Table of Contents

Preface ix

1. Language and Computation 1
The Data Science Paradigm 2
Language-Aware Data Products 4
The Data Product Pipeline 5
Language as Data 8
A Computational Model of Language 8
Language Features 10
Contextual Features 13
Structural Features 15
Conclusion 16

2. Building a Custom Corpus 19
What Is a Corpus? 19
Domain-Specific Corpora 20
The Baleen Ingestion Engine 21
Corpus Data Management 22
Corpus Disk Structure 24
Corpus Readers 27
Streaming Data Access with NLTK 28
Reading an HTML Corpus 31
Reading a Corpus from a Database 34
Conclusion 36

3. Corpus Preprocessing and Wrangling 37
Breaking Down Documents 38
Identifying and Extracting Core Content 38

Deconstructing Documents into Paragraphs 39
Segmentation: Breaking Out Sentences 42
Tokenization: Identifying Individual Tokens 43
Part-of-Speech Tagging 44
Intermediate Corpus Analytics 45
Corpus Transformation 47
Intermediate Preprocessing and Storage 48
Reading the Processed Corpus 51
Conclusion 53

4. Text Vectorization and Transformation Pipelines. 55
Words in Space 56
Frequency Vectors 57
One-Hot Encoding 59
Term Frequency–Inverse Document Frequency 62
Distributed Representation 65
The Scikit-Learn API 68
The BaseEstimator Interface 68
Extending TransformerMixin 70
Pipelines 74
Pipeline Basics 75
Grid Search for Hyperparameter Optimization 76
Enriching Feature Extraction with Feature Unions 77
Conclusion 79

5. Classification for Text Analysis. 81
Text Classification 82
Identifying Classification Problems 82
Classifier Models 84
Building a Text Classification Application 85
Cross-Validation 86
Model Construction 89
Model Evaluation 91
Model Operationalization 94
Conclusion 95

6. Clustering for Text Similarity. 97
Unsupervised Learning on Text 97
Clustering by Document Similarity 99
Distance Metrics 99
Partitive Clustering 102
Hierarchical Clustering 107

Modeling Document Topics 111
Latent Dirichlet Allocation 111
Latent Semantic Analysis 119
Non-Negative Matrix Factorization 121
Conclusion 123

7. Context-Aware Text Analysis..... 125
Grammar-Based Feature Extraction 126
Context-Free Grammars 126
Syntactic Parsers 127
Extracting Keyphrases 128
Extracting Entities 131
n-Gram Feature Extraction 132
An n-Gram-Aware CorpusReader 133
Choosing the Right n-Gram Window 135
Significant Collocations 136
n-Gram Language Models 139
Frequency and Conditional Frequency 140
Estimating Maximum Likelihood 143
Unknown Words: Back-off and Smoothing 145
Language Generation 147
Conclusion 149

8. Text Visualization..... 151
Visualizing Feature Space 152
Visual Feature Analysis 152
Guided Feature Engineering 162
Model Diagnostics 170
Visualizing Clusters 170
Visualizing Classes 172
Diagnosing Classification Error 173
Visual Steering 177
Silhouette Scores and Elbow Curves 177
Conclusion 180

9. Graph Analysis of Text..... 183
Graph Computation and Analysis 185
Creating a Graph-Based Thesaurus 185
Analyzing Graph Structure 186
Visual Analysis of Graphs 187
Extracting Graphs from Text 189
Creating a Social Graph 189

Insights from the Social Graph 192
Entity Resolution 200
Entity Resolution on a Graph 201
Blocking with Structure 202
Fuzzy Blocking 202
Conclusion 205

10. Chatbots..... 207
Fundamentals of Conversation 208
Dialog: A Brief Exchange 210
Maintaining a Conversation 213
Rules for Polite Conversation 215
Greetings and Salutations 216
Handling Miscommunication 220
Entertaining Questions 222
Dependency Parsing 223
Constituency Parsing 225
Question Detection 227
From Tablespoons to Grams 229
Learning to Help 233
Being Neighborly 235
Offering Recommendations 238
Conclusion 240

11. Scaling Text Analytics with Multiprocessing and Spark..... 241
Python Multiprocessing 242
Running Tasks in Parallel 244
Process Pools and Queues 249
Parallel Corpus Preprocessing 251
Cluster Computing with Spark 253
Anatomy of a Spark Job 254
Distributing the Corpus 255
RDD Operations 257
NLP with Spark 259
Conclusion 270

12. Deep Learning and Beyond..... 273
Applied Neural Networks 274
Neural Language Models 274
Artificial Neural Networks 275
Deep Learning Architectures 280
Sentiment Analysis 284

Deep Structure Analysis 286
The Future Is (Almost) Here 291

Glossary **293**

Index **303**