

Table of Contents

About the Author	xv
About the Technical Reviewer	xvii
Foreword	xix
Acknowledgments	xxi
Introduction	xxiii
Chapter 1: Natural Language Processing Basics	1
Natural Language.....	3
What Is Natural Language?	3
The Philosophy of Language.....	3
Language Acquisition and Usage	6
Linguistics.....	10
Language Syntax and Structure.....	13
Words	15
Phrases.....	17
Clauses	20
Grammar.....	21
Word-Order Typology	33
Language Semantics	35
Lexical Semantic Relations	35
Semantic Networks and Models.....	39
Representation of Semantics	41
Text Corpora	51
Corpora Annotation and Utilities	52
Popular Corpora.....	53
Accessing Text Corpora	55

TABLE OF CONTENTS

Natural Language Processing	62
Machine Translation	62
Speech Recognition Systems	63
Question Answering Systems	64
Contextual Recognition and Resolution	64
Text Summarization	65
Text Categorization	65
Text Analytics	66
Machine Learning	67
Deep Learning	68
Summary.....	68
Chapter 2: Python for Natural Language Processing.....	69
Getting to Know Python.....	70
The Zen of Python	71
Applications: When Should You Use Python?	73
Drawbacks: When Should You Not Use Python?	75
Python Implementations and Versions	76
Setting Up a Robust Python Environment	78
Which Python Version?	78
Which Operating System?	79
Integrated Development Environments	79
Environment Setup	80
Package Management.....	84
Virtual Environments	85
Python Syntax and Structure	88
Working with Text Data	89
String Literals	89
Representing Strings.....	91
String Operations and Methods.....	93

Basic Text Processing and Analysis: Putting It All Together	106
Natural Language Processing Frameworks	111
Summary.....	113
Chapter 3: Processing and Understanding Text.....	115
Text Preprocessing and Wrangling.....	117
Removing HTML Tags	117
Text Tokenization	119
Removing Accented Characters.....	135
Expanding Contractions.....	136
Removing Special Characters.....	138
Case Conversions	138
Text Correction.....	139
Stemming	148
Lemmatization	152
Removing Stopwords.....	154
Bringing It All Together—Building a Text Normalizer	155
Understanding Text Syntax and Structure.....	157
Installing Necessary Dependencies.....	159
Important Machine Learning Concepts.....	162
Parts of Speech Tagging	163
Shallow Parsing or Chunking	172
Dependency Parsing.....	183
Constituency Parsing.....	190
Summary.....	199
Chapter 4: Feature Engineering for Text Representation.....	201
Understanding Text Data	202
Building a Text Corpus	203
Preprocessing Our Text Corpus	205
Traditional Feature Engineering Models.....	208
Bag of Words Model.....	208

TABLE OF CONTENTS

Bag of N-Grams Model	210
TF-IDF Model	211
Extracting Features for New Documents.....	220
Document Similarity	220
Topic Models.....	226
Advanced Feature Engineering Models.....	231
Loading the Bible Corpus.....	233
Word2Vec Model.....	234
Robust Word2Vec Models with Gensim	255
Applying Word2Vec Features for Machine Learning Tasks	258
The GloVe Model	263
Applying <i>GloVe</i> Features for Machine Learning Tasks.....	265
The FastText Model.....	269
Applying FastText Features to Machine Learning Tasks	270
Summary.....	273
Chapter 5: Text Classification	275
What Is Text Classification?	277
Formal Definition	277
Major Text Classification Variants	278
Automated Text Classification	279
Formal Definition	281
Text Classification Task Variants.....	282
Text Classification Blueprint.....	282
Data Retrieval.....	285
Data Preprocessing and Normalization.....	287
Building Train and Test Datasets	292
Feature Engineering Techniques	293
Traditional Feature Engineering Models	294
Advanced Feature Engineering Models	295

Classification Models	296
Multinomial Naïve Bayes	298
Logistic Regression	301
Support Vector Machines.....	303
Ensemble Models	306
Random Forest	307
Gradient Boosting Machines.....	308
Evaluating Classification Models	309
Confusion Matrix	310
Building and Evaluating Our Text Classifier.....	315
Bag of Words Features with Classification Models.....	315
TF-IDF Features with Classification Models	319
Comparative Model Performance Evaluation	322
Word2Vec Embeddings with Classification Models.....	323
GloVe Embeddings with Classification Models.....	326
FastText Embeddings with Classification Models.....	327
Model Tuning	328
Model Performance Evaluation.....	334
Applications	341
Summary.....	341
Chapter 6: Text Summarization and Topic Models	343
Text Summarization and Information Extraction	344
Keyphrase Extraction.....	346
Topic Modeling	346
Automated Document Summarization.....	346
Important Concepts.....	347
Keyphrase Extraction	350
Collocations	351
Weighted Tag-Based Phrase Extraction.....	357

TABLE OF CONTENTS

Topic Modeling.....	362
Topic Modeling on Research Papers	364
The Main Objective	364
Data Retrieval	365
Load and View Dataset	366
Basic Text Wrangling	367
Topic Models with Gensim	368
Text Representation with Feature Engineering.....	369
Latent Semantic Indexing.....	372
Implementing LSI Topic Models from Scratch	382
Latent Dirichlet Allocation.....	389
LDA Models with MALLET	399
LDA Tuning: Finding the Optimal Number of Topics.....	402
Interpreting Topic Model Results	409
Predicting Topics for New Research Papers.....	415
Topic Models with Scikit-Learn.....	418
Text Representation with Feature Engineering.....	419
Latent Semantic Indexing	419
Latent Dirichlet Allocation.....	425
Non-Negative Matrix Factorization.....	428
Predicting Topics for New Research Papers.....	432
Visualizing Topic Models.....	434
Automated Document Summarization	435
Text Wrangling	439
Text Representation with Feature Engineering.....	440
Latent Semantic Analysis	441
TextRank.....	445
Summary.....	450

Chapter 7: Text Similarity and Clustering	453
Essential Concepts.....	455
Information Retrieval (IR).....	455
Feature Engineering	455
Similarity Measures.....	456
Unsupervised Machine Learning Algorithms	457
Text Similarity	457
Analyzing Term Similarity.....	458
Hamming Distance	461
Manhattan Distance	462
Euclidean Distance	464
Levenshtein Edit Distance	465
Cosine Distance and Similarity.....	471
Analyzing Document Similarity	475
Building a Movie Recommender	476
Load and View Dataset.....	477
Text Preprocessing	480
Extract TF-IDF Features	481
Cosine Similarity for Pairwise Document Similarity	482
Find Top Similar Movies for a Sample Movie.....	483
Build a Movie Recommender.....	484
Get a List of Popular Movies	485
Okapi BM25 Ranking for Pairwise Document Similarity.....	488
Document Clustering	497
Clustering Movies	500
Feature Engineering	500
K-Means Clustering	501
Affinity Propagation	508
Ward's Agglomerative Hierarchical Clustering	512
Summary.....	517

TABLE OF CONTENTS

Chapter 8: Semantic Analysis	519
Semantic Analysis.....	520
Exploring WordNet	521
Understanding Synsets.....	522
Analyzing Lexical Semantic Relationships	523
Word Sense Disambiguation	533
Named Entity Recognition.....	536
Building an NER Tagger from Scratch	544
Building an End-to-End NER Tagger with Our Trained NER Model	554
Analyzing Semantic Representations	558
Propositional Logic	558
First Order Logic	560
Summary.....	566
Chapter 9: Sentiment Analysis	567
Problem Statement	568
Setting Up Dependencies.....	569
Getting the Data	569
Text Preprocessing and Normalization.....	570
Unsupervised Lexicon-Based Models	572
Bing Liu’s Lexicon.....	574
MPQA Subjectivity Lexicon	574
Pattern Lexicon.....	575
TextBlob Lexicon.....	575
AFINN Lexicon	578
SentiWordNet Lexicon	580
VADER Lexicon.....	584
Classifying Sentiment with Supervised Learning	587
Traditional Supervised Machine Learning Models	590
Newer Supervised Deep Learning Models.....	593
Advanced Supervised Deep Learning Models.....	602

Analyzing Sentiment Causation	614
Interpreting Predictive Models	614
Analyzing Topic Models	622
Summary.....	629
Chapter 10: The Promise of Deep Learning	631
Why Are We Crazy for Embeddings?	633
Trends in Word-Embedding Models	635
Trends in Universal Sentence-Embedding Models.....	636
Understanding Our Text Classification Problem	642
Universal Sentence Embeddings in Action.....	643
Load Up Dependencies	643
Load and View the Dataset.....	644
Building Train, Validation, and Test Datasets	645
Basic Text Wrangling	645
Build Data Ingestion Functions.....	647
Build Deep Learning Model with Universal Sentence Encoder.....	648
Model Training	649
Model Evaluation	651
Bonus: Transfer Learning with Different Universal Sentence Embeddings.....	652
Summary and Future Scope	659
Index.....	661