Natural Language Processing : A Textbook with Python Implementation

Book · November 2023

DOI: 10.1007/978-981-99-1999-4

1 author:

Raymond Lee
United International College
119 PUBLICATIONS **1,006** CITATIONS

SEE PROFILE

Raymond S. T. Lee

Natural Language Processing

A Textbook with Python Implementations

Springer

Natural Language Processing

Raymond S. T. Lee

Natural Language Processing

A Textbook with Python Implementation

Springer

Raymond S. T. Lee
United International College
Beijing Normal University-Hong Kong Baptist University
Zhuhai, China

ISBN 978-981-99-1998-7 ISBN 978-981-99-1999-4 (eBook)
https://doi.org/10.1007/978-981-99-1999-4

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

This book is dedicated to all readers and students taking my undergraduate and postgraduate courses in Natural Language Processing, your enthusiasm in seeking knowledge incited me to write this book.

Preface

Motivation of This Book

Natural Language Processing (NLP) and its related applications become part of daily life with exponential growth of Artificial Intelligence (AI) in past decades. NLP applications including Information Retrieval (IR) systems, Text Summarization System, and Question-and-Answering (Chatbot) System became one of the prevalent topics in both industry and academia that had evolved routines and benefited immensely to a wide array of day-to-day services.

The objective of this book is to provide NLP concepts and knowledge to readers with a 14-h 7 step-by-step workshops to practice various core Python-based NLP tools: NLTK, spaCy, TensorFlow Keras, Transformer, and BERT Technology to construct NLP applications.

Organization and Structure of This Book

This book consists of two parts:

Part I Concepts and Technology (Chaps. 1–9)

Discuss concepts and technology related to NLP including: Introduction, N-gram Language Model, Part-of-Speech Tagging, Syntax and Parsing, Meaning Representation, Semantic Analysis, Pragmatic Analysis, Transfer Learning and Transformer Technology, Major NLP Applications.

Part II Natural Language Processing Workshops with Python Implementation (Chaps. 10–16)

7 Python workshops to provide step-by-step Python implementation tools including: NLTK, spaCy, TensorFlow Keras, Transformer, and BERT Technology.

This book is organized and structured as follows:

Part I: Concepts and Technology

- Chapter 1: Introduction to Natural Language Processing

This introductory chapter begins with human language and intelligence constituting six levels of linguistics followed by a brief history of NLP with major components and applications. It serves as the cornerstone to the NLP concepts and technology discussed in the following chapters. This chapter also serves as the conceptual basis for Workshop#1: Basics of Natural Language Toolkit (NLTK) in Chap. 10.

- Chapter 2: N-gram Language Model

  Language model is the foundation of NLP. This chapter introduces N-gram language model and Markov Chains using classical literature *The Adventures of Sherlock Holmes* by Sir Conan Doyle (1859–1930) to illustrate how N-gram model works that form NLP basics in text analysis followed by Shannon's model and text generation with evaluation schemes. This chapter also serves as the conceptual basis for Workshop#2 on N-gram modelling with NLTK in Chap. 11.
- Chapter 3: Part-of-Speech Tagging

  Part-of-Speech (POS) Tagging is the foundation of text processing in NLP. This chapter describes how it relates to NLP and Natural Language Understanding (NLU). There are types and algorithms for POS Tagging including Rule-based POS Tagging, Stochastic POS Tagging, and Hybrid POS Tagging with Brill Tagger and evaluation schemes. This chapter also serves as the conceptual basis for Workshop#3: Part-of-Speech using Natural Language Toolkit in Chap. 12.
- Chapter 4—Syntax and Parsing

  As another major component of Natural Language Understanding (NLU), this chapter explores syntax analysis and introduces different types of constituents in English language followed by the main concept of context-free grammar (CFG) and CFG parsing. It also studies different major parsing techniques, including lexical and probabilistic parsing with live examples for illustration.
- Chapter 5: Meaning Representation

  Before the study of Semantic Analysis, this chapter explores meaning representation, a vital component in NLP. It studies four major meaning representation techniques which include: first-order predicate calculus (FOPC), semantic net, conceptual dependency diagram (CDD), and frame-based representation. After that it explores canonical form and introduces Fillmore's theory of universal cases followed by predicate logic and inference work using FOPC with live examples.
- Chapter 6: Semantic Analysis

  This chapter studies Semantic Analysis, one of the core concepts for learning NLP. First, it studies the two basic schemes of semantic analysis: lexical and compositional semantic analysis. After that it explores word senses and six commonly used lexical semantics followed by word sense disambiguation (WSD) and various WSD schemes. Further, it also studies WordNet and online thesauri for word similarity and various distributed similarity measurement including Point-wise Mutual Information (PMI) and Positive Point-wise Mutual information (PPMI) models with live examples for illustration. Chapters 4 and 5 also

serve as the conceptual basis for Workshop#4: Semantic Analysis and Word Vectors using spaCy in Chap. 13.

- Chapter 7: Pragmatic Analysis

  After the discussion of semantic meaning and analysis, this chapter explores pragmatic analysis in linguistics and discourse phenomena. It also studies coherence and coreference as the key components of pragmatics and discourse critical to NLP, followed by discourse segmentation with different algorithms on Coreference Resolution including Hobbs Algorithm, Centering Algorithm, Log-Linear Model, the latest machine learning methods, and evaluation schemes. This chapter also serves as the conceptual basis for Workshop#5: Sentiment Analysis and Text Classification in Chap. 14.
- Chapter 8: Transfer Learning and Transformer Technology

  Transfer learning is a commonly used deep learning model to minimize computational resources. This chapter explores: (1) Transfer Learning (TL) against traditional Machine Learning (ML); (2) Recurrent Neural Networks (RNN), a significant component of transfer learning with core technologies such as Long Short-Term Memory (LSTM) Network and Bidirectional Recurrent Neural Networks (BRNNs) in NLP applications, and (3) Transformer technology architecture, Bidirectional Encoder Representation from Transformers (BERT) Model, and related technologies including Transformer-XL and ALBERT technologies. This chapter also serves as the conceptual basis for Workshop#6: Transformers with spaCy and Tensorflow in Chap. 15.
- Chapter 9: Major Natural Language Processing Applications

  This is a summary of Part I with three core NLP applications: Information Retrieval (IR) systems, Text Summarization (TS) systems, and Question-and-Answering (Q&A) chatbot systems, how they work and related R&D in building NLP applications. This chapter also serves as the conceptual basis for Workshop#7: Building Chatbot with TensorFlow and Transformer Technology in Chap. 16.

Part II: Natural Language Processing Workshops with Python Implementation in 14 h

- Chapter 10: Workshop#1 Basics of Natural Language Toolkit (Hour 1–2)

  With the basic NLP concept being learnt in Chap. 1, this introductory workshop gives a NLTK overview and system installation procedures are the foundations of Python NLP development tool used for text processing which include simple text analysis, text analysis with lexical dispersion plot, text tokenization, and basic statistical tools in NLP.
- Chapter 11: Workshop#2 N-grams Modelling with Natural Language Toolkit (Hour 3–4)

  This is a coherent workshop of Chap. 2 using NTLK technology for N-gram generation and statistics. This workshop consists of two parts. Part I introduces N-gram language model using NLTK in Python and N-grams class to generate N-gram statistics on any sentence, text objects, whole document, literature to

provide a foundation technique for text analysis, parsing and semantic analysis in subsequent workshops. Part II introduces spaCy, the second important NLP Python implementation tools not only for teaching and learning (like NLTK) but also widely used for NLP applications including text summarization, information extraction, and Q&A chatbot. It is a critical mass to integrate with Transformer Technology in subsequent workshops.

- Chapter 12: Workshop#3 Part-of-Speech Tagging with Natural Language Toolkit (Hour 5–6)

  In Chap. 3, we studied basic concepts and theories related to Part-of-Speech (POS) and various POS tagging techniques. This workshop explores how to implement POS tagging by using NLTK starting from a simple recap on tokenization techniques and two fundamental processes in word-level progressing: stemming and stop-word removal, which will introduce two types of stemming techniques: Porter Stemmer and Snowball Stemmer that can be integrated with WordCloud commonly used in data visualization followed by the main theme of this workshop with the introduction of PENN Treebank Tagset and to create your own POS tagger.
- Chapter 13: Workshop#4 Semantic Analysis and Word Vectors using spaCy (Hour 7–8)

  In Chaps. 5 and 6, we studied the basic concepts and theories related to meaning representation and semantic analysis. This workshop explores how to use spaCy technology to perform semantic analysis starting from a revisit on word vectors concept, implement and pre-train them followed by the study of similarity method and other advanced semantic analysis.
- Chapter 14: Workshop#5 Sentiment Analysis and Text Classification (Hour 9–10)

  This is a coherent workshop of Chap. 7, this workshop explores how to position NLP implementation techniques into two important NLP applications: text classification and sentiment analysis. TensorFlow and Kera are two vital components to implement Long Short-Term Memory networks (LSTM networks), a commonly used Recurrent Neural Networks (RNN) on machine learning especially in NLP applications.
- Chapter 15: Workshop#6 Transformers with spaCy and TensorFlow (Hour 11–12)

  In Chap. 8, the basic concept about Transfer Learning, its motivation and related background knowledge such as Recurrent Neural Networks (RNN) with Transformer Technology and BERT model are introduced. This workshop explores how to put these concepts and theories into practice. More importantly, is to implement Transformers, BERT Technology with the integration of spaCy's Transformer Pipeline Technology and TensorFlow. First, it gives an overview and summation on Transformer and BERT Technology. Second, it explores Transformer implementation with TensorFlow by revisiting Text Classification using BERT model as example. Third, it introduces spaCy's Transformer Pipeline Technology and how to implement Sentiment Analysis and Text Classification system using Transformer Technology.

- Chapter 16: Workshop#7 Building Chatbot with TensorFlow and Transformer Technology (Hour 13–14)

  In previous six NLP workshops, we studied NLP implementation tools and techniques ranging from tokenization, N-gram generation to semantic and sentiment analysis with various key NLP Python enabling technologies: NLTK, spaCy, TensorFlow and contemporary Transformer Technology. This final workshop explores how to integrate them for the design and implementation of a live domain-based chatbot system on a movie domain. First, it explores the basis of chatbot system and introduce a knowledge domain—the Cornell Large Movie Conversation Dataset. Second, it conducts a step-by-step implementation of movie chatbot system which involves dialog preprocessing, model construction, attention learning implementation, system integration, and performance evaluation followed by live tests. Finally, it introduces a mini project for this workshop and present related chatbot datasets with resources in summary.

Readers of This Book

This book is both an NLP textbook and NLP Python implementation book tailored for:

- Undergraduates and postgraduates of various disciplines including AI, Computer Science, IT, Data Science, etc.
- Lecturers and tutors teaching NLP or related AI courses.
- NLP, AI scientists and developers who would like to learn NLP basic concepts, practice and implement via Python workshops.
- Readers who would like to learn NLP concepts, practice Python-based NLP workshops using various NLP implementation tools such as NLTK, spaCy, TensorFlow Keras, BERT, and Transformer technology.

How to Use This book?

This book can be served as a textbook for undergraduates and postgraduate courses on Natural Language Processing, and a reference book for general readers who would like to learn key technologies and implement NLP applications with contemporary implementation tools such as NLTK, spaCy, TensorFlow, BERT, and Transformer technology.

Part I (Chaps. 1–9) covers the main course materials of basic concepts and key technologies which include N-gram Language Model, Part-of-Speech Tagging, Syntax and Parsing, Meaning Representation, Semantic Analysis, Pragmatic

Analysis, Transfer Learning and Transformer Technology, and Major NLP Applications. Part II (Chaps. 10–16) provides materials for a 14-h, step-by-step Python-based NLP implementation in 7 workshops.

For readers and AI scientists, this book can be served as both reference in learning NLP and Python implementation toolbook on NLP applications by using the latest Python-based NLP development tools, platforms, and libraries.

For seven NLP Workshops in Part II (Chaps. 10–16), readers can download all JupyterNB files and data files from my NLP GitHub directory: https://github.com/raymondshtlee/nlp/. For any query, please feel free to contact me via email: raymondshtlee@uic.edu.cn.

Zhuhai, China Raymond S. T. Lee

Acknowledgements

I would like to express my gratitude:

To my wife Iris for her patience, encouragement, and understanding, especially during my time spent on research and writing in the past 30 years.

To Ms. Celine Cheng, executive editor of Springer NATURE and her professional editorial and book production team for their support, valuable comments, and advice.

To Prof. Tang Tao, President of UIC, for the provision of excellent environment for research, teaching, and writing this book.

To Prof. Weijia Jia, Vice President (Research and Development) of UIC for their supports for R&D of NLP and related AI projects.

To Prof. Jianxin Pan, Dean of Faculty of Science and Technology of UIC, and Prof. Weifeng Su, Head of Department of Computer Science of UIC for their continuous supports for AI and NLP courses.

To research assistant Mr. Zihao Huang for the help of NLP workshops preparation. To research student Ms. Clarissa Shi and student helpers Ms. Siqi Liu, Mr. Mingjie Wang, and Ms. Jie Lie to help with literature review on major NLP applications and Transformer technology, and Mr. Zhuohui Chen to help for bugs fixing and version update for the workshop programs.

To UIC for the prominent support in part by the Guangdong Provincial Key Laboratory IRADS (2022B1212010006, R0400001-22), Key Laboratory for Artificial Intelligence and Multi-Model Data Processing of Department of Education of Guangdong Province and Guangdong Province F1 project grant on Curriculum Development and Teaching Enhancement on course development UICR0400050-21 CTL for the provision of an excellent environment and computer facilities for the preparation of this book.

Dr. Raymond Lee

December 2022

Beijing Normal University-Hong Kong Baptist University United International College

Zhuhai

China

About the Book

This textbook presents an up-to-date and comprehensive overview of Natural Language Processing (NLP) from basic concepts to core algorithms and key applications. It contains 7 step-by-step workshops (total 14 h) to practice essential Python tools like NLTK, spaCy, TensorFlow Kera, Transformer, and BERT.

The objective of this book is to provide readers with fundamental knowledge, core technologies, and enable to build their own applications (e.g. Chatbot systems) using Python-based NLP tools. It is both a textbook and toolbook intended for undergraduate students from various disciplines who want to learn, lecturers and tutors who want to teach courses or tutorials for undergraduate/graduate students on the subject and related AI topics, and readers with various backgrounds who want to learn and build practicable applications after completing 14 h Python-based workshops.

Contents

About the Author

Raymond Lee is the founder of the Quantum Finance Forecast System (QFFC) (https://qffc.uic.edu.cn) and currently an Associate Professor at United International College (UIC) with 25+ years' experience in AI research and consultancy, Chaotic Neural Networks, NLP, Intelligent Fintech Systems, Quantum Finance, and Intelligent E-Commerce Systems. He has published over 100 publications and authored 8 textbooks in the fields of AI, chaotic neural networks, AI-based fintech systems, intelligent agent technology, chaotic cryptosystems, ontological agents, neural oscillators, biometrics, and weather simulation and forecasting systems. Upon completion of the QFFC project, in 2018 he joined United International College (UIC), China, to pursue further R&D work on AI-Fintech and to share his expertise in AI-Fintech, chaotic neural networks, and related intelligent systems with fellow students and the community. His three latest textbooks, Quantum Finance: Intelligent Forecast and Trading Systems (2019), Artificial Intelligence in Daily Life (2020), and this NLP book have been adopted as the main textbooks for various AI courses in UIC.

Abbreviations

AI	Artificial intelligence
ASR	Automatic speech recognition
BERT	Bidirectional encoder representations from transformers
BRNN	Bidirectional recurrent neural networks
CDD	Conceptual dependency diagram
CFG	Context-free grammar
CFL	Context-free language
CNN	Convolutional neural networks
CR	Coreference resolution
DNN	Deep neural networks
DT	Determiner
FOPC	First-order predicate calculus
GRU	Gate recurrent unit
HMM	Hidden Markov model
IE	Information extraction
IR	Information retrieval
KAI	Knowledge acquisition and inferencing
LSTM	Long short-term memory
MEMM	Maximum entropy Markov model
MeSH	Medical subject thesaurus
ML	Machine learning
NER	Named entity recognition
NLP	Natural language processing
NLTK	Natural language toolkit
NLU	Natural language understanding
NN	Noun
NNP	Proper noun
Nom	Nominal
NP	Noun phrase
PCFG	Probabilistic context-free grammar
PMI	Pointwise mutual information

POS	Part-of-speech
POST	Part-of-speech tagging
PPMI	Positive pointwise mutual information
Q&A	Question-and-answering
RNN	Recurrent neural networks
TBL	Transformation-based learning
VB	Verb
VP	Verb phrase
WSD	Word sense disambiguation