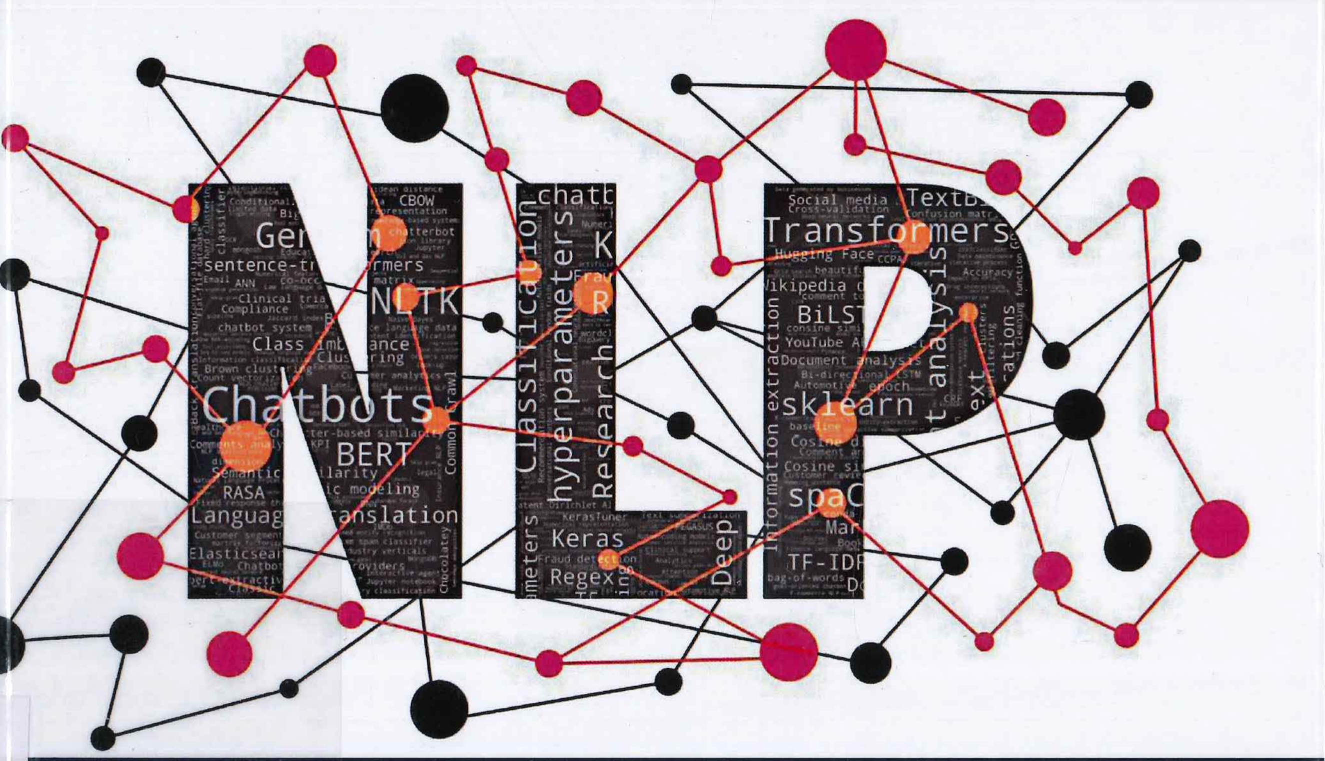
จากงาน CMU Book

DATA SCIENCE SERIES

NATURAL LANGUAGE PROCESSING IN THE REAL WORLD

Text Processing, Analytics, and Classification



JYOTIKA SINGH

CRC Press
Taylor & Francis Group

สำนักหอสมุด มหาวิทยาลัยเชียงใหม่

b16703686
012575884
i22684128

Natural Language Processing in the Real World

Text Processing, Analytics, and Classification

Jyotika Singh

สำนักหอสมุดมหาวิทยาลัยเชียงใหม่
CHIANG MAI UNIVERSITY LIBRARY

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

Contents

List of Figures xv

List of Tables xxi

Preface xxiii

Author Bio xxvii

Acknowledgments xxix

Section I NLP Concepts

Chapter 1 ▪ NLP Basics 5

1.1 NATURAL LANGUAGE PROCESSING 5

1.2 LANGUAGE CONCEPTS 11

1.2.1 Understanding language 11

1.2.2 Components of language 13

1.3 USING LANGUAGE AS DATA 14

1.3.1 Look-up 15

1.3.2 Linguistics 15

1.3.3 Data quantity and relevance 16

1.3.4 Preprocessing 17

1.3.5 Numerical representation 18

1.4 NLP CHALLENGES 19

1.4.1 Language diversity 19

1.4.1.1 Writing styles 19

1.4.1.2 Sentence ambiguities 20

1.4.1.3 Different languages 21

1.4.2 Language evolution 22

1.4.3 Context awareness 22

1.4.4 Not always a one-size-fits-all 23

1.5 SETUP 26
1.6 TOOLS 27

SECTION II Data Curation

CHAPTER 2 ▪ Data Sources and Extraction 35

2.1 SOURCES OF DATA 35
2.1.1 Generated by businesses 35
2.1.2 Openly accessible 36
2.1.3 Conditionally available 38
2.2 DATA EXTRACTION 40
2.2.1 Reading from a PDF 40
2.2.2 Reading from a scanned document 41
2.2.3 Reading from a JSON 43
2.2.4 Reading from a CSV 44
2.2.5 Reading from HTML page (web scraping) 45
2.2.6 Reading from a Word document 46
2.2.7 Reading from APIs 46
2.2.8 Closing thoughts 52
2.3 DATA STORAGE 52
2.3.1 Flat-file database 54
2.3.2 Elasticsearch 54
2.3.2.1 Query examples 57
2.3.3 MongoDB 58
2.3.3.1 Query samples 59
2.3.4 Google BigQuery 61
2.3.4.1 Query examples 61

SECTION III Data Processing and Modeling

CHAPTER 3 ▪ Data Preprocessing and Transformation 75

3.1 DATA CLEANING 75
3.1.1 Segmentation 76
3.1.2 Cleaning 78
3.1.3 Standardization 82
3.1.4 Example scenario 84

3.2 VISUALIZATION 85
3.3 DATA AUGMENTATION 87
3.4 DATA TRANSFORMATION 89
3.4.1 Encoding 90
3.4.2 Frequency-based vectorizers 92
3.4.3 Co-occurrence matrix 94
3.4.4 Word embeddings 95

CHAPTER 4 ■ Data Modeling 105

4.1 DISTANCE METRICS 105
4.1.1 Character-based similarity 105
4.1.2 Phonetic matching 106
4.1.3 Semantic similarity metrics 107
4.2 MODELING 108
4.2.1 Classic ML models 111
4.2.1.1 Clustering 111
4.2.1.2 Classification 113
4.2.2 Deep learning 117
4.2.2.1 Convolutional neural network (CNN) 117
4.2.2.2 Recurrent neural network (RNN) 120
4.2.2.3 Long short term memory (LSTM) 121
4.2.2.4 Bi-directional LSTMs (BiLSTMs) 122
4.2.3 Transformers 123
4.2.3.1 Main innovations behind transformers 124
4.2.3.2 Types of transformer models 125
4.2.3.3 Using transformer models 128
4.2.4 Model hyperparameters 131
4.3 MODEL EVALUATION 132
4.3.1 Metrics 136
4.3.2 Hyperparameter tuning 138

SECTION IV NLP Applications across Industry Verticals

CHAPTER 5 ■ NLP Applications - Active Usage 149

5.1 SOCIAL MEDIA 149
5.1.1 What is social media? 149

5.1.2 Language data generated 149
5.1.3 NLP in social media 149
5.2 FINANCE 153
5.2.1 What is finance? 153
5.2.2 Language data generated 154
5.2.3 NLP in finance 154
5.3 E-COMMERCE 155
5.3.1 What is e-commerce? 155
5.3.2 Language data generated 156
5.3.3 NLP in e-commerce 156
5.4 TRAVEL AND HOSPITALITY 160
5.4.1 What is travel and hospitality? 160
5.4.2 Language data generated 161
5.4.3 NLP in travel and hospitality 161
5.5 MARKETING 163
5.5.1 What is marketing? 163
5.5.2 Language data generated 163
5.5.3 NLP in marketing 163
5.6 INSURANCE 166
5.6.1 What is insurance? 166
5.6.2 Language data generated 166
5.6.3 NLP in insurance 166
5.7 OTHER COMMON USE CASES 170
5.7.1 Writing and email 170
5.7.2 Home assistants 170
5.7.3 Recruiting 171

CHAPTER 6 ■ NLP Applications - Developing Usage 173

6.1 HEALTHCARE 173
6.1.1 What is healthcare? 173
6.1.2 Language data generated 173
6.1.3 NLP in healthcare 173
6.2 LAW 176
6.2.1 What is law? 176
6.2.2 Language data generated 176
6.2.3 NLP in law 176

6.3 REAL ESTATE 178
6.3.1 What is real estate? 178
6.3.2 Language data generated 179
6.3.3 NLP in real estate 179
6.4 OIL AND GAS 181
6.4.1 What is oil and gas? 181
6.4.2 Language data generated 181
6.4.3 NLP in oil and gas 182
6.5 SUPPLY CHAIN 184
6.5.1 What is supply chain? 184
6.5.2 Language data generated 184
6.5.3 NLP in supply chain 184
6.6 TELECOMMUNICATION 186
6.6.1 What is telecom? 186
6.6.2 Language data generated 186
6.6.3 NLP in telecom 186
6.7 AUTOMOTIVE 188
6.7.1 What is automotive? 188
6.7.2 Language data generated 188
6.7.3 NLP in automotive 189
6.8 SERIOUS GAMES 191
6.8.1 What is a serious game? 191
6.8.2 Language data generated 191
6.8.3 NLP in serious games 192
6.9 EDUCATION AND RESEARCH 194
6.9.1 What is education and research? 194
6.9.2 Language data generated 194
6.9.3 NLP in education and research 194

SECTION V Implementing Advanced NLP Applications

CHAPTER 7 ■ Information Extraction and Text Transforming Models 203

7.1 INFORMATION EXTRACTION 203
7.1.1 Named entity recognition (NER) 204
7.1.1.1 Rule-based approaches 204
7.1.1.2 Open-source pre-trained models 205

7.1.1.3 Training your own model 210
7.1.1.4 Fine-tuning on custom datasets using `transformers` 213
7.1.2 Keyphrase extraction (KPE) 217
7.1.2.1 textacy 218
7.1.2.2 rake-nltk 219
7.1.2.3 KeyBERT 219
7.2 TEXT SUMMARIZATION 221
7.2.1 Extractive summarization 221
7.2.1.1 Classic open-source models 221
7.2.1.2 Transformers 223
7.2.2 Abstractive summarization 224
7.2.2.1 Transformers 225
7.3 LANGUAGE DETECTION AND TRANSLATION 227
7.3.1 Language detection 227
7.3.2 Machine translation 227
7.3.2.1 Paid services 227
7.3.2.2 Labeled open-source 228
7.3.2.3 Transformers 229

CHAPTER 8 ▪ Text Categorization and Affinities 231

8.1 TOPIC MODELING 231
8.1.1 Latent dirichlet allocation (LDA) 231
8.2 TEXT SIMILARITY 235
8.2.1 Elasticsearch 235
8.2.2 Classic TF-IDF approach 236
8.2.3 Pre-trained word embedding models 237
8.3 TEXT CLASSIFICATION 239
8.3.1 Off-the-shelf content classifiers 239
8.3.1.1 Zero-shot classification 239
8.3.2 Classifying with available labeled data 241
8.3.2.1 Classic ML 241
8.3.2.2 Deep learning 248
8.3.3 Classifying unlabeled data 253
8.3.3.1 Solution 1: Labeling 253
8.3.3.2 Solution 2: Clustering 253
8.3.3.3 Solution 3: Hybrid approach 253

8.4 SENTIMENT ANALYSIS 254
8.4.1 Classic open-source models 254
8.4.2 Transformers 256
8.4.3 Paid services 257

SECTION VI Implementing NLP Projects in the Real-World

CHAPTER 9 ■ Chatbots 263

9.1 TYPES OF CHATBOTS 263
9.2 COMPONENTS OF A CHATBOT 265
9.3 BUILDING A RULE-BASED CHATBOT 266
9.4 BUILDING A GOAL-ORIENTED CHATBOT 271
9.4.1 Chatbots using service providers 273
9.4.2 Create your own chatbot 274
9.4.3 Using RASA 275
9.5 CLOSING THOUGHTS 285

CHAPTER 10 ■ Customer Review Analysis 287

10.1 HOTEL REVIEW ANALYSIS 287
10.1.1 Sentiment analysis 289
10.1.2 Extracting comment topic themes 291
10.1.3 Unlabeled comment classification into categories 296

CHAPTER 11 ■ Recommendations and Predictions 311

11.1 CONTENT RECOMMENDATION SYSTEM 311
11.1.1 Approaches 311
11.1.2 Building a social media post recommendation system 311
11.1.2.1 Evaluating a classic TF-IDF method, spaCy model, and BERT model 313
11.1.3 Conclusion and closing thoughts 318
11.2 NEXT-WORD PREDICTION 318
11.2.1 Building a next-word prediction for the data science topic 318
11.2.1.1 Training a BiLSTM model 320

CHAPTER 12 ■ More Real-World Scenarios and Tips 325

12.1 DATA SCENARIOS 325
12.2 MODELING SCENARIOS 327
12.3 DEPLOYING YOUR MODEL 331
12.4 MODEL AND OUTCOME EXPLAINABILITY 334

Bibliography 337

Index 357