



# TEXT MINING IN PRACTICE WITH R

**A reliable, cost-effective approach to extracting priceless business information from all sources of text**

Excavating actionable business insights from data is a complex undertaking, and that complexity is magnified by an order of magnitude when the focus is on documents and other text information. This book takes a practical, hands-on approach to teaching you a reliable, cost-effective approach to mining the vast, untold riches buried within all forms of text using R.

Author Ted Kwartler clearly describes all of the tools needed to perform text mining and shows you how to use them to identify practical business applications to get your creative text mining efforts started right away. With the help of numerous real-world examples and case studies from industries ranging from healthcare to entertainment to telecommunications, he demonstrates how to execute an array of text mining processes and functions, including sentiment scoring, topic modelling, predictive modelling, extracting clickbait from headlines, and more. You'll learn how to:

- Identify actionable social media posts to improve customer service
- Use text mining in HR to identify candidate perceptions of an organisation, match job descriptions with resumes, and more
- Extract priceless information from virtually all digital and print sources, including the news media, social media sites, PDFs, and even JPEG and GIF image files
- Make text mining an integral component of marketing in order to identify brand evangelists, impact customer propensity modelling, and much more

Most companies' data mining efforts focus almost exclusively on numerical and categorical data, while text remains a largely untapped resource. Especially in a global marketplace where being first to identify and respond to customer needs and expectations imparts an unbeatable competitive advantage, text represents a source of immense potential value. Unfortunately, there is no reliable, cost-effective technology for extracting analytical insights from the huge and ever-growing volume of text available online and other digital sources, as well as from paper documents—until now.

**TED KWARTLER** is a data science instructor at DataCamp.com. He has worked in analytical and executive roles at DataRobot, Liberty Mutual Insurance and Amazon.com.

Cover Design: Wiley  
Cover Image: © ChrisPole/Gettyimages

[www.wiley.com](http://www.wiley.com)

# WILEY



Also available  
as an e-book

ISBN 978-1-119-28201-3



9 781119 282013



# Contents

## Foreword *xi*

<b>1</b>	<b>What is Text Mining?</b>	<b>1</b>
1.1	What is it?	1
1.1.1	What is Text Mining in Practice?	2
1.1.2	Where Does Text Mining Fit?	2
1.2	Why We Care About Text Mining	2
1.2.1	What Are the Consequences of Ignoring Text?	3
1.2.2	What Are the Benefits of Text Mining?	5
1.2.3	Setting Expectations: When Text Mining Should (and Should Not) Be Used	6
1.3	A Basic Workflow – How the Process Works	9
1.4	What Tools Do I Need to Get Started with This?	12
1.5	A Simple Example	12
1.6	A Real World Use Case	13
1.7	Summary	15
<b>2</b>	<b>Basics of Text Mining</b>	<b>17</b>
2.1	What is Text Mining in a Practical Sense?	17
2.2	Types of Text Mining: Bag of Words	20
2.2.1	Types of Text Mining: Syntactic Parsing	22
2.3	The Text Mining Process in Context	24
2.4	String Manipulation: Number of Characters and Substitutions	25
2.4.1	String Manipulations: Paste, Character Splits and Extractions	29
2.5	Keyword Scanning	33
2.6	String Packages stringr and stringi	36
2.7	Preprocessing Steps for Bag of Words Text Mining	37
2.8	Spellcheck	44
2.9	Frequent Terms and Associations	47
2.10	DeltaAssist Wrap Up	49
2.11	Summary	49



3	<b>Common Text Mining Visualizations</b>	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	<b>Sentiment Scoring</b>	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
4.3.3	Revisiting Word Clouds – Sentiment Word Clouds	96
4.4	Emoticons – Dealing with These Perplexing Clues	103
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	<b>Hidden Structures: Clustering, String Distance, Text Vectors and Topic Modeling</b>	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
<b>6</b>	<b>Document Classification: Finding Clickbait from Headlines</b>	<b>181</b>
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
<b>7</b>	<b>Predictive Modeling: Using Text for Classifying and Predicting Outcomes</b>	<b>209</b>
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
<b>8</b>	<b>The OpenNLP Project</b>	<b>237</b>
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



8.3.3	Using OpenNLP on a single email	246
8.3.4	Using OpenNLP on Multiple Documents	251
8.3.5	Revisiting the Text Mining Workflow	254
8.4	Analyzing the Named Entities	255
8.4.1	Worldwide Map of Hillary Clinton's Location Mentions	256
8.4.2	Mapping Only European Locations	259
8.4.3	Entities and Polarity: How Does Hillary Clinton Feel About an Entity?	262
8.4.4	Stock Charts for Entities	266
8.4.5	Reach an Insight or Conclusion About Hillary Clinton's Emails	268
8.6	Summary	269
<b>9</b>	<b>Text Sources</b>	<b>271</b>
9.1	Sourcing Text	271
9.2	Web Sources	272
9.2.1	Web Scraping a Single Page with rvest	272
9.2.2	Web Scraping Multiple Pages with rvest	276
9.2.3	Application Program Interfaces (APIs)	282
9.2.4	Newspaper Articles from the Guardian Newspaper	283
9.2.5	Tweets Using the twitterR Package	285
9.2.6	Calling an API Without a Dedicated R Package	287
9.2.7	Using Jsonlite to Access the New York Times	288
9.2.8	Using RCurl and XML to Parse Google Newsfeeds	290
9.2.9	The tm Library Web-Mining Plugin	292
9.3	Getting Text from File Sources	293
9.3.1	Individual CSV, TXT and Microsoft Office Files	294
9.3.2	Reading Multiple Files Quickly	296
9.3.3	Extracting Text from PDFs	298
9.3.4	Optical Character Recognition: Extracting Text from Images	299
9.4	Summary	302
	<b>Index</b>	<b>305</b>