# Contents

**Foreword**  
†

1  **What is Text Mining?**  
1.1  What is it?  
1.1.1  What is Text Mining in Practice?  
1.1.2  Where Does Text Mining Fit?  
1.2  Why We Care About Text Mining  
1.2.1  What Are the Consequences of Ignoring Text?  
1.2.2  What Are the Benefits of Text Mining?  
1.2.3  Setting Expectations: When Text Mining Should (and Should Not) Be Used  
1.3  A Basic Workflow – How the Process Works  
1.4  What Tools Do I Need to Get Started with This?  
1.5  A Simple Example  
1.6  A Real World Use Case  
1.7  Summary

2  **Basics of Text Mining**  
2.1  What is Text Mining in a Practical Sense?  
2.2  Types of Text Mining: Bag of Words  
2.2.1  Types of Text Mining: Syntactic Parsing  
2.3  The Text Mining Process in Context  
2.4  String Manipulation: Number of Characters and Substitutions  
2.4.1  String Manipulations: Paste, Character Splits and Extractions  
2.5  Keyword Scanning  
2.6  String Packages stringr and stringi  
2.7  Preprocessing Steps for Bag of Words Text Mining  
2.8  Spellcheck  
2.9  Frequent Terms and Associations  
2.10  DeltaAssist Wrap Up  
2.11  Summary
3 Common Text Mining Visualizations 51
3.1 A Tale of Two (or Three) Cultures 51
3.2 Simple Exploration: Term Frequency, Associations and Word Networks 53
3.2.1 Term Frequency 54
3.2.2 Word Associations 57
3.2.3 Word Networks 59
3.3 Simple Word Clusters: Hierarchical Dendrograms 67
3.4 Word Clouds: Overused but Effective 73
3.4.1 One Corpus Word Clouds 74
3.4.2 Comparing and Contrasting Corpora in Word Clouds 75
3.4.3 Polarized Tag Plot 79
3.5 Summary 83

4 Sentiment Scoring 85
4.1 What is Sentiment Analysis? 85
4.2 Sentiment Scoring: Parlor Trick or Insightful? 88
4.3 Polarity: Simple Sentiment Scoring 89
4.3.1 Subjectivity Lexicons 89
4.3.2 Qdap’s Scoring for Positive and Negative Word Choice 93
4.4 Revisiting Word Clouds – Sentiment Word Clouds 96
4.4.1 Symbol-Based Emoticons Native to R 105
4.4.2 Punctuation Based Emoticons 106
4.4.3 Emoji 108
4.5 R’s Archived Sentiment Scoring Library 113
4.6 Sentiment the Tidytext Way 118
4.7 Airbnb.com Boston Wrap Up 126
4.8 Summary 126

5 Hidden Structures: Clustering, String Distance, Text Vectors and Topic Modeling 129
5.1 What is clustering? 129
5.1.1 K-Means Clustering 130
5.1.2 Spherical K-Means Clustering 139
5.1.3 K-Mediod Clustering 144
5.1.4 Evaluating the Cluster Approaches 145
5.2 Calculating and Exploring String Distance 147
5.2.1 What is String Distance? 148
5.2.2 Fuzzy Matching – Amatch, Ain 151
5.2.3 Similarity Distances – Stringdist, Stringdistmatrix 152
5.3 LDA Topic Modeling Explained 154
5.3.1 Topic Modeling Case Study 156
5.3.2 LDA and LDAvis 158
5.4 Text to Vectors using text2vec 169
5.4.1 Text2vec 171
5.5 Summary 179

6 Document Classification: Finding Clickbait from Headlines 181
6.1 What is Document Classification? 181
6.2 Clickbait Case Study 183
6.2.1 Session and Data Set-Up 185
6.2.2 GLMNet Training 188
6.2.3 GLMNet Test Predictions 196
6.2.4 Test Set Evaluation 198
6.2.5 Finding the Most Impactful Words 200
6.2.6 Case Study Wrap Up: Model Accuracy and Improving Performance Recommendations 206
6.3 Summary 207

7 Predictive Modeling: Using Text for Classifying and Predicting Outcomes 209
7.1 Classification vs Prediction 209
7.2 Case Study I: Will This Patient Come Back to the Hospital? 210
7.2.1 Patient Readmission in the Text Mining Workflow 211
7.2.2 Session and Data Set-Up 211
7.2.3 Patient Modeling 214
7.2.4 More Model KPIs: AUC, Recall, Precision and F1 216
7.2.4.1 Additional Evaluation Metrics 218
7.2.5 Apply the Model to New Patients 222
7.2.6 Patient Readmission Conclusion 223
7.3 Case Study II: Predicting Box Office Success 224
7.3.1 Opening Weekend Revenue in the Text Mining Workflow 225
7.3.2 Session and Data Set-Up 225
7.3.3 Opening Weekend Modeling 228
7.3.4 Model Evaluation 231
7.3.5 Apply the Model to New Movie Reviews 234
7.3.6 Movie Revenue Conclusion 235
7.4 Summary 236

8 The OpenNLP Project 237
8.1 What is the OpenNLP project? 237
8.2 R’s OpenNLP Package 238
8.3 Named Entities in Hillary Clinton’s Email 242
8.3.1 R Session Set-Up 243
8.3.2 Minor Text Cleaning 245
## Contents

8.3.3 Using OpenNLP on a single email  
8.3.4 Using OpenNLP on Multiple Documents  
8.3.5 Revisiting the Text Mining Workflow  
8.4 Analyzing the Named Entities  
8.4.1 Worldwide Map of Hillary Clinton’s Location Mentions  
8.4.2 Mapping Only European Locations  
8.4.3 Entities and Polarity: How Does Hillary Clinton Feel About an Entity?  
8.4.4 Stock Charts for Entities  
8.4.5 Reach an Insight or Conclusion About Hillary Clinton’s Emails  
8.6 Summary  

9 Text Sources  
9.1 Sourcing Text  
9.2 Web Sources  
9.2.1 Web Scraping a Single Page with rvest  
9.2.2 Web Scraping Multiple Pages with rvest  
9.2.3 Application Program Interfaces (APIs)  
9.2.4 Newspaper Articles from the Guardian Newspaper  
9.2.5 Tweets Using the twitteR Package  
9.2.6 Calling an API Without a Dedicated R Package  
9.2.7 Using Jsonlite to Access the New York Times  
9.2.8 Using RCurl and XML to Parse Google Newsfeeds  
9.2.9 The tm Library Web-Mining Plugin  
9.3 Getting Text from File Sources  
9.3.1 Individual CSV, TXT and Microsoft Office Files  
9.3.2 Reading Multiple Files Quickly  
9.3.3 Extracting Text from PDFs  
9.3.4 Optical Character Recognition: Extracting Text from Images  
9.4 Summary  

Index