

Advances in Analytics and Data Science

Murugan Anandarajan
Chelsey Hill
Thomas Nolan

Practical Text Analytics

Maximizing the Value of Text Data

 Springer

Advances in Analytics and Data Science

Volume 2

Series Editors

Ramesh Sharda

Oklahoma State University, Stillwater, OK, USA

Hsinchun Chen

University of Arizona, Tucson, AZ, USA

More information about this series at <http://www.springer.com/series/15876>

Murugan Anandarajan • Chelsey Hill
Thomas Nolan

Practical Text Analytics

Maximizing the Value of Text Data

Murugan Anandarajan
LeBow College of Business
Drexel University
Philadelphia, PA, USA

Chelsey Hill
Feliciano School of Business
Montclair State University
Montclair, NJ, USA

Thomas Nolan
Mercury Data Science
Houston, TX, USA

ISSN 2522-0233 ISSN 2522-0241 (electronic)
Advances in Analytics and Data Science
ISBN 978-3-319-95662-6 ISBN 978-3-319-95663-3 (eBook)
<https://doi.org/10.1007/978-3-319-95663-3>

Library of Congress Control Number: 2018955905

© Springer Nature Switzerland AG 2019

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors, and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

To my aunts and uncle—MA
To my angel mother, Deborah—CH
To my dad—TN

Preface

The oft-cited statistic that “80% of data is unstructured” reminds industry leaders and analytics professionals about the vast volume of untapped text data resources. Predictably, there has been an increasing focus on text analytics to generate information resources. In fact, the expected growth in this market is projected to be \$5.93 billion by 2020!¹

Whenever businesses capture text data, they want to capitalize on the information hidden within. Beginning with early adopters in the intelligence and biomedical communities, text analytics has expanded to include applications across industries, including manufacturing, insurance, healthcare, education, safety and security, publishing, telecommunications, and politics. The broad range of applied text analytics requires practitioners in this field.

Our goal is to democratize text analytics and increase the number of people using text data for research. We hope this book lowers the barrier of entry for analyzing text data, making it more accessible for people to uncover value-added text information.

This book covers the elements involved in creating a text mining pipeline. While analysts will not use every element in every project, each tool provides a potential segment in the final pipeline. Understanding the options is key to choosing the appropriate elements in designing and conducting text analysis.

The book is divided into five parts. The first part provides an overview of the text analytics process by introducing text analytics, discussing the relationship with content analysis, and providing a general overview of the process.

Next, the chapter moves on to the actual practice of text analytics, beginning with planning the project. The next part covers the methods of data preparation and pre-processing. Once the data is prepared, the next step is the analysis. Here, we describe the array of analysis options. The part concludes with a discussion about reporting options, indicating the benefits of various choices for convincing others about the value of the analysis.

¹ <http://www.marketsandmarkets.com/PressReleases/text-analytics.asp>

The last part of the book demonstrates the use of various software programs and programming languages for text analytics. We hope these examples provide the reader with practical examples of how information hidden within text data can be mined.

Philadelphia, PA, USA
Montclair, NJ, USA
Houston, TX, USA

Murugan Anandarajan
Chelsey Hill
Thomas Nolan

Acknowledgments

The authors wish to acknowledge the invaluable contributions of several individuals to the preparation of the manuscript of this book.

Diana Jones, Director of the Center for Business Analytics and the Dornsife Office for Experiential Learning, at the LeBow College of Business, Drexel University, for her chapter on *Storytelling Using Text Data*.

Jorge Fresneda Fernandez, Assistant Professor of Marketing at the Martin Tuchman School of Management, New Jersey Institute of Technology, for his chapters on *Latent Semantic Analysis (LSA) in Python* and *SAS Visual Text Analytics*.

We thank Diana and Jorge for their expertise and invaluable contributions to this book.

We also thank Irena Nedelcu, Rajiv Nag, and Stacy Boyer, all of the LeBow College of Business, Drexel University, for providing valuable comments on various chapters.

Our appreciation to Matthew Amboy and his team at Springer who made the publication of this book possible.

Murugan Anandarajan
Chelsey Hill
Thomas Nolan

Contents

1	Introduction to Text Analytics	1
1.1	Introduction	1
1.2	Text Analytics: What Is It?	1
1.3	Origins and Timeline of Text Analytics	3
1.4	Text Analytics in Business and Industry	5
1.5	Text Analytics Skills	6
1.6	Benefits of Text Analytics	8
1.7	Text Analytics Process Road Map	8
1.7.1	Planning	8
1.7.2	Text Preparing and Preprocessing	9
1.7.3	Text Analysis Techniques	9
1.7.4	Communicating the Results	10
1.8	Examples of Text Analytics Software	10
	References	10
 Part I Planning the Text Analytics Project		
2	The Fundamentals of Content Analysis	15
2.1	Introduction	15
2.2	Deductive Versus Inductive Approaches	16
2.2.1	Content Analysis for Deductive Inference	16
2.2.2	Content Analysis for Inductive Inference	16
2.3	Unitizing and the Unit of Analysis	18
2.3.1	The Sampling Unit	19
2.3.2	The Recording Unit	19
2.3.3	The Context Unit	20
2.4	Sampling	20
2.5	Coding and Categorization	20
2.6	Examples of Inductive and Deductive Inference Processes	23
2.6.1	Inductive Inference	23

2.6.2	Deductive Inference	24
	References	25
3	Planning for Text Analytics	27
3.1	Introduction	27
3.2	Initial Planning Considerations	28
3.2.1	Drivers	29
3.2.2	Objectives	29
3.2.3	Data	30
3.2.4	Cost	30
3.3	Planning Process	30
3.4	Problem Framing	31
3.4.1	Identifying the Analysis Problem	31
3.4.2	Inductive or Deductive Inference	32
3.5	Data Generation	32
3.5.1	Definition of the Project's Scope and Purpose	32
3.5.2	Text Data Collection	33
3.5.3	Sampling	35
3.6	Method and Implementation Selection	38
3.6.1	Analysis Method Selection	38
3.6.2	The Selection of Implementation Software	39
	References	40
 Part II Text Preparation		
4	Text Preprocessing	45
4.1	Introduction	45
4.2	The Preprocessing Process	46
4.3	Unitize and Tokenize	46
4.3.1	N-Grams	48
4.4	Standardization and Cleaning	50
4.5	Stop Word Removal	50
4.5.1	Custom Stop Word Dictionaries	53
4.6	Stemming and Lemmatization	53
4.6.1	Syntax and Semantics	53
4.6.2	Stemming	54
4.6.3	Lemmatization	56
4.6.4	Part-of-Speech (POS) Tagging	58
	References	59
5	Term-Document Representation	61
5.1	Introduction	61
5.2	The Inverted Index	61
5.3	The Term-Document Matrix	64
5.4	Term-Document Matrix Frequency Weighting	65
5.4.1	Local Weighting	66

5.4.2	Global Weighting	68
5.4.3	Combinatorial Weighting: Local and Global Weighting . .	71
5.5	Decision-Making	73
	References	73

Part III Text Analysis Techniques

6	Semantic Space Representation and Latent Semantic Analysis	77
6.1	Introduction	77
6.2	Latent Semantic Analysis (LSA)	79
6.2.1	Singular Value Decomposition (SVD)	79
6.2.2	LSA Example	80
6.3	Cosine Similarity	84
6.4	Queries in LSA	87
6.5	Decision-Making: Choosing the Number of Dimensions	88
	References	91
7	Cluster Analysis: Modeling Groups in Text	93
7.1	Introduction	93
7.2	Distance and Similarity	94
7.3	Hierarchical Cluster Analysis	98
7.3.1	Hierarchical Cluster Analysis Algorithm	99
7.3.2	Graph Methods	99
7.3.3	Geometric Methods	101
7.3.4	Advantages and Disadvantages of HCA	103
7.4	k -Means Clustering	103
7.4.1	kMC Algorithm	104
7.4.2	The kMC Process	104
7.4.3	Advantages and Disadvantages of kMC	108
7.5	Cluster Analysis: Model Fit and Decision-Making	109
7.5.1	Choosing the Number of Clusters	109
7.5.2	Naming/Describing Clusters	112
7.5.3	Evaluating Model Fit	113
7.5.4	Choosing the Cluster Analysis Model	114
	References	115
8	Probabilistic Topic Models	117
8.1	Introduction	117
8.2	Latent Dirichlet Allocation (LDA)	119
8.3	Correlated Topic Model (CTM)	120
8.4	Dynamic Topic Model (DT)	122
8.5	Supervised Topic Model (sLDA)	123
8.6	Structural Topic Model (STM)	124
8.7	Decision Making in Topic Models	124
8.7.1	Assessing Model Fit and Number of Topics	124

8.7.2	Model Validation and Topic Identification	126
8.7.3	When to Use Topic Models	128
	References	129
9	Classification Analysis: Machine Learning Applied to Text	131
9.1	Introduction	131
9.2	The General Text Classification Process	132
9.3	Evaluating Model Fit	132
9.3.1	Confusion Matrices/Contingency Tables	132
9.3.2	Overall Model Measures	134
9.3.3	Class-Specific Measures	135
9.4	Classification Models	137
9.4.1	Naïve Bayes	137
9.4.2	<i>k</i> -Nearest Neighbors (kNN)	138
9.4.3	Support Vector Machines (SVM)	140
9.4.4	Decision Trees	141
9.4.5	Random Forests	143
9.4.6	Neural Networks	144
9.5	Choosing a Classification	146
9.5.1	Model Fit	146
	References	148
10	Modeling Text Sentiment: Learning and Lexicon Models	151
10.1	Lexicon Approach	152
10.2	Machine Learning Approach	158
10.2.1	Naïve Bayes (NB)	159
10.2.2	Support Vector Machines (SVM)	160
10.2.3	Logistic Regression	161
10.3	Sentiment Analysis Performance: Considerations and Evaluation	162
	References	164
Part IV Communicating the Results		
11	Storytelling Using Text Data	167
11.1	Introduction	167
11.2	Telling Stories About the Data	168
11.3	Framing the Story	170
11.3.1	Storytelling Framework	170
11.3.2	Applying the Framework	171
11.4	Organizations as Storytellers	173
11.4.1	United Parcel Service	173
11.4.2	Zillow	173
11.5	Data Storytelling Checklist	174
	References	175

12 Visualizing Analysis Results	177
12.1 Strategies for Effective Visualization	178
12.1.1 Be Purposeful	178
12.1.2 Know the Audience	178
12.1.3 Solidify the Message	178
12.1.4 Plan and Outline	179
12.1.5 Keep It Simple	179
12.1.6 Focus Attention	180
12.2 Visualization Techniques in Text Analytics	180
12.2.1 Corpus/Document Collection-Level Visualizations	180
12.2.2 Theme and Category-Level Visualizations	182
12.2.3 Document-Level Visualizations	188
References	190

Part V Text Analytics Examples

13 Sentiment Analysis of Movie Reviews Using R	193
13.1 Introduction to R and RStudio	193
13.2 SA Data and Data Import	194
13.3 Objective of the Sentiment Analysis	197
13.4 Data Preparation and Preprocessing	199
13.4.1 Tokenize	199
13.4.2 Remove Stop Words	200
13.5 Sentiment Analysis	201
13.6 Sentiment Analysis Results	204
13.7 Custom Dictionary	206
13.8 Out-of-Sample Comparison	217
References	219
14 Latent Semantic Analysis (LSA) in Python	221
14.1 Introduction to Python and IDLE	221
14.2 Preliminary Steps	222
14.3 Getting Started	224
14.4 Data and Data Import	225
14.5 Analysis	227
Further Reading	242
15 Learning-Based Sentiment Analysis Using RapidMiner	243
15.1 Introduction	243
15.2 Getting Started in RapidMiner	244
15.3 Text Data Import	247
15.4 Text Preparation and Preprocessing	249
15.5 Text Classification Sentiment Analysis	256
Reference	261

16 SAS Visual Text Analytics	263
16.1 Introduction	263
16.2 Getting Started	264
16.3 Analysis	266
Further Reading	282
Index	283

About the Authors

Murugan Anandarajan is a Professor of MIS at Drexel University. His current research interests lie in the intersections of crime, IoT, and analytics. His work has been published in journals such as *Decision Sciences*, *Journal of Management Information Systems*, and *Journal of International Business Studies*. He co-authored eight books, including *The Internet and Workplace Transformation* (2006) and its follow-up volume, *The Internet of People, Things and Services* (2018). He has been awarded over \$2.5 million in research grants from various government agencies including the National Science Foundation, the US Department of Justice, the National Institute of Justice, and the State of PA.

Chelsey Hill is an Assistant Professor of Business Analytics in the Information Management and Business Analytics Department of the Feliciano School of Business at Montclair State University. She holds a BA in Political Science from the College of New Jersey, an MS in Business Intelligence from Saint Joseph's University, and a PhD in Business Administration with a concentration in Decision Sciences from Drexel University. Her research interests include consumer product recalls, online consumer reviews, safety and security, public policy, and humanitarian operations. Her research has been published in the *Journal of Informetrics* and the *International Journal of Business Intelligence Research*.

Tom Nolan completed his undergraduate work at Kenyon College. After Kenyon, he attended Drexel University where he graduated with an MS in Business Analytics. From there, he worked at Independence Blue Cross in Philadelphia, PA, and Anthem Inc. in Houston, TX. Currently, he works with all types of data as a Data Scientist for Mercury Data Science.

List of Abbreviations

ANN	Artificial neural networks
BOW	Bag-of-words
CA	Content analysis
CTM	Correlated topic model
df	Document frequency
DM	Data mining
DTM	Document-term matrix
HCA	Hierarchical cluster analysis
idf	Inverse document frequency
IoT	Internet of Things
KDD	Knowledge discovery in databases
KDT	Knowledge discovery in text
kMC	k-means clustering
kNN	k-nearest neighbors
LDA	Latent Dirichlet allocation
LSA	Latent semantic analysis
LSI	Latent semantic indexing
NB	Naive Bayes
NLP	Natural language processing
NN	Neural networks
OM	Opinion mining
pLSI	Probabilistic latent semantic indexing
RF	Random forest
SA	Sentiment analysis
sLDA	Supervised latent Dirichlet allocation
STM	Structural topic model
SVD	Singular value decomposition
SVM	Support vector machines
TA	Text analytics
TDM	Term-document matrix
tf	Term frequency
tfidf	Term frequency-inverse document frequency
TM	Text mining

List of Figures

Fig. 1.1	Text analytics timeline	3
Fig. 1.2	Frequency of text analytics articles by year.	4
Fig. 1.3	Word cloud of the titles and abstracts of articles on text analytics	5
Fig. 1.4	Article frequency by industry for top 25 industry classifications	6
Fig. 1.5	Word cloud of text analytics job titles	7
Fig. 1.6	Word cloud of the skills required for text analytics jobs	7
Fig. 1.7	Guide to the text analytics process and the book	9
Fig. 2.1	Content analysis framework.	17
Fig. 2.2	Features of deductive inference	18
Fig. 2.3	Manifest and latent variables	18
Fig. 2.4	Deductive and inductive coding approaches	21
Fig. 2.5	Four open coding strategies	22
Fig. 2.6	The three-step inductive inference process	22
Fig. 3.1	Four initial planning considerations.	28
Fig. 3.2	Text analytics planning tasks	31
Fig. 3.3	Three characteristics of good research problems.	31
Fig. 3.4	Quality dimensions of text data	34
Fig. 3.5	Simple random sampling	36
Fig. 3.6	Systematic sampling	37
Fig. 3.7	Stratified sampling	37
Fig. 3.8	Choosing the analysis method based on the focus of the analysis	38
Fig. 4.1	Hierarchy of terms and documents	46
Fig. 4.2	Example document collection	47
Fig. 4.3	The text data pre-processing process	48

Fig. 4.4	Tokenized example documents	49
Fig. 4.5	Cleansed and standardized document collection	51
Fig. 4.6	Documents after stop word removal	52
Fig. 4.7	Document 9 tokenized text before and after stemming	55
Fig. 4.8	Stemmed example document collection	55
Fig. 4.9	Document 9 before and after stemming and lemmatization	56
Fig. 4.10	Lemmatized example document collection	57
Fig. 5.1	Basic document-term and term-document matrix layouts.	64
Fig. 5.2	Heat map visualizing the term-document matrix.	66
Fig. 5.3	Document frequency weighting.	69
Fig. 5.4	Global frequency weighting.	70
Fig. 5.5	Inverse document frequency weighting	71
Fig. 6.1	Two-dimensional representation of the first five documents in term space for the terms <i>brown</i> and <i>dog</i>	78
Fig. 6.2	Three-dimensional representation of the ten documents in term space for the terms <i>brown</i> , <i>coat</i> and <i>favorite</i>	79
Fig. 6.3	SVD process in LSA, based on Martin and Berry (2007)	80
Fig. 6.4	Terms and documents in three-dimensional LSA vector space	84
Fig. 6.5	Terms and documents of a two-dimensional LSA solution across the first two dimensions	85
Fig. 6.6	Rotated plot of the query and Document 6 vectors in three-dimensional LSA vector space	89
Fig. 6.7	Scree plot showing variance explained by number of singular vectors	90
Fig. 7.1	Visualization of cluster analysis.	94
Fig. 7.2	Two-dimensional representation of terms in document space for Documents 3 and 6	95
Fig. 7.3	Three-dimensional representation of terms in document space for Documents 1, 3, and 7	96
Fig. 7.4	<i>fluffy</i> and <i>brown</i> in document space for Documents 3 and 6	96
Fig. 7.5	Dendrogram example with cluster groupings	98
Fig. 7.6	HCA algorithm.	99
Fig. 7.7	Single linkage document—HCA example.	100
Fig. 7.8	Complete linkage document HCA example	101
Fig. 7.9	Centroid linkage document—HCA example.	102
Fig. 7.10	Ward's method for linking documents—HCA example	102
Fig. 7.11	kMC algorithm.	104
Fig. 7.12	<i>k</i> -Means process example plot.	105
Fig. 7.13	<i>k</i> -Means initial cluster seed designation of <i>cat</i> , <i>coat</i> , and <i>hat</i>	106
Fig. 7.14	Cluster assignments based on cluster seeds for a three-cluster solution.	107

Fig. 7.15	First iteration cluster assignment and calculated centroids	107
Fig. 7.16	Ward's method hierarchical cluster analysis solutions for $k = 2, 3, 4, 5, 6$, and 7	110
Fig. 7.17	Scree plot total within cluster SSE for k Values $1-9$	111
Fig. 7.18	Silhouette plot for Ward's method hierarchical clustering analysis.	112
Fig. 8.1	LSA and topic models (Griffiths et al. 2007, p. 216).	118
Fig. 8.2	Plate representation of the random variables in the LDA model (Blei 2012, p. 23)	119
Fig. 8.3	Top ten terms per topic, four-topic model	120
Fig. 8.4	Plate representation of the random variables in the CTM model (Blei et al. 2007, p. 21)	121
Fig. 8.5	Expected topic proportions of four categories in the CTM model with no covariates	121
Fig. 8.6	CTM topic correlation plot	122
Fig. 8.7	Plate diagram of DT model (Blei and Lafferty 2006, p. 2)	123
Fig. 8.8	Plate representation of the sLDA model (McAuliffe and Blei 2008, p. 3).	123
Fig. 8.9	Plate diagram representation of the structural topic model (Roberts et al. 2013, p. 2).	125
Fig. 8.10	Top ten terms in topics for STM model.	125
Fig. 8.11	Dog type content across topics	126
Fig. 8.12	Four measures across a number of topics, k , for $2-30$ LDA topics	127
Fig. 8.13	Topic frequency and the five most probable terms per topic	128
Fig. 9.1	Classification analysis process.	133
Fig. 9.2	Sample contingency table with two classifications, Yes and No . .	133
Fig. 9.3	Contingency table example	134
Fig. 9.4	Two-dimensional representation of support vector machine classification of ten documents (Sebastiani 2002).	141
Fig. 9.5	Splitting criteria for decision trees.	142
Fig. 9.6	Decision tree created from training data using deviance as the splitting criteria	143
Fig. 9.7	Random forest plot of the importance of variables	145
Fig. 9.8	Neural network example with one hidden layer and three classes	146
Fig. 10.1	Levels of sentiment Analysis	152
Fig. 10.2	Sample of positive and negative words that coincide and are consistent across the four lexicons	154
Fig. 10.3	Positive review example: text preparation and preprocessing . . .	155
Fig. 10.4	Negative review example: text preparation and preprocessing . .	156
Fig. 10.5	Ambiguous review example: text preparation and preprocessing .	157
Fig. 10.6	Word clouds of positive and negative words in review sample . .	158

Fig. 10.7	Examples of Accurate and Inaccurate Predictions using NB	159
Fig. 10.8	Examples of accurate and inaccurate predictions using SVM . . .	160
Fig. 10.9	Examples of accurate and inaccurate predictions using logistic regression.	162
Fig. 11.1	Questions to ask to identify the key components of the analysis.	168
Fig. 11.2	Storytelling framework.	171
Fig. 12.1	Strategies for text analytics visualizations	179
Fig. 12.2	Heat map visualization	181
Fig. 12.3	Word cloud of dog descriptions in the shape of a paw	182
Fig. 12.4	Plots of top terms in first four LSA dimensions	183
Fig. 12.5	Two dendrogram versions of the same HCA solution	184
Fig. 12.6	Top ten terms per topic, four-topic model	185
Fig. 12.7	Plot of expected topic proportions over time	186
Fig. 12.8	Two versions of the same document network plot	187
Fig. 12.9	Word clouds of positive and negative words in review sample . .	188
Fig. 12.10	Five-star word cloud of reviews	188
Fig. 13.1	RStudio IDE	194
Fig. 13.2	RStudio workspace with imdb dataframe in the global environment	195
Fig. 13.3	Table formatted data using the view function.	196
Fig. 13.4	Frequency of negative (0) and positive (1) reviews	196
Fig. 13.5	Diagram of an inner join between stop words and sentiments . . .	201
Fig. 13.6	Number of classifications per matching stop word and lexicon lists	201
Fig. 13.7	Diagram of a left join between imdb and the aggregated lexicons	204
Fig. 14.1	Python IDLE Shell file	223
Fig. 14.2	Python IDLE script and .py file	224
Fig. 14.3	NLTK downloader	226
Fig. 14.4	First two latent factors in 25-factor LSA solution	234
Fig. 14.5	Variance explained for increasing k values.	235
Fig. 14.6	First two latent factors in 25-factor LSA solution with <i>tfidf</i> weighting.	237
Fig. 14.7	Scree plot for up to 25 latent factors	239
Fig. 14.8	Top 10 terms by weight for dimension 1	240
Fig. 14.9	Top 10 terms in first four LSA dimensions	242
Fig. 15.1	RapidMiner welcome screen	244
Fig. 15.2	Main RapidMiner view.	245
Fig. 15.3	RapidMiner view with operations, process, and parameters panels	245
Fig. 15.4	Marketplace extension drop-down menu	246

Fig. 15.5	Text Processing extension	246
Fig. 15.6	Read CSV node	247
Fig. 15.7	Data import wizard	248
Fig. 15.8	Tab delimited text	248
Fig. 15.9	Row name designations	249
Fig. 15.10	Data type designations	249
Fig. 15.11	Read CSV connection to Results port	250
Fig. 15.12	Read CSV results	250
Fig. 15.13	Process Documents from Data operator	251
Fig. 15.14	Process Documents from Data operator warning	251
Fig. 15.15	Tokenize operator	252
Fig. 15.16	Tokenize operator results	252
Fig. 15.17	Transform Cases operator	253
Fig. 15.18	Filter Stop words (English) operator	253
Fig. 15.19	Filter Tokens (by Length) operator	254
Fig. 15.20	Stem (Porter) operator	254
Fig. 15.21	Process Documents operator results	255
Fig. 15.22	Term and document frequency node connections	255
Fig. 15.23	Term and document frequency results	256
Fig. 15.24	Term occurrences sorted in descending order	256
Fig. 15.25	Document Occurrences sorted in descending order	257
Fig. 15.26	Validation operator	257
Fig. 15.27	Validation node view drop-down	258
Fig. 15.28	kNN operator	258
Fig. 15.29	Apply model and performance operators in validation view	259
Fig. 15.30	kNN results contingency table	259
Fig. 15.31	Remove kNN operator	260
Fig. 15.32	Naïve Bayes operator	260
Fig. 15.33	Naïve Bayes results contingency table	260
Fig. 16.1	SAS Viya Welcome page	265
Fig. 16.2	Data import option for data management	265
Fig. 16.3	The <i>Available</i> option shows the dataset loaded	266
Fig. 16.4	First 13 rows in the dataset	266
Fig. 16.5	Assigning the text variable role	267
Fig. 16.6	Customizing the pipeline	267
Fig. 16.7	Default pipeline available on VTA	268
Fig. 16.8	Predefined concepts available	269
Fig. 16.9	Options available for text parsing	269
Fig. 16.10	Topic node options	270
Fig. 16.11	Run pipeline icon to implement the analysis	271
Fig. 16.12	The pipeline tasks completed	272
Fig. 16.13	Predefined concepts <i>nlpPerson</i>	272
Fig. 16.14	List of kept and dropped terms	273
Fig. 16.15	Results of topic analysis	273

Fig. 16.16	Most relevant terms associated with topics and their use in documents.	274
Fig. 16.17	Category definition: code and validation	275
Fig. 16.18	Category rule validated over a sample document.	275
Fig. 16.19	Example of matched documents in the disease/condition category	276
Fig. 16.20	<i>Explore and Visualize Data</i> selection for advanced analysis: main menu	277
Fig. 16.21	<i>Add Data Source</i> selection	278
Fig. 16.22	Tile by category and color by frequency percent selected under <i>Roles</i>	278
Fig. 16.23	Category popularity visualization Treemap	279
Fig. 16.24	Category popularity visualization Treemap after removing missing values.	279
Fig. 16.25	Category popularity visualization: line charts	279
Fig. 16.26	Category popularity visualization: pie charts	280
Fig. 16.27	Word by keywords and color by frequency percent selected under <i>Roles</i>	280
Fig. 16.28	Word cloud of key terms in the dataset	281
Fig. 16.29	Visualizations can be customized through the <i>Options</i> tag	281

List of Tables

Table 4.1	Document 9 related words, POS, and lemmatization for the word fluffy	56
Table 4.2	Document 9 related words, POS, and lemmatization for the word favorite	57
Table 5.1	Unprocessed and preprocessed text.	62
Table 5.2	Inverted index for document collection.	62
Table 5.3	Document frequency of the term <i>brown</i>	63
Table 5.4	Term-postings frequency table for the term <i>brown</i>	63
Table 5.5	Term-document matrix example	65
Table 5.6	Log frequency matrix	67
Table 5.7	Binary frequency matrix	68
Table 5.8	<i>tfidf</i> -weighted TDM.	72
Table 6.1	The LSA space	83
Table 6.2	Cosine similarity measures for <i>fluffy</i> , in descending order	86
Table 6.3	Term-term cosine similarity measures.	87
Table 6.4	Cosine values between the query (<i>brown, pink, tan</i>) and documents in descending order by cosine similarity value .	88
Table 7.1	Distance matrix of terms	97
Table 7.2	Distance matrix of documents.	97
Table 7.3	<i>Tfidf</i> term values for Documents 3 and 6.	105
Table 7.4	Squared distance from cluster seeds	106
Table 7.5	Squared distance from terms to cluster centroids	108
Table 9.1	Naïve Bayes contingency table	138
Table 9.2	Goodness of fit measures, naïve Bayes model	138
Table 9.3	1NN testing document actual classifications, 1NN documents and 1NN predicted classifications.	139
Table 9.4	Contingency table, kNN classification, $k = 1$ (1NN).	139

Table 9.5	Goodness of fit measures, k-nearest neighbors, $k = 1$	140
Table 9.6	Support vector machines contingency table	141
Table 9.7	Goodness of fit measures, SVM	142
Table 9.8	Decision tree confusion matrix	143
Table 9.9	Goodness of fit measures, decision tree	143
Table 9.10	Random forest contingency table	144
Table 9.11	Goodness of fit measures, random forest	144
Table 9.12	Neural network contingency matrix with five hidden nodes in one hidden layer	146
Table 9.13	Goodness of fit measures, neural network with five hidden nodes in one hidden layer	147
Table 9.14	Classification model accuracy	147
Table 10.1	Positive review word-, sentence-, and document-level sentiment	155
Table 10.2	Negative review word- and document-level sentiment	156
Table 10.3	Ambiguous review word and document-level sentiment	158
Table 10.4	Naïve Bayes contingency matrix classification analysis	159
Table 10.5	SVM contingency matrix classification analysis	160
Table 10.6	Logistic regression contingency matrix classification analysis . .	161
Table 10.7	Examples of consistent accurate and inaccurate predictions across learning methods for negative and positive sentiments	163
Table 13.1	Misclassified sentiment reviews with actual and predicted sentiment	206
Table 13.2	Reviews with <i>IO</i> , including review number, text, and sentiment labels	211
Table 13.3	Reviews with <i>time</i> , including review number, text, and sentiment labels	211

Chapter 1

Introduction to Text Analytics



Abstract In this chapter we define text analytics, discuss its origins, cover its current usage, and show its value to businesses. The chapter describes examples of current text analytics uses to demonstrate the wide array of real-world impacts. Finally, we present a process road map as a guide to text analytics and to the book.

Keywords Text analytics · Text mining · Data mining · Content analysis

1.1 Introduction

Recent estimates maintain that 80% of all data is text data. A recent article in *USA Today* asserts that even the Internal Revenue Service is using text, in the form of US citizens' social media posts, to help them make auditing decisions.¹ The importance of text data has created a veritable industry comprised of firms dedicated solely to the storage, analysis, and extraction of text data. One such company, Crimson Hexagon, has created the world's largest database of text data from social media sites including one trillion public social media posts spanning more than a decade.²

1.2 Text Analytics: What Is It?

Hearst (1999a, b) defines text analytics, sometimes referred to as text mining or text data mining, as the automatic discovery of new, previously unknown, information from unstructured textual data. The terms text analytics and text mining are often used interchangeably. Text mining can also be described as the process of deriving

¹Agency breaking law by mining social media. (2017, 12). *USA Today*, 146, 14–15.

²<http://www.businessinsider.com/analytics-firm-crimson-hexagon-uses-social-media-to-predict-stock-movements-2017-4>

high-quality information from text. This process involves three major tasks: information retrieval (gathering the relevant documents), information extraction (unearthing information of interest from these documents), and data mining (discovering new associations among the extracted pieces of information).

Text analytics has been influenced by many fields and has made significant contributions to many disciplines. Modern text analytics applications span many disciplines and objectives. In addition to having multidisciplinary origins, text analytics continues to have applications in and make advancements to many fields of study. It intersects many research fields, including:

- Library and information science
- Social sciences
- Computer science
- Databases
- Data mining
- Statistics
- Artificial intelligence
- Computational linguistics

Although many of these areas contributed to modern day text analytics, Hearst (1999a, b) posited that text mining came to fruition as an extension of data mining. The similarities and differences between text mining and data mining have been widely discussed and debated (Gupta and Lehal 2009; Hearst 2003). Data mining uses structured data, typically from databases, to uncover “patterns, associations, changes, anomalies, and significant structures” (Bose 2009, p. 156). The major difference between the two is the types of data that they use for analysis. Data mining uses structured data, found in most business databases, while text mining uses unstructured or semi-structured data from a variety of sources, including media, the web, and other electronic data sources. The two methods are similar because they (i) are equipped for handling large data sets; (ii) look for patterns, insight, and discovery; and (iii) apply similar or the same techniques. Additionally, text mining draws on techniques used in data mining for the analysis of the numeric representation of text data.

The complexities associated with the collection, preparation, and analysis of unstructured text data make text analytics a unique area of research and application. Unstructured data are particularly difficult for computers to process. The data itself cover a wide range of possibilities, each with its own challenges. Some examples of the sources of text data used in text mining are blogs, web pages, emails, social media, message board posts, newspaper articles, journal articles, survey text, interview transcripts, resumes, corporate reports and letters, insurance claims, customer complaint letters, patents, recorded phone calls, contracts, and technical documentation (Bose 2009; Dörre et al. 1999).

1.3 Origins and Timeline of Text Analytics

Text-based analysis has its roots in the fields of computer science and the social sciences as a means of converting qualitative data into quantitative data for analysis. The field of computer science is in large part responsible for the text analytics that we know today. In contrast, the social sciences built the foundation of the analysis of text as a means of understanding literature, discourse, documents, and surveys. Text analytics combines the computational and humanistic elements of both fields and uses technology to analyze unstructured data text data by “turning text into numbers.” The text analytics process includes the structuring of input text, deriving patterns within the structured data, and evaluating and interpreting the output.

Figure 1.1 presents a timeline of text analytics by decade. In the 1960s, computational linguistics was developed to describe computer-aided natural language processing (Miner et al. 2012). Natural language processing techniques are outlined in Chaps. 5 and 6. During this decade, content analysis, the focus of Chap. 2, emerged in the social sciences as a means of analyzing a variety of content, including text and media (Krippendorff 2012).

In the late 1980s and early 1990s, latent semantic indexing, or latent semantic analysis, introduced in Chap. 6 arrived as a dimension reduction and latent factor identification method applied to text (Deerwester et al. 1990; Dumais et al. 1988). At this time, knowledge discovery in databases developed as a means of making sense of data (Fayyad et al. 1996; Frawley et al. 1992). Building on this advancement, Feldman and Dagan (1995) created a framework for text, known as knowledge discovery in texts to do the same with unstructured text data.

Data mining emerged in the 1990s as the analysis step in the knowledge discovery in databases process (Fayyad et al. 1996). In the 1990s, machine learning methods, covered in Chaps. 7 and 9, gained prominence in the analysis of text data (Sebastiani 2002). Around that time, text mining became a popular buzzword but lacked practitioners (Hearst 1999a, b). Nagarkar and Kumbhar (2015) reviewed text mining-related publications and citations from 1999 to 2013 and found that the number of publications consistently increased throughout this period.

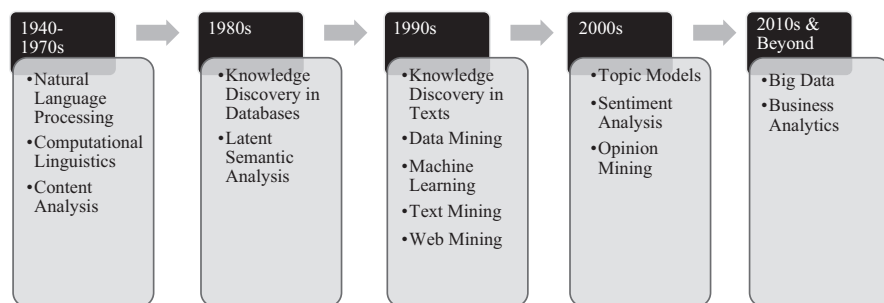


Fig. 1.1 Text analytics timeline

Building on a probabilistic form of latent semantic analysis (Hofmann 1999) introduced the late 1990s, topic models, discussed in Chap. 8, were created in the early 2000s with the development of the latent Dirichlet allocation model (Blei et al. 2002, 2003). Around the same time, sentiment analysis (Nasukawa and Yi 2003) and opinion mining (Dave et al. 2003), the focus of Chap. 10, were introduced as methods to understand and analyze opinions and feelings (Liu 2012).

The 2010s were the age of big data analytics. During this period, the foundational concepts preceding this time were adapted and applied to big data. According to Davenport (2013), although the field of business analytics has been around for over 50 years, the current era of analytics, Analytics 3.0, has witnessed the widespread use of corporate data for decision-making across many organizations and industries. More specifically, four key features define the current generation of text analytics and text mining: foundation, speed, logic, and output. As these characteristics indicate, text analysis and mining are data driven, conducted in real time, rely on probabilistic inference and models, and provide interpretable output and visualization (Müller et al. 2016). According to *IBM Tech Trends Report* (2011), business analytics was deemed one of the major trends in technology in the 2010s (Chen et al. 2012).

One way to better understand the area of text analytics is by examining relevant journal articles. We analyzed 3,264 articles, 2,315 published in scholarly journals and 949 published in trade journals. In our article sample, there are 704 journals, 187 trade journals and 517 scholarly journals. Each article belongs to one or more article classifications. The articles in our sample have 170 distinct classifications based on information about them such as industry, geographical location, content, and article type. Figure 1.2 displays the total number of text analytics-related

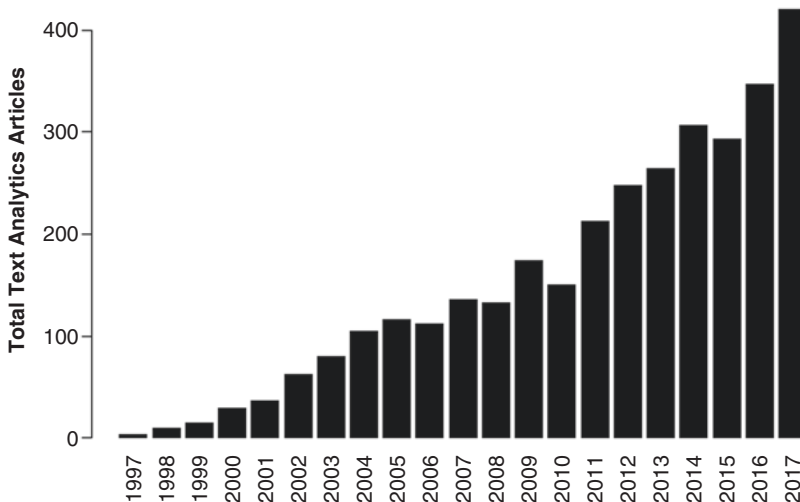


Fig. 1.2 Frequency of text analytics articles by year

articles over time. As the figure illustrates, there has been a considerable growth in the publications since 1997, with 420 articles about text mining published in 2017.

1.4 Text Analytics in Business and Industry

Text analytics has numerous practical uses, including but not limited to email filtering, product suggestions, fraud detection, opinion mining, trend analysis, search engines, and bankruptcy predictions (Talib et al. 2016). The field has a wide range of goals and objectives, including understanding semantic information, text summarization, classification, and clustering (Bolasco et al. 2005). We explore some examples of text analytics and text mining in the areas of business and industry. Text analytics applications require clear, interpretable results and actionable outcomes to achieve the desired result. Indeed, the technique can be used in almost every business department to increase productivity, efficiency, and understanding.

Figure 1.3 is a word cloud depicting the most popular terms and phrases in the text analytics-related articles' abstracts and titles. As the figure shows, research and applications in this area are as diverse as the area's history. The close relationship between data mining and text analytics is also evident in the figure.



Fig. 1.3 Word cloud of the titles and abstracts of articles on text analytics