

TABLE OF CONTENTS

Series Editor's Introduction	xiii
Preface	xv
Acknowledgments	xvii
About the Authors	xix
 Chapter 1: Introduction	 1
1.1 Text Data	1
1.1.1 <i>Introducing the Definitions</i>	2
1.1.2 <i>Types of Text Data</i>	4
1.1.3 <i>File Formats to Save and Store Text Information</i>	5
1.2 The Two Applications Considered in This Book	6
1.3 Introductory Example and Its Analysis Using the R Statistical Software	7
1.4 The Introductory Example Revisited, Illustrating Concordance and Collocation Using Alternative Software	22
1.5 Concluding Remarks	24
1.6 References	25
 Chapter 2: A Description of the Studied Text Corpora and A Discussion of Our Modeling Strategy	 26
2.1 Introduction to the Corpora: Selecting the Texts	26
2.2 Debates of the 39th U.S. Congress, as recorded in the <i>Congressional Globe</i>	27
2.3 <i>The Territorial Papers of the United States</i>	29
2.4 Analyzing Text Data: Bottom-Up or Top-Down Analysis	32
2.5 References	34
Appendix to Chapter 2: The Complete Congressional Record	35
 Chapter 3: Preparing Text for Analysis: Text Cleaning and Formatting	 36
3.1 Text Cleaning	36
3.1.1 <i>Compacting Multiple Word Sets Into a Single Word</i>	42

3.2 Text Formatting	43
3.2.1 <i>Formatting by Marking Versus Formatting by Deleting</i>	44
3.2.2 <i>Formatting Beyond Metavariables: Telling the Computer What Sections to Skip When Running the Analysis</i>	44
3.3 Concluding Remarks	46
3.4 References	48
Chapter 4: Word Distributions: Document-Term Matrices of Word Frequencies and the “Bag of Words” Representation	49
4.1 Document-Term Matrices of Frequencies	49
4.1.1 <i>Creating the Document-Term Matrix in R</i>	51
4.1.2 <i>Dropping Sparse Words That Do Not Occur in Many Documents</i>	52
4.2 Displaying Word Frequencies	53
4.3 Co-Occurrence of Terms in the Same Document	56
4.4 The Zipf Law: An Interesting Fact About the Distribution of Word Frequencies	59
4.5 References	61
Chapter 5: Metavariables and Text Analysis Stratified on Metavariables	62
5.1 The Significance of Stratification and the Importance of Metavariables	62
5.2 Analysis of the <i>Territorial Papers</i>	63
5.2.1 <i>Territorial Papers: Visualization of the Metavariables</i>	64
5.2.2 <i>Territorial Papers: Stratified Text Analysis</i>	69
5.3 Analysis of <i>Speeches From the 39th Congress</i>	72
5.3.1 <i>Speeches From the 39th Congress: Visualization of the Metavariables</i>	73
5.3.2 <i>Speeches From the 39th Congress: Stratified Text Analysis</i>	77
5.4 References	83
Chapter 6: Sentiment Analysis	84
6.1 Lexicons of Sentiment-Charged Words	84
6.1.1 <i>Attaching Sentiment to a Document</i>	85
6.1.2 <i>Sentiment Analysis for the Corpus and Its Documents</i>	87
6.1.3 <i>Importance of Sentiment Analysis</i>	88
6.2 Applying Sentiment Analysis to the Letters of the <i>Territorial Papers</i>	88

6.3 Using Other Sentiment Dictionaries and the R Software tidytext for Sentiment Analysis	91
6.4 Concluding Remarks: An Alternative Approach for Sentiment Analysis	94
6.5 References	95
Chapter 7: Clustering of Documents	97
7.1 Clustering Documents	97
7.2 Measures for the Closeness and the Distance of Documents	98
7.3 Methods for Clustering Documents	101
7.3.1 Hierarchical Agglomerative Clustering and Dendrograms	101
7.3.2 <i>k</i> -Means Clustering	103
7.3.3 Additional Remarks	105
7.4 Illustrating Clustering Methods on a Simulated Example	106
7.5 References	109
Chapter 8: Classification of Documents	110
8.1 Introduction	110
8.2 Classification Procedures	111
8.2.1 The <i>k</i> -Nearest Neighbor Algorithm	111
8.2.2 Naive Bayesian Analysis	113
8.2.3 Fisher Linear Discriminant Method and Linear Scoring (SVM) Methods	115
8.2.4 Evaluating Classification Rules on Hold-Out Samples	116
8.3 Two Examples Using the Congressional Speech Database	116
8.4 Concluding Remarks on Authorship Attribution: Commenting on the Field of Stylometry	119
8.5 References	120
Chapter 9: Modeling Text Data: Topic Models	121
9.1 Topic Models	121
9.1.1 Some More Technical Details and a Brief Primer on Dirichlet Distributions	126
9.1.2 Model Extensions and Useful Software, With a Tip of the Hat to Their Developers	128
9.1.3 Further Comments	129
9.2 Fitting Topic Models to the Two Corpora Studied in This Book	130
9.2.1 Topic Models for the Corpus of the Territorial Papers	130
9.2.2 Topic Models for the Corpus of Speeches From the 39th U.S. Congress	134
9.3 References	140

Chapter 10: <i>n</i>-Grams and Other Ways of Analyzing	
Adjacent Words	142
10.1 Analysis of Bigrams	142
10.2 Text Windows to Measure Word Associations Within a Neighborhood of Words and a Discussion of the R Package <code>text2vec</code>	143
10.3 Illustrating the Use of <i>n</i> -Grams: <i>Speeches of the 39th Congress</i>	146
Chapter 11: Concluding Remarks	151
Appendix: Listing of Website Resources	155
Index	161