

Manika Lamba  
Margam Madhusudhan

# Text Mining for Information Professionals

An Uncharted Territory

 Springer

# Text Mining for Information Professionals

Manika Lamba • Margam Madhusudhan

# Text Mining for Information Professionals

An Uncharted Territory

 Springer

Manika Lamba  
Library and Information Science  
University of Delhi  
Delhi, India

Margam Madhusudhan  
Library and Information Science  
University of Delhi  
Delhi, India

ISBN 978-3-030-85084-5      ISBN 978-3-030-85085-2 (eBook)  
<https://doi.org/10.1007/978-3-030-85085-2>

© The Editor(s) (if applicable) and The Author(s), under exclusive license to Springer Nature Switzerland AG 2022

This work is subject to copyright. All rights are solely and exclusively licensed 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, expressed 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 grandparents and brother for their  
endless support and constant encouragement  
– Manika*

# Preface

Machine learning and artificial intelligence are the futuristic approaches dominating all disciplines currently. Data analytics, data mining, and data science are some of the significant sub-domains with a strong market of research and jobs at the moment in the world. With a long leap from paper to digitization, the burden on libraries and librarians has increased to manage, organize, and generate knowledge from such a massive amount of data stored in their repositories/databases/websites. As libraries deal with a higher percentage of textual data daily, this book focuses primarily on textual data and presents various text mining approaches through a new lens. Text mining is a very efficient, fast, and effective way of managing and extracting knowledge from existing data stored in the archives of libraries. This book will make every library and information professional competent to use text mining in their daily life and get the best out of it by serving their patrons, researchers, faculty, or scientists with new services. Text mining techniques can be applied to any library type, be it a school, university, or special library by the librarians. It will help to provide the *right information* to the *right user* at the *right time* by providing services like recommendation services, current awareness services, or selective dissemination services to its users.

From understanding different types and forms of data to case studies covering primary research showing the application of each text mining approach on data retrieved from various resources, this book will be a must-read for all library professionals interested in text mining and its application in libraries. Additionally, this book will be helpful to archivists, digital curators, or any other humanities and social science professionals who want to understand the basic theory behind text data, text mining, and various tools and techniques available to solve and visualize their research problems.

## Key points of the book

1. Contains 14 demonstrative step-by-step case studies which show how to conduct 8 different text mining and visualization approaches on 9 distinct data type sources
2. Provides case studies demonstrating the use of five open-source software for both non-programmers and programmers

3. Reproduces six case studies using R programming in the cloud without having to install any software
4. Story section presenting 17 real-life experiences of the application of text mining methods and tools by 24 librarians/researchers/faculty/publishers
5. Elucidates 19 open-source text mining and visualization tools with their advantages and disadvantages
6. Illustrates various use cases that show how text mining strategies have been used in different ways in libraries across the globe

The book contains *11 chapters*, *14 case studies* showing 8 different text mining and visualization approaches, and *17 stories*. A website (<https://textmining-infopros.github.io/>) and a GitHub account (<https://github.com/textmining-infopros>) are also maintained for the book. They contain the code, data, and notebooks for the case studies; a summary of all the stories shared by the librarians/faculty; and hyperlinks to open an interactive virtual RStudio/Jupyter Notebook environment. The interactive virtual environment runs case studies based on the R programming language for hands-on practice in the cloud without installing any software. Text mining is a topic of international interest, and this book has been written to meet the reading interests of both national and international audiences. It will be appropriate for both beginners and advanced-level readers as it has been written keeping their information needs in mind.

Many books in the market are written to meet the need of computer science professionals on text mining, whereas there are very few books on text mining for librarians. Also, the books present in the market on this topic are very difficult for non-programmers to understand. They may contain lots of jargon, which may not be easily understood by a library professional. In contrast, this book focuses on a basic theoretical framework dealing with the problems, solutions, and applications of text mining and its various facets in the form of *case studies*, *use cases*, and *stories*.

Delhi, India  
Delhi, India  
June 2021

Manika Lamba  
Margam Madhusudhan

# Acknowledgments

The authors are immensely grateful to Alex Wermer-Colan, Amy J. Kirchhoff, Anne M. Brown, C. Cozette Comer, Carady DeSimone, Chreston Miller, Cody Hennesy, Issac Williams, Jacob Lahne, Jonathan Briganti, Jordan Patterson, Karen Harker, Leah Hamilton, Lighton Phiri, Manisha Bolina, Manoj Kumar Verma, Marcella Fredriksson, Maya Deori, Michael J. Stamper, Nathaniel D. Porter, Parthasarathi Mukhopadhyay, Rachel Miles, Sephra Byrne, and Vinit Kumar for sharing their personal experience of using different text mining approaches for the *story section* of the book.



# Contents

<b>1</b>	<b>The Computational Library</b> .....	1
1.1	Computational Thinking .....	1
1.2	Genealogy of Text Mining in Libraries .....	6
1.3	What Is Text Mining? .....	8
1.3.1	Text Characteristics .....	10
1.3.2	Different Text Mining Tasks .....	12
1.3.3	Supervised vs. Unsupervised Learning Methods .....	15
1.3.4	Cost, Benefits, and Barriers .....	17
1.3.5	Limitations .....	17
1.4	Case Study: Clustering of Documents Using Two Different Tools .....	17
	References .....	30
<b>2</b>	<b>Text Data and Where to Find Them?</b> .....	33
2.1	Data .....	33
2.1.1	Digital Trace Data .....	34
2.2	Different Types of Data .....	38
2.3	Data File Types .....	39
2.3.1	Plain Text .....	39
2.3.2	CSV .....	41
2.3.3	JSON .....	42
2.3.4	XML .....	45
2.3.5	Binary Files .....	46
2.4	Metadata .....	48
2.4.1	What Is a Metadata Standard? .....	49
2.4.2	Steps to Create Quality Metadata .....	50
2.5	Digital Data Creation .....	51
2.6	Different Ways of Getting Data .....	54
2.6.1	Downloading Digital Data .....	56
2.6.2	Downloading Data from Online Repositories .....	56
2.6.3	Downloading Data from Relational Databases .....	56

2.6.4	Web APIs .....	63
2.6.5	Web Scraping/Screen Scraping .....	66
References	.....	77
<b>3</b>	<b>Text Pre-Processing</b> .....	79
3.1	Introduction .....	79
3.1.1	Level of Text Representation .....	81
3.2	Text Transformation .....	81
3.2.1	Corpus Creation .....	81
3.2.2	Dictionary Creation .....	82
3.3	Text Pre-Processing .....	82
3.3.1	Case Normalization .....	82
3.3.2	Morphological Normalization .....	83
3.3.3	Tokenization .....	83
3.3.4	Stemming .....	84
3.3.5	Lemmatization .....	84
3.3.6	Stopwords .....	85
3.3.7	Object Standardization .....	85
3.4	Feature Engineering .....	86
3.4.1	Semantic Parsing .....	86
3.4.2	Bag of Words (BOW) .....	86
3.4.3	N-Grams .....	87
3.4.4	Creation of Matrix .....	88
3.4.5	Term Frequency-Inverse Document Frequency (TF-IDF) .....	89
3.4.6	Syntactical Parsing .....	90
3.4.7	Parts-of-Speech Tagging (POS) .....	91
3.4.8	Named Entity Recognition (NER) .....	93
3.4.9	Similarity Computation Using Distances .....	94
3.4.10	Word Embedding .....	95
3.5	Case Study: An Analysis of Tolkien's Books .....	96
References	.....	103
<b>4</b>	<b>Topic Modeling</b> .....	105
4.1	What Is Topic Modeling? .....	105
4.1.1	Topic Evolution .....	106
4.1.2	Application and Visualization .....	107
4.1.3	Available Tools and Packages .....	108
4.1.4	When to Use Topic Modeling .....	109
4.1.5	When <i>Not</i> to Use Topic Modeling .....	110
4.2	Methods and Algorithms .....	110
4.3	Topic Modeling and Libraries .....	113
4.3.1	Use Cases .....	117
4.4	Case Study: Topic Modeling of Documents Using Three Different Tools .....	119
References	.....	136

- 5 Network Text Analysis** ..... 139
  - 5.1 What Is Network Text Analysis? ..... 139
    - 5.1.1 Two-Mode Networks ..... 141
    - 5.1.2 Centrality Measures ..... 142
    - 5.1.3 Graph Algorithms ..... 145
    - 5.1.4 Comparison of Network Text Analysis with Others ..... 145
    - 5.1.5 How to Perform Network Text Analysis? ..... 146
    - 5.1.6 Available Tools and Packages ..... 147
    - 5.1.7 Applications ..... 147
    - 5.1.8 Advantages ..... 148
    - 5.1.9 Limitations ..... 149
  - 5.2 Topic Maps ..... 149
    - 5.2.1 Constructs of Topic Maps ..... 150
    - 5.2.2 Topic Map Software Architecture ..... 151
    - 5.2.3 Typical Uses ..... 152
    - 5.2.4 Advantages of Topic Maps ..... 152
    - 5.2.5 Disadvantages of Topic Maps ..... 153
  - 5.3 Network Text Analysis and Libraries ..... 153
    - 5.3.1 Use Cases ..... 156
  - 5.4 Case Study: Network Text Analysis of Documents Using  
Two Different R Packages ..... 158
  - References ..... 171
- 6 Burst Detection** ..... 173
  - 6.1 What Is Burst Detection? ..... 173
    - 6.1.1 How to Detect a Burst? ..... 174
    - 6.1.2 Comparison of Burst Detection with Others ..... 175
    - 6.1.3 How to Perform Burst Detection? ..... 176
    - 6.1.4 Available Tools and Packages ..... 177
    - 6.1.5 Applications ..... 178
    - 6.1.6 Advantages ..... 178
    - 6.1.7 Limitations ..... 178
  - 6.2 Burst Detection and Libraries ..... 179
    - 6.2.1 Use Cases ..... 179
    - 6.2.2 Marketing ..... 180
    - 6.2.3 Reference Desk Service ..... 180
  - 6.3 Case Study: Burst Detection of Documents Using Two  
Different Tools ..... 180
  - References ..... 188
- 7 Sentiment Analysis** ..... 191
  - 7.1 What Is Sentiment Analysis? ..... 191
    - 7.1.1 Levels of Granularity ..... 192
    - 7.1.2 Approaches for Sentiment Analysis ..... 193
    - 7.1.3 How to Perform Sentiment Analysis? ..... 194
    - 7.1.4 Available Tools and Packages ..... 195

7.1.5	Applications .....	196
7.1.6	Advantages .....	196
7.1.7	Limitations .....	196
7.2	Sentiment Analysis and Libraries .....	197
7.2.1	Use Cases .....	200
7.3	Case Study: Sentiment Analysis of Documents Using Two Different Tools .....	201
	References .....	210
<b>8</b>	<b>Predictive Modeling</b> .....	213
8.1	What Is Predictive Modeling? .....	213
8.1.1	Why Use Machine Learning? .....	215
8.1.2	Machine Learning Methods .....	215
8.1.3	Feature Selection and Representation .....	216
8.1.4	Machine Learning Algorithms .....	216
8.1.5	Classification Task .....	219
8.1.6	How to Perform Predictive Modeling on Text Documents? .....	221
8.1.7	Available Tools and Packages .....	227
8.1.8	Advantages .....	227
8.1.9	Limitations .....	228
8.2	Machine Learning and Libraries .....	228
8.2.1	Challenges .....	230
8.2.2	Use Cases .....	234
8.3	Case Study: Predictive Modeling of Documents Using RapidMiner .....	236
	References .....	240
<b>9</b>	<b>Information Visualization</b> .....	243
9.1	What Is Information Visualization? .....	243
9.1.1	Information Visualization Framework .....	244
9.1.2	Data Scale Types .....	245
9.1.3	Graphic Variable Types .....	246
9.1.4	Types of Datasets .....	247
9.1.5	Attribute Semantics .....	248
9.1.6	What Is an Appropriate Visual Representation for a Given Dataset? .....	248
9.1.7	Graphical Decoding .....	248
9.1.8	How Does One Know How Good a Visual Encoding Is? .....	249
9.1.9	Main Purpose of Visualization .....	249
9.1.10	Modes of Visualization .....	250
9.1.11	Methods of Graphic Visualization .....	250
9.2	Fundamental Graphs .....	251
9.3	Networks and Trees .....	254
9.4	Advanced Graphs .....	255

- 9.5 Rules on Visual Design ..... 261
- 9.6 Text Visualization ..... 262
- 9.7 Document Visualization..... 269
- 9.8 Information Visualization and Libraries ..... 270
  - 9.8.1 Use Cases ..... 282
  - 9.8.2 Information Visualization Skills for Librarians ..... 289
  - 9.8.3 Conclusion ..... 289
- 9.9 Case Study: To Build a Dashboard Using R ..... 290
- References ..... 292
  
- 10 Tools and Techniques for Text Mining and Visualization ..... 295**
  - 10.1 Introduction ..... 295
  - 10.2 Text Mining Tools ..... 296
    - 10.2.1 R ..... 296
    - 10.2.2 Topic-Modeling-Tool ..... 297
    - 10.2.3 RapidMiner ..... 299
    - 10.2.4 Waikato Environment for Knowledge Analysis (WEKA) ..... 301
    - 10.2.5 Orange ..... 302
    - 10.2.6 Voyant Tools ..... 304
    - 10.2.7 Science of Science (Sci2) Tool ..... 306
    - 10.2.8 LancsBox ..... 307
    - 10.2.9 ConText ..... 308
    - 10.2.10 Overview Docs ..... 309
  - 10.3 Visualization Tools ..... 310
    - 10.3.1 Gephi..... 310
    - 10.3.2 Tableau Public ..... 311
    - 10.3.3 Infogram ..... 312
    - 10.3.4 Microsoft Power BI ..... 312
    - 10.3.5 Datawrapper ..... 314
    - 10.3.6 RAWGraphs ..... 315
    - 10.3.7 WORDij ..... 315
    - 10.3.8 Palladio ..... 316
    - 10.3.9 Chart Studio ..... 317
  - References ..... 318
  
- 11 Text Data and Mining Ethics ..... 319**
  - 11.1 Text Data Management ..... 319
    - 11.1.1 Plan ..... 320
    - 11.1.2 Lifecycle ..... 320
    - 11.1.3 Citation ..... 325
    - 11.1.4 Sharing..... 326
    - 11.1.5 Need of Data Management for Text Mining ..... 326
    - 11.1.6 Benefits of Data Management for Text Mining ..... 327
    - 11.1.7 Ethical and Legal Rules Related to Text Data ..... 327

- 11.2 Social Media Ethics ..... 330
  - 11.2.1 Framework for Ethical Research with Social Media Data ..... 332
- 11.3 Ethical and Legal Issues Related to Text Mining ..... 332
  - 11.3.1 Copyright ..... 337
  - 11.3.2 License Conditions ..... 337
  - 11.3.3 Algorithmic Confounding/Biasness ..... 338
- References ..... 347
  
- Index** ..... 349