



Mastering Natural Language Processing using Python

From Fundamentals to Advanced Techniques

Dr. Goonjan Jain

Dr. Kanika Garg

SULTAN CHAND & SONS

Mastering Natural Language Processing using Python

From Fundamentals to Advanced Techniques

Dr. Goonjan Jain

Assistant Professor

Mathematics and Computing

Delhi Technological University (DTU)

Delhi

Dr. Kanika Garg

Associate Professor

Computer Science Department

Geetanjali Institute of Technical Studies

Dabok, Udaipur-313022



SULTAN CHAND & SONS[®]

Educational Publishers

New Delhi

SULTAN CHAND & SONS®

Educational Publishers

23, Daryaganj, New Delhi-110 002

Phones : 011-23281876, 23266105, 41625022 (*Showroom & Shop*)

011-23247051, 40234454 (*Office*)

E-Mail : sultanchand74@yahoo.com; info@sultanchandandsons.com

Fax : 011-23266357; Website : www.sultanchandandsons.com

ISBN : 978-93-49290-62-4 (TC-1319)

Price : ₹ 295.00

First Edition: 2025

EVERY GENUINE COPY OF THIS BOOK HAS A HOLOGRAM



In our endeavour to protect you against counterfeit/fake books, we have pasted a copper hologram over the cover of this book. The hologram displays the full visual image, unique 3D multi-level, multi-colour effects of our logo from different angles when tilted or properly illuminated under a single light source, such as 3D depth effect, kinetic effect, pearl effect, gradient effect, trailing effect, emboss effect, glitter effect, randomly sparking tiny dots, micro text, laser numbering, etc.

A fake hologram does not display all these effects.

Always ask the bookseller to put his stamp on the first page of this book.

All Rights Reserved: No part of this book, including its style and presentation, can be reproduced, stored in a retrieval system, or transmitted in any form or by any means – electronic, mechanical, photocopying, recording or otherwise without the prior written consent of the publishers. Exclusive publication, promotion and distribution rights reserved with the Publishers.

Warning: An unauthorised act done in relation to a copyright work may result in both civil claim for damages and criminal prosecution.

Special Note: Photocopy or Xeroxing of educational books without the written permission of publishers is illegal and against Copyright Act. Buying and Selling of pirated books is a criminal offence. Publication of a key to this book is strictly prohibited.

General: While every effort has been made to present authentic information and avoid errors, the author and the publishers are not responsible for the consequences of any action taken on the basis of this book.

Limits of Liability/Disclaimer of Warranty: The publisher and the author make no representation or warranties with respect to the accuracy or completeness of the contents of this work and specifically disclaim all warranties, including without limitation warranties of fitness for a particular purpose. No warranty may be created or extended by sales or promotional materials. The advice and strategies contained therein may not be suitable for every situation. This work is sold with the understanding that the publisher is not engaged in rendering legal, accounting, or other professional services. If professional assistance is required, the services of a competent professional person should be sought. Neither the publishers nor the author shall be liable for damages arising herefrom.

Disclaimer: The publishers have taken all care to ensure highest standard of quality as regards typesetting, proof-reading, accuracy of textual material, printing and binding. However, they accept no responsibility for any loss occasioned as a result of any misprint or mistake found in this publication.

Author's Acknowledgement: The writing of a Textbook always involves creation of a huge debt towards innumerable authors and publications. We owe our gratitude to all of them. We acknowledge our indebtedness in extensive footnotes throughout the book. If, for any reason, any acknowledgement has been left out we beg to be excused. We assure to carry out correction in the subsequent edition, as and when it is known.

Preface

“When the student is ready, the teacher will appear. When the student is truly ready, the teacher will disappear.”
— *Lao Tzu*

This book offers a comprehensive introduction to Natural Language Processing (NLP) and its diverse range of applications, from machine translation to sentiment analysis. This book provides both theoretical foundations and practical insights into the methods and technologies behind NLP systems. It aims to give readers a well-rounded understanding of how machines process natural language, the associated challenges, and the ways NLP techniques can solve real-world problems.

NLP has emerged as one of the most transformative subfields within artificial intelligence, bridging the gap between human communication and machine understanding. As NLP continues to advance, it is making profound impacts across various sectors, including healthcare, business, and beyond. The potential of this field seems boundless, particularly as it evolves alongside developments in machine learning and large language models.

Human language is the most fundamental means of communication, and enabling machines to understand and process it is a research-worthy endeavor. From healthcare to business, NLP allows for automation, enhances human-computer interaction, and facilitates the extraction of valuable insight. Research in NLP is pivotal for building the next generation of intelligent systems that can truly understand human language in all its complexity.

The scope of this book encompasses the fundamental concepts, methodologies, and advanced techniques in the field of NLP. It covers a wide range of topics, from the basics of language processing, such as tokenization and morphological analysis, to more advanced areas like machine translation, information retrieval, and deep learning-based NLP models. The book explores classical methods (*e.g.*, rule-based systems, probabilistic models) and modern approaches (*e.g.*, neural networks, transformer-based models like BERT and GPT). With hands-on implementations and practical examples throughout, it serves as a comprehensive guide for readers.

This book is intended for students, researchers, and industry professionals seeking a deeper understanding of Natural Language Processing. Whether you are new to the field or looking to enhance your expertise, this book offers a structured approach to learning both the theory and practical techniques in NLP. It is ideal for university courses on NLP, AI, and machine learning, as well as professionals and enthusiasts looking to explore the applications of NLP in real-world scenarios.

By the end of the book, readers will have a solid foundation in NLP and be equipped to tackle a variety of tasks, from basic text preprocessing to advanced applications like machine translation, question answering, and sentiment analysis.

We hope this book not only deepens your understanding of NLP but also inspires you to explore its many possibilities.

Goonjan Jain

Kanika Garg

Acknowledgement

For my daughters – Arshi & Ameyaa

– *Goonjan Jain*

I would like to express my heartfelt gratitude to my family and friends and my husband who have supported and encouraged me throughout the process.

– *Kanika Garg*

Snapshot of the Book

<i>S. No.</i>	<i>Chapter Name</i>	<i>Pages</i>	<i>Figures</i>	<i>Tables</i>	<i>Code Snippets</i>	<i>Exercise Questions</i>	<i>Practical Questions</i>
1.	Introduction to Language Processing	1-19	13	1	–	10	6
2.	Language Modeling	21-44	3	11	5	9	10
3.	Lexical Analysis	45-91	9	3	24	23	11
4.	Syntactic Analysis	93-129	9	10	2	21	6
5.	Semantic Analysis	131-162	10	5	7	13	7
6.	Discourse Processing	163-190	10	5	3	30	4
7.	Natural Language Generation	191-213	14	4	4	20	7
8.	Tasks in NLP	215-268	10	7	27	38	10
9.	Advanced Deep Learning and Large Language Models	269-309	26	8	4	30	10
	Previous Year Questions papers	311-316	–	–	–	–	–
	Bibliography	317-324	–	–	–	–	–
	Total	324	104	54	76	194	71

Contents

1. Introduction to Language Processing	1-19
1.1. Introduction	2
1.1.1. Why NLP?	2
1.1.2. Origins and History of NLP	3
1.2. Language Analysis	4
1.2.1. Phases in Language Analysis	4
1.3. Major Challenges of NLP	5
1.4. NLP and Machine Learning	7
1.5. Natural Language and Grammar	8
1.5.1. What is Grammar?	8
1.5.2. Context-Free Grammar	9
1.5.3. Dependency Grammar	9
1.5.4. Constituency Grammar	10
1.6. Applications of NLP	11
1.7. Tasks of NLP	12
1.7.1. Parsing	12
1.7.2. Sentiment Analysis	13
1.7.3. Machine Translation	13
1.7.4. Automatic Summarization	13
1.7.5. Document Classification	14
1.7.6. Question Answering	14
1.7.7. Named Entity Recognition	15
1.7.8. Word Sense Disambiguation	15
1.7.9. Keyword Extraction	16
1.8. Simple NLP Implementation using Python	17
1.8.1. Install Python	17
1.8.2. NLTK Toolkit	17
1.8.3. Other Common Python Libraries for NLP	18

1.9. Summary	18
Exercise	18
Practical Exercises	19
2. Language Modeling	21-44
2.1. Introduction	22
2.2. Statistical Language Models	23
2.2.1. N-gram Models	23
2.2.2. Smoothing Techniques	27
2.2.3. Maximum-Likelihood Estimation	30
2.2.4. Markov Models	31
2.2.5. Hidden Markov Models	31
2.2.6. Continuous Space Models	35
2.3. Grammar Based Language Models	41
2.4. Summary	43
Exercise	43
Practical Exercise	43
3. Lexical Analysis	45-91
3.1. Introduction	46
3.1.1. Key Terms	46
3.1.2. Other Useful Concepts	47
3.2. Initial Steps of Lexical Analysis	49
3.2.1. Tokenization	49
3.2.2. Stop Word Removal	49
3.2.3. Stemming	50
3.2.4. Lemmatization	51
3.3. Regular Expressions	52
3.4. Finite State Automata	53
3.5. Morphological Analysis	54
3.5.1. Morphological Analysis Techniques	54
3.5.2. Limitations of Morphological Parsing	59
3.6. Part-of-speech Tagging	60
3.6.1. POS Tagsets	61
3.6.2. Rule-based POS Tagging	61
3.6.3. Popular Rule-based POS Taggers	64
3.6.4. Machine Learning Based Tagging	70
3.6.5. Hidden Markov Model (HMM) based POS Tagging	70
3.6.6. Maximum Entropy (MaxEnt) Model	72
3.6.7. Rule-based Transformational Tagger	74
3.6.8. Hybrid Approaches	76
3.7. Spelling Error Detection and Correction	78
3.7.1. Error Detection Techniques	79
3.7.2. Error Correction Techniques	83

3.8. Summary	90
Exercise	90
Practical Exercise	91
4. Syntactic Analysis	93-129
4.1. Introduction	94
4.1.1. Syntax in Natural Language	94
4.1.2. Applications of Syntactic Parsing	96
4.2. Constituency	96
4.2.1. Phrase Level Constructions	97
4.2.2. Sentence Level Constructions	97
4.3. Context-free Grammar (CFG)	98
4.3.1. Definition of Context-free Grammar	98
4.3.2. Normal Forms of CFG	101
4.3.3. Benefits and Drawbacks of CFG	104
4.4. Probabilistic Context Free Grammar (PCFG)	105
4.4.1. Benefits and Drawbacks of PCFGs	107
4.4.2. Treebanks	108
4.4.3. Training PCFGs from Treebanks	109
4.5. Parsing	111
4.5.1. Top-down Parsing	113
4.5.2. Bottom-up Parsing	114
4.5.3. Bidirectional Parsing	115
4.6. Probabilistic Parsing	116
4.6.1. Introduction to Probabilistic Parsing	116
4.6.2. Inside-Outside Algorithm for PCFGs	117
4.6.3. Viterbi Algorithm for PCFGs	120
4.7. Structural Ambiguity and Resolution	123
4.8. The Cocke–Younger–Kasami (CYK) Algorithm	124
4.8.1. Overview of CYK Algorithm for Parsing CFG	124
4.8.2. Complexity Analysis and Limitations of CYK Algorithm	127
4.9. Summary	128
Exercise	128
Practical Exercise	129
5. Semantic Analysis	131-162
5.1. Introduction	132
5.1.1. Overview of Semantic Analysis	132
5.1.2. Importance and Applications	133
5.2. Meaning Representation	133
5.2.1. Approaches to Meaning Representation	133
5.2.2. Formal Representation of Meanings	136
5.2.3. Distributed Representations of Meaning	138
5.3. Lexical Semantics	140

5.3.1. Definition of Lexical Semantics	140
5.3.2. Word Sense Inventory	141
5.3.3. WordNet	141
5.3.4. Polysemy and Homonymy	142
5.3.5. Hyponymy and Hypernymy	145
5.3.6. Synonymy and Antonymy	146
5.3.7. Some Other Related Terms	147
5.4. Semantic Ambiguity	148
5.5. Semantic Relatedness	149
5.5.1. Similarity Measures	150
5.5.2. Measures of WordNet Similarity	152
5.5.3. Resnick's work on WordNet Similarity	156
5.5.4. Challenges and Limitations of Semantic Relatedness	157
5.5.5. Applications of Semantic Relatedness	159
5.6. Real Life Examples of Semantic Analysis	160
5.7. Summary	160
Exercise	161
Practical Exercise	162
6. Discourse Processing	163-190
6.1. Introduction	164
6.1.1. Overview of Discourse Processing	164
6.1.2. Importance and Applications	165
6.2. Cohesion	166
6.2.1. Cohesion vs. Coherence	167
6.2.2. Types of Cohesion	167
6.2.3. Techniques for Cohesion Analysis	168
6.2.4. Role of Cohesion in Discourse Understanding	170
6.3. Reference Resolution	171
6.3.1. Importance of Reference Resolution	172
6.3.2. Anaphora and Cataphora Resolution	172
6.3.3. Techniques for Reference Resolution	175
6.3.4. Challenges in Reference Resolution	178
6.4. Discourse Coherence and Structure	179
6.4.1. Discourse Coherence Models	181
6.4.2. Discourse Parsing	183
6.4.3. Applications of Discourse Coherence	184
6.4.4. Discourse Structure and Analysis Tools	185
6.5. Discourse Relation Recognition	186
6.5.1. Types of Discourse Relations	186
6.5.2. Discourse Relation Recognition Approaches	187
6.6. Summary	189
Exercise	189
Practical Exercise	190

7. Natural Language Generation	191-213
7.1. Introduction	192
7.1.1. Use Cases of NLG	192
7.1.2. Overview of Generative AI	194
7.2. Architectures of NLG Systems	196
7.2.1. Rule-Based NLG	197
7.2.2. Statistical NLG	199
7.2.3. Neural NLG	200
7.3. Generation Tasks and Representations	205
7.3.1. Text Generation	205
7.3.2. Image Captioning	207
7.3.3. Representations in NLG	209
7.4. Famous NLG Models	211
7.5. Summary	211
Exercise	212
Practical Exercise	212
8. Tasks in NLP	215-268
8.1. Word Sense Disambiguation	216
8.1.1. Introduction	216
8.1.2. Approaches to Word Sense Disambiguation	216
8.1.3. Evaluation and Benchmarking	222
8.1.4. Challenges and Limitations	224
8.1.5. Applications of WSD	225
8.2. Machine Translation	226
8.2.1. Introduction	227
8.2.2. Machine Translation Approaches	228
8.2.3. Evaluation Metrics	233
8.2.4. Machine Translation Involving Indian Languages	235
8.2.5. Challenges and Limitations in Machine Translation	236
8.3. Keyphrase Extraction	238
8.3.1. Introduction	238
8.3.2. Types of Keyword/Keyphrase Extraction Techniques	240
8.3.3. Evaluating different Keyword/Keyphrase Extraction Techniques	249
8.3.4. Challenges and Limitations	253
8.3.5. Applications	253
8.4. Sentiment Analysis	254
8.4.1. Introduction	254
8.4.2. Approaches to Sentiment Analysis	256
8.4.3. Emotion Analysis	260
8.4.4. Challenges and Limitations	263
8.4.5. Applications of Sentiment Analysis	264
8.5. Summary	266
Exercise	266

Practical Exercise	267
9. Advanced Deep Learning and Large Language Models	269-309
9.1. Introduction to Deep Learning Models	270
9.1.1. Importance of Deep Learning in NLP	270
9.1.2. Applications of Deep Learning in NLP	271
9.2. Recurrent Neural Network (RNN)	272
9.2.1. Introduction to RNN	272
9.2.2. Architecture of RNN	273
9.2.3. Applications of RNNs	277
9.2.4. Limitations of RNNs	277
9.3. Long Short-Term Memory (LSTM)	278
9.3.1. Introduction to LSTMs	279
9.3.2. LSTM Architecture and Components	279
9.3.3. Advantages of LSTMs over RNNs	281
9.3.4. Sentiment Analysis using LSTM	282
9.4. Large Language Models	284
9.4.1. Evolution of Large Language Models	284
9.4.2. Introduction to Large Language Models	285
9.4.3. Bidirectional Encoder Representations from Transformers (BERT)	288
9.4.4. Large Language Model Meta AI (LLAMA)	295
9.4.5. Generative Pre-trained Transformer (GPT)	299
9.4.6. Importance and Applications of Large Language Models	303
9.5. Opportunities and Challenges	305
9.6. Summary	307
Exercise	307
Practical Exercise	309
Previous Year Question Papers	311-316
Bibliography	317-323

About the Book

This book is a comprehensive guide to Natural Language Processing (NLP), designed for both beginners and advanced learners. Whether you're just starting or looking to refine your skills, this book takes you through every aspect of NLP—from the basics of text processing to cutting-edge machine learning techniques used in NLP today. It combines theoretical foundations with practical examples using Python, making complex NLP concepts accessible and actionable.

The book is rich with practical exercises, hands-on Python code snippets, and visual aids, ensuring that readers not only understand the concepts but also see how they apply in real-world scenarios. By the end of the book, readers will be proficient in using NLP libraries and will have a clear understanding of how to implement NLP solutions in Python.

Salient Features

- **Comprehensive Coverage:** From tokenization, stemming, and lemmatization to more advanced topics like named entity recognition (NER), sentiment analysis, and deep learning for NLP.
- **Python-Powered:** Every concept is paired with Python code examples, ensuring readers can practice and apply their knowledge immediately.
- **Beginner-Friendly to Advanced:** The book is structured to cater to all levels, beginning with foundational topics before advancing to complex NLP models.
- **Visual Aids:** Includes a variety of images, diagrams, and visual representations to simplify complex ideas.
- **Real-World Examples:** Contains numerous real-world NLP projects and use-cases to demonstrate practical applications.
- **Hands-On Exercises:** End-of-chapter exercises encourage readers to put their skills to the test and deepen their understanding.
- **Focus on Popular NLP Libraries:** Introduces popular Python NLP libraries like NLTK, spaCy, and Hugging Face, providing a well-rounded toolkit for solving NLP tasks.

About the Authors



Dr. Goonjan Jain has over a decade of experience in teaching, research, and software development. She is currently an Assistant Professor at Delhi Technological University, Delhi (formerly DCE). Prior to joining academia, she worked with a multinational technology company. Dr. Jain holds a Ph.D. with a specialization in Natural Language Processing (NLP) and an M.Tech. in Computer Science and Technology,

both from Jawaharlal Nehru University (JNU), Delhi. Dr. Jain has made significant contributions to the field of Natural Language Processing, particularly in the areas of language understanding and computational linguistics. She has published many research papers in prestigious international journals and conferences, including *Nature Portfolio's Scientific Reports* and COLING.



Dr. Kanika Garg is a PhD holder in Natural Language Processing (NLP) and currently serves as an Associate Professor at Geetanjali Institute of Technical Studies, Udaipur. With deep expertise in NLP, she has successfully bridged the gap between academia and industry, bringing real-world experience into the classroom. In addition to teaching, Dr. Kanika has authored several

publications in SCI-indexed and Scopus-indexed journals. Her research focuses on advancing NLP techniques and applications in diverse domains. Her work reflects a deep commitment to furthering the understanding of language technologies, both in theory and in practical application. With years of experience in both academic research and industry, Dr. Kanika is passionate about making NLP accessible to a wider audience. This passion is evident in her writing, where she skilfully blends academic rigor with practical insights, providing readers with the tools they need to succeed in the rapidly growing field of NLP.



Sultan Chand & Sons

Publishers of Standard Educational Textbooks

23 Daryaganj, New Delhi-110002
Phones (S) : 011-23281876, 23266105, 41625022
(O) : 011-23247051, 40234454
Email : sultanchand74@yahoo.com
info@sultanchandandsons.com



TC 1319

ISBN 978-93-91820-60-2

