225
How is word2vec used in real-life applications?	226