2023 ASIS&T Meet the Authors Webinar Series December 7, 2023

"Text Mining for Information Professionals: An Uncharted Territory"

Presenter: Dr. Manika Lamba

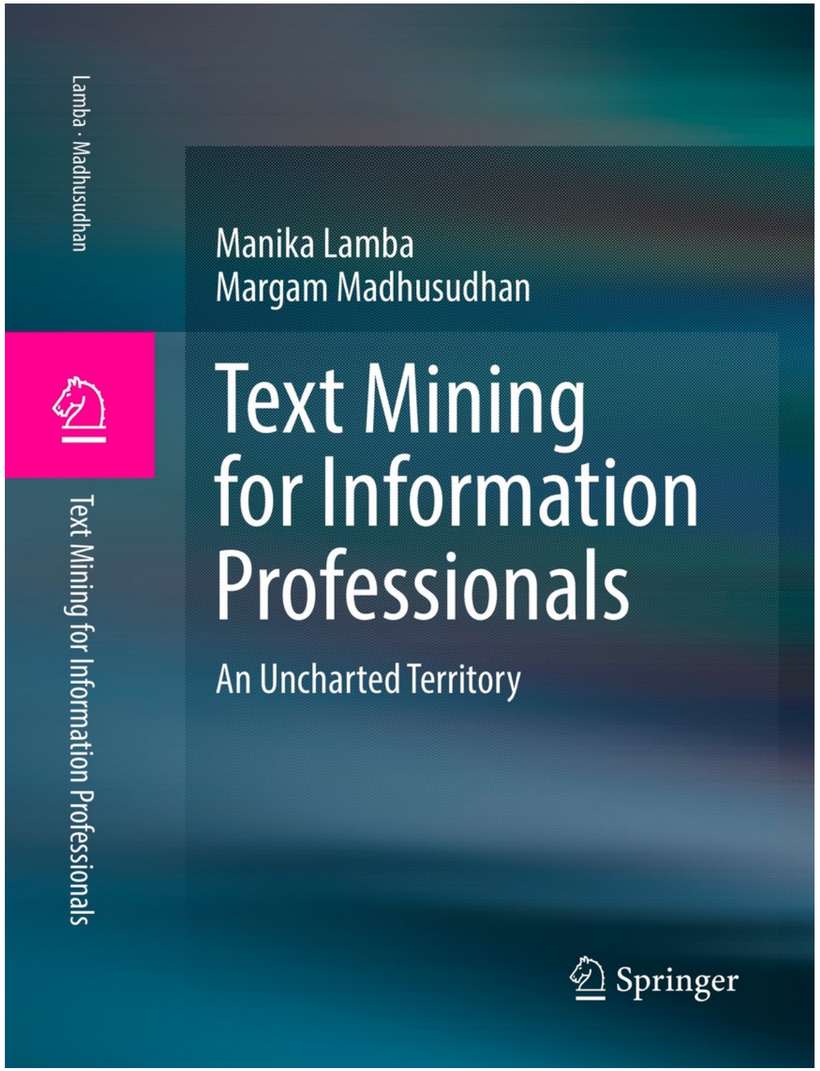
webinars@asist.org

asis&t

Association for Information Science and Technology

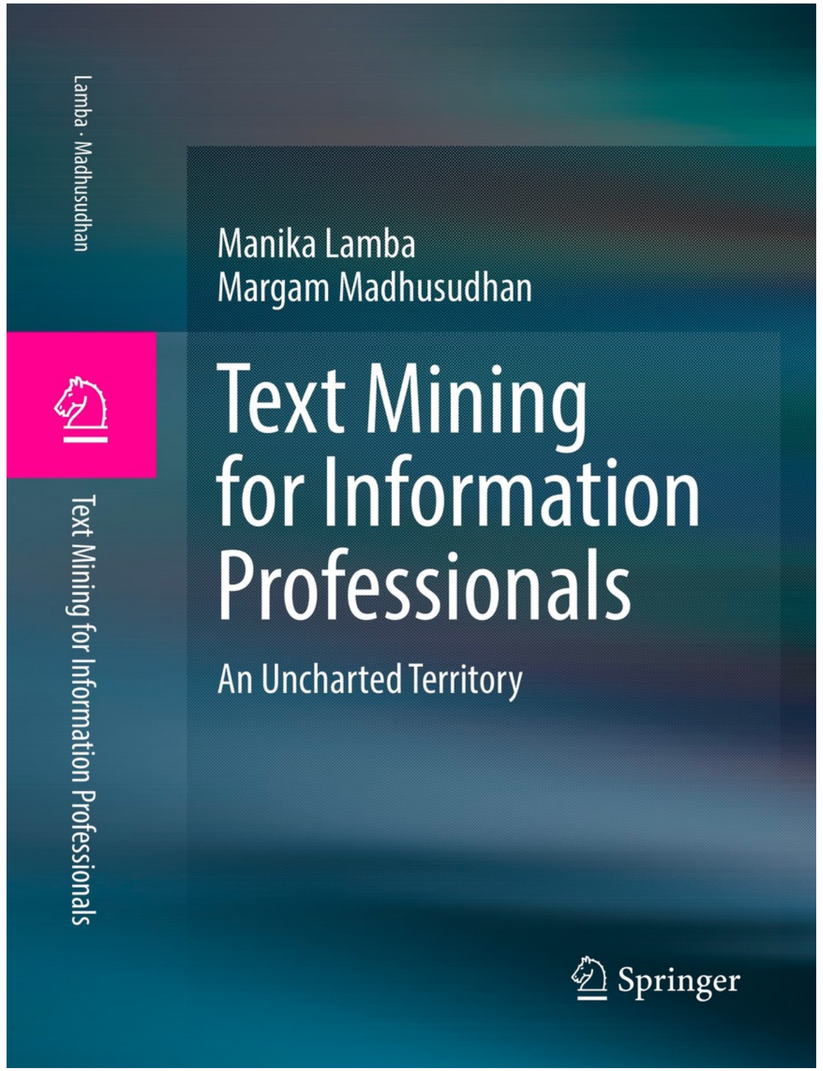
I am a Postdoctoral Research Associate at the School of Information Sciences, University of Illinois Urbana-Champaign. I have earned my Ph.D. in Library & Information Science from the University of Delhi, India. My research focuses on information organization, health & social informatics, and science of science using text mining, NLP, and machine learning techniques.

Dr. Margam Madhusudhan is a Professor in the Department of Library and Information Science, University of Delhi, India. He has worked as Deputy Dean Academics and Member of Academic Council at the University of Delhi. He has more than 20 years of teaching, administration, and research experience at the University level.



ABOUT THE BOOK

HTTPS://TEXTMINING-INFOPROS.GITHUB.IO/

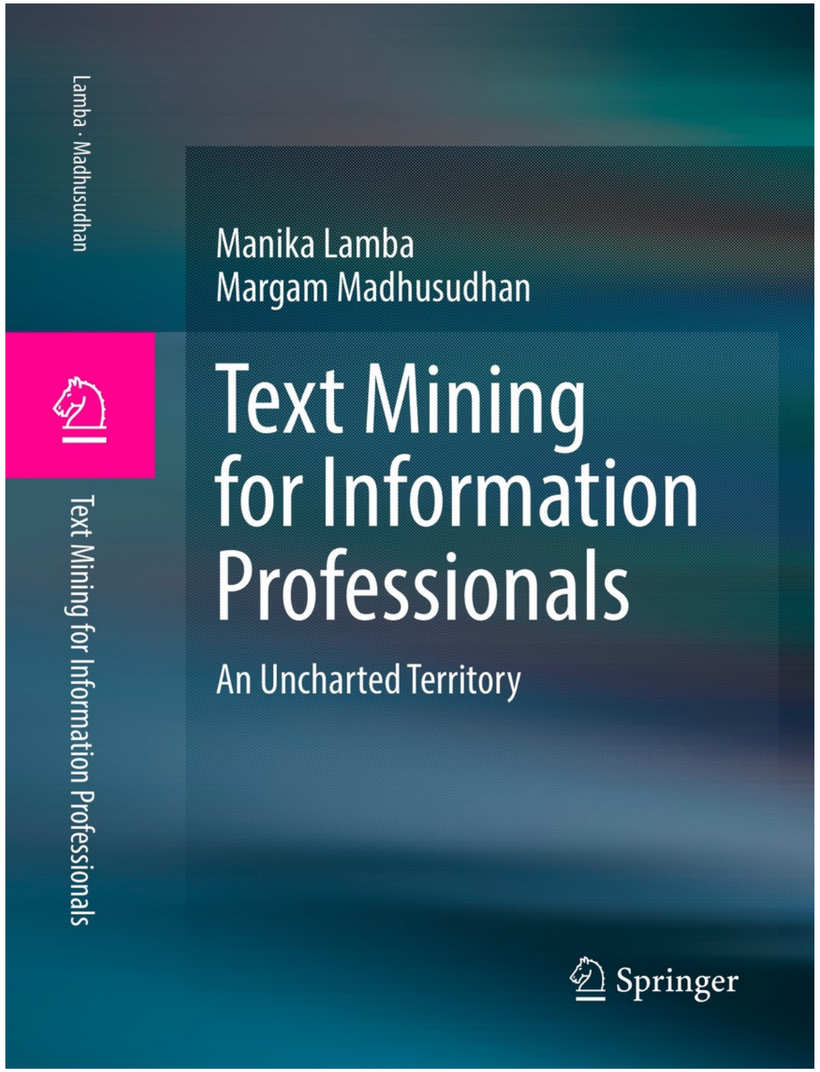


The book contains 11 chapters, 14 case studies showing 8 different text mining and visualization approaches, and 17 stories.

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 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.

Key points of the book

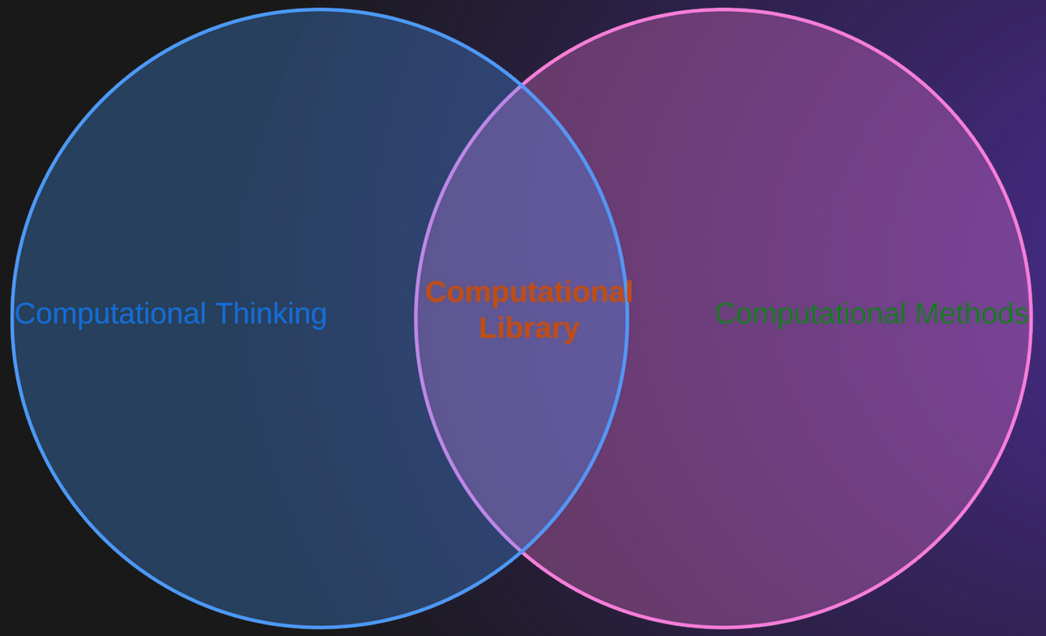
- Contains 14 demonstrative step-by-step case studies which show how to conduct 8 different text mining and visualization approaches on 9 distinct data type sources
- Provides case studies demonstrating the use of five open-source software for both non- programmers and programmers
- Reproduces six case studies using R programming in the cloud without having to install any software
- Story section presenting 17 real-life experiences of the application of text mining methods and tools by 24 librarians/researchers/faculty/publishers
- Elucidates 19 open-source text mining and visualization tools with their advantages and disadvantages
- Illustrates various use cases that show how text mining strategies have been used in different ways in libraries across the globe



Chapters

!! BONUS -- Curated Datasets: This repository contains some of the additional datasets which are in open-access and can be used to practice or teach text mining. The goal of this repository is to act as a collection of textual data set to be used for training and practice in text mining/NLP.

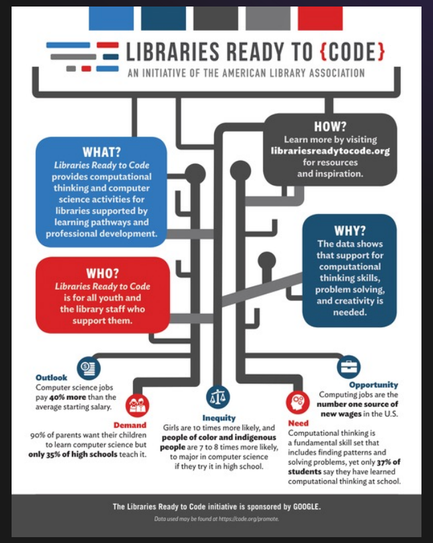
Computational Library



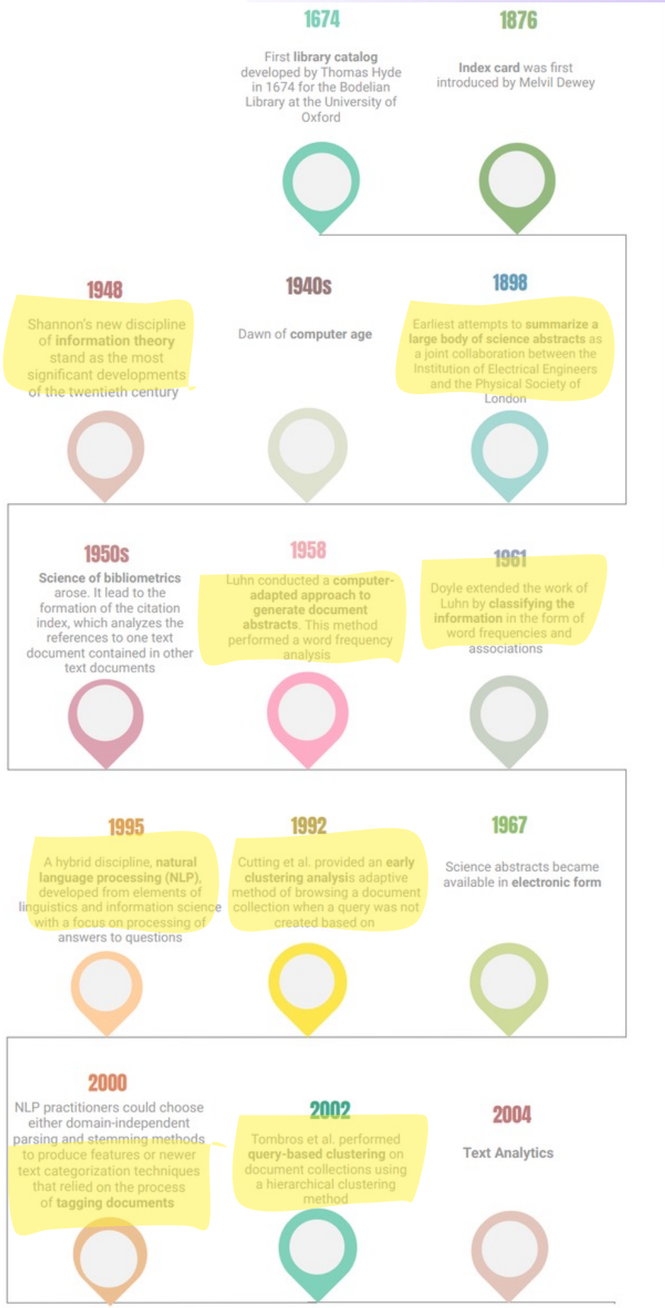
A computational library helps in making a library's collection accessible by applications, algorithms, machines, and people

Libraries have been involved in various CT literacies (like designing a website) and provide a programmable environment to their community by using computational techniques and methods

Examples: Libraries Ready to Code (RtC) by ALA
TIMDEX by MIT Libraries, 2019



Text Mining in Libraries



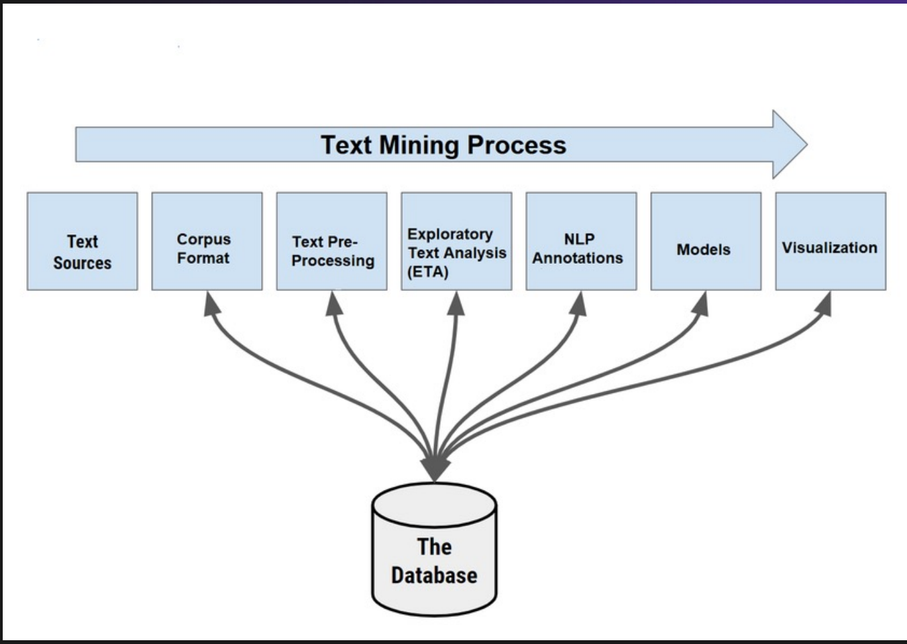
What is Text Mining?

Text mining is a process of automatically extracting information from the text with the aim of generating new knowledge

It is a specialized interdisciplinary field combining techniques from linguistics, computer science, and statistics to build tools that can efficiently retrieve and extract information from digital text

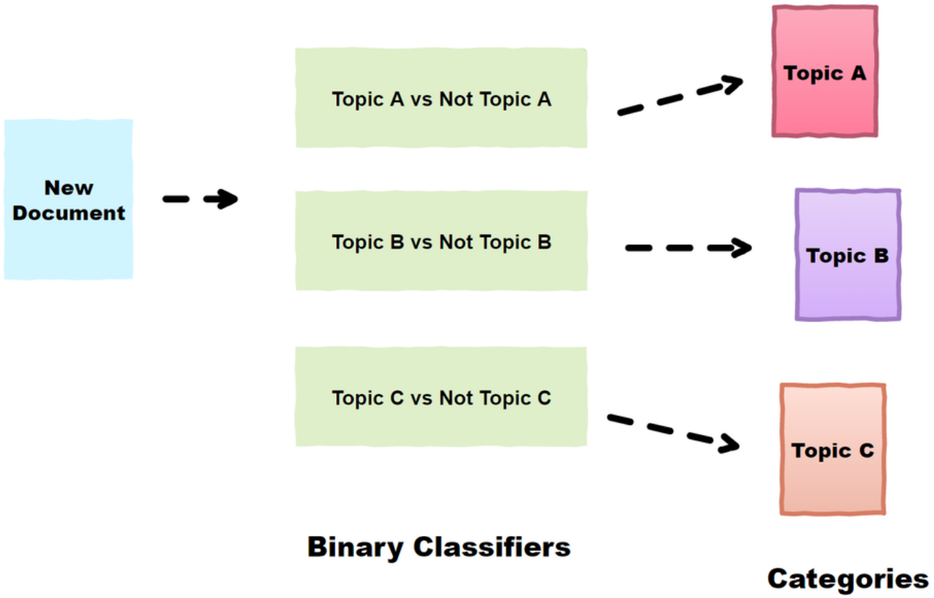
It assists in the automatic classification of documents

In text mining, “**words** are *attributes or predictors* and **documents** are *cases or records*, together these form a sample of data that can feed in well-known learning methods” (Weiss et al., 2005)

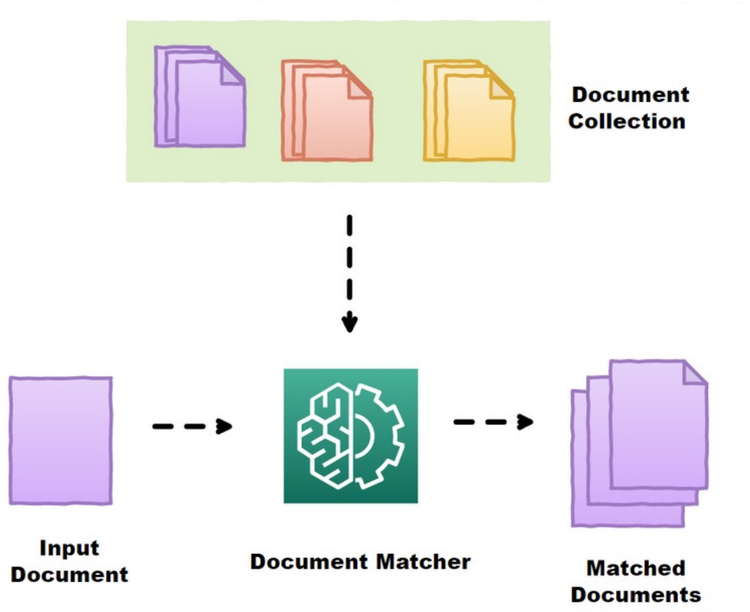


Weiss SM, Indurkhya N, Zhang T, Damerau FJ (2005) Overview of text mining. In: Weiss SM, Indurkhya N, Zhang T, Damerau FJ (eds) Text mining: Predictive methods for analyzing unstructured information. Springer, New York, NY, pp 1–13

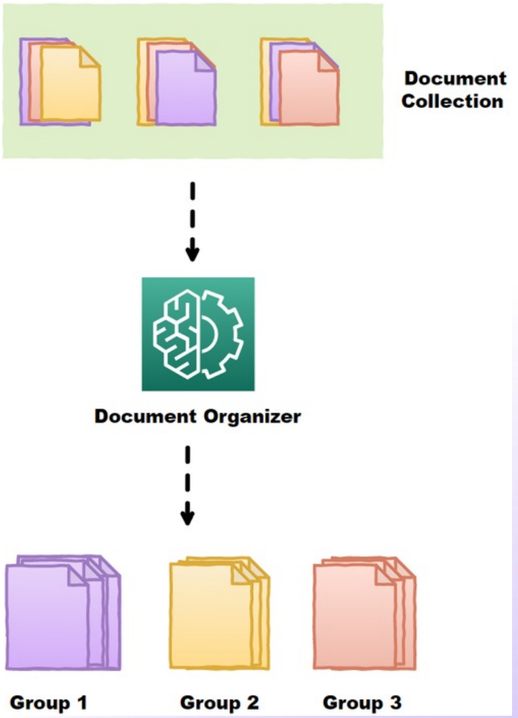
Different Text Mining Tasks



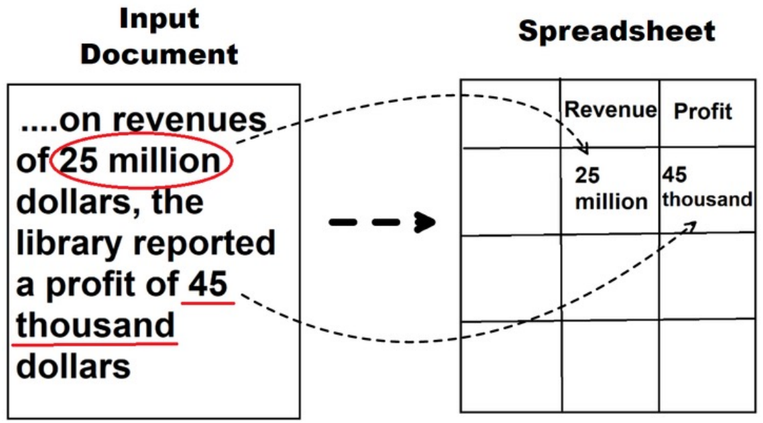
Document Classification/Text Categorization



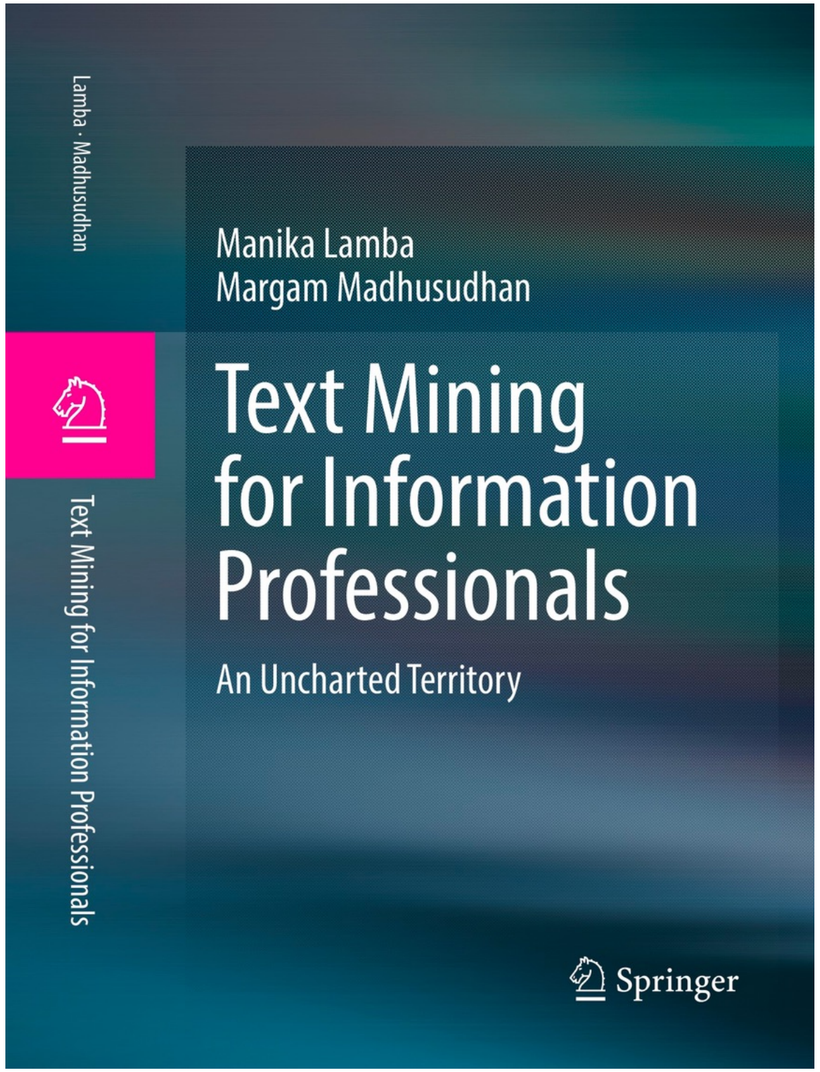
Information Retrieval



Clustering



Information Extraction



Chapters

- Chapter 1: The Computational Library
  - Case Study: Clustering of Documents using Two Different Tools DOI 10.5281/zenodo.5203303
- Chapter 2: Text Data and Where to Find Them?
- Chapter 3: Text Pre-Processing
  - Case Study: An Analysis of Tolkien's Books
- Chapter 4: Topic Modeling
  - Case Study: Topic Modeling of Documents using Three Different Tools DOI 10.5281/zenodo.5203494
- Chapter 5: Network Text Analysis
  - Case Study: Network Text Analysis of Documents using Two Different R Packages DOI 10.5281/zenodo.5203302
- Chapter 6: Burst Detection
  - Case Study: Burst Detection of Documents using Two Different Tools DOI 10.5281/zenodo.5203298
- Chapter 7: Sentiment Analysis
  - Case Study: Sentiment Analysis of Documents using Two Different Tools DOI 10.5281/zenodo.5203347
- Chapter 8: Predictive Modeling
  - Case Study: Predictive Modeling of Documents using RapidMiner DOI 10.5281/zenodo.5203567
- Chapter 9: Information Visualization
- Chapter 10: Tools and Techniques for Text Mining and Visualizations
- Chapter 11: Text Data and Mining Ethics
- Appendix A: Online Repositories Available for Text Mining DOI 10.5281/zenodo.5104488
- Appendix B: Language Corpora Available for Text Mining DOI 10.5281/zenodo.5104678
- Appendix C: Text Data and Mining Licensing Conditions DOI 10.5281/zenodo.5104740

!! BONUS -- Curated Datasets: This repository contains some of the additional datasets which are in open-access and can be used to practice or teach text mining. The goal of this repository is to act as a collection of textual data set to be used for training and practice in text mining/NLP.

Digital Trace Data

Past decade has witnessed an increasingly voluminous amount of digital data that is produced on the internet which describes human behavior and other objects of scholarly inquiry

Recent decades have not only witnessed an increase in the amount of text-based data but also increased computing power which is increasingly necessary to analyze it

Together, these two shifts hold the potential to significantly expand the scope of research in many different fields

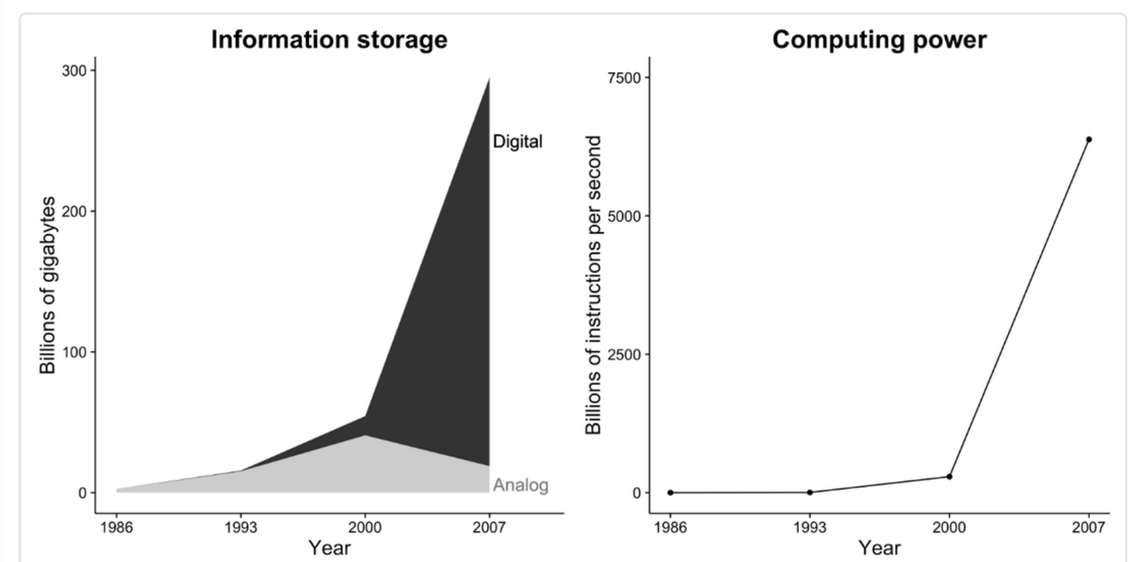


Figure 1.1: Information storage capacity and computing power are increasing dramatically. Further, information storage is now almost exclusively digital (Hilbert and López 2011). These changes create incredible opportunities for social researchers.

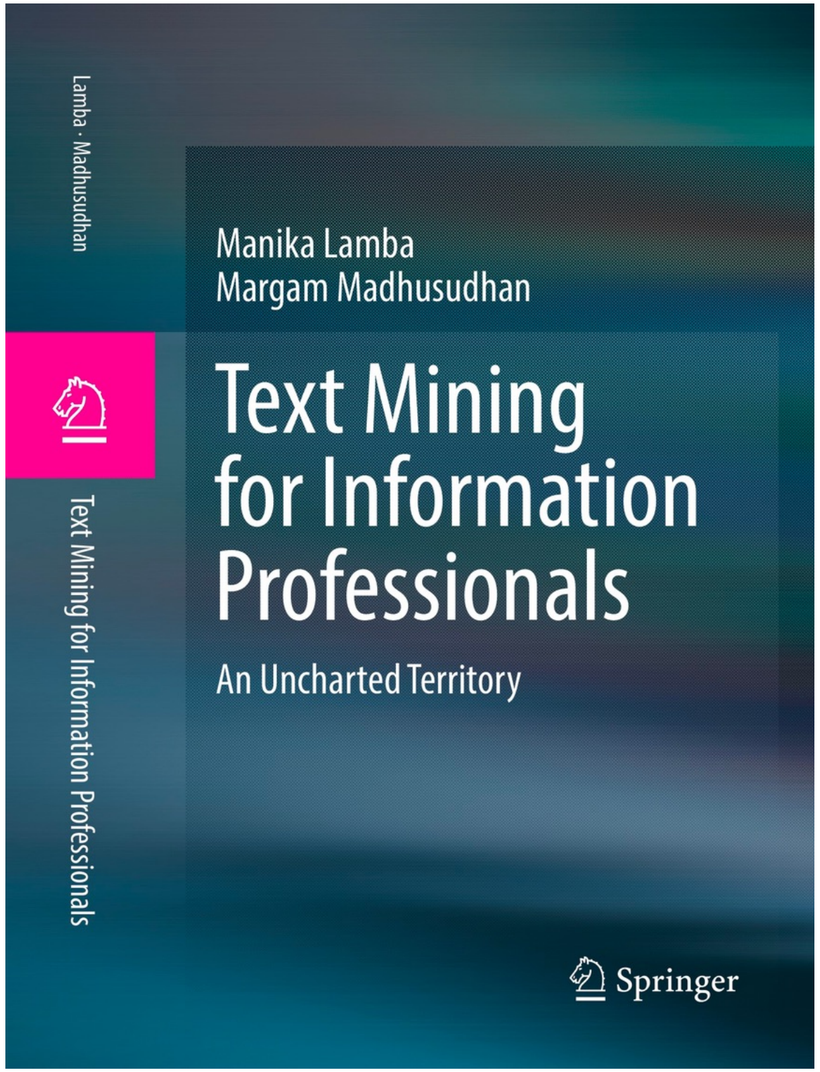
Where to Get Digital Data?

Table 2.6 Example of tools available for collecting digital trace data

Platform/tool	Description	Data	Access	Limit
Digital Scholar Workbench—Constellate https://constellate.org/	Provides access to CSV containing only metadata and JSON file containing metadata and textual data that includes unigrams, bigrams, trigrams, and full text (where available) for text analysis	Articles, books, chapters, and newspapers from JSTOR, Portico, Chronicling America, CORD-19, and DocSouth databases	Free	Up to 25,000 documents in a dataset where a single IP can create 10 datasets per day
TDM Studio https://tdmstudio.proquest.com/home	Provide access to ProQuest's collection across all disciplines and formats. Researchers can also incorporate their own datasets such as institutional repositories and journal articles and make use of additional content such as social media and blogs. The platform further allows researchers to perform text mining and visualization techniques	Current and historical newspapers, full-text dissertations and theses, scholarly journals, and primary sources	Subscription-based	10 datasets where one dataset limits to half a million records with a maximum limit of 15MB export of code and resultant dataset per week
Netlytic https://netlytic.org/	Cloud-based tool that uses public API to collect data and also provide basic text analysis and visualization	Twitter, Instagram, YouTube, Facebook	Free access up to 2500 records but paid if want data up to 100,000	1000 most recent tweets
Twitter Archiving Google Sheet (TAGS) https://tags.hawksey.info/	Google spreadsheet template that allows to download tweets for different hashtags automatically	Twitter	Free	API limit
RapidMiner https://rapidminer.com/	Data mining tool that retrieves data from Twitter using its API and performs various data and text mining analyses	Twitter	Free	API limit
Orange https://orangedatamining.com/	Data mining tool that retrieves data from selected databases using their API and performs various data and text mining analyses	Twitter, Wikipedia, PubMed, The Guardian, NY Times	Free	API limit
Harzing's Publish or Perish https://harzing.com/resources/publish-or-perish	Standalone software that retrieves and analyzes citations from selected databases using their APIs	Google Scholar*, Crossref*, PubMed*, Microsoft Academic**, Web of Science***, Scopus***	* Free; ** Require free registration; *** Require subscription	Google Scholar has 1000 limit per query

Table 2.7 Example of various libraries from R and Python for collecting digital trace Data

Library	Description	Programming language
crossrefapi https://github.com/fabiobatalha/crossrefapi	Retrieves data from Crossref	Python
twitter https://pypi.org/project/twitter/	Retrieves data from Twitter	Python
tweepy https://www.tweepy.org/	Retrieves data from Twitter	Python
python-youtube https://pypi.org/project/python-youtube/	Retrieves data from YouTube	Python
Goodreads https://pypi.org/project/Goodreads/	Retrieves data from Goodreads	Python
arxiv https://pypi.org/project/arxiv/	Retrieves data from arXiv preprints	Python
scihub https://github.com/zaytoun/scihub.py	Retrieves data from Sci-Hub	Python
gtrendsR https://github.com/PMassicotte/gtrendsR	Retrieves data from Google Trends	R
rtweet https://www.rdocumentation.org/packages/rtweet/versions/0.4.0	Retrieves data from Twitter	R
aRxiv https://github.com/ropensci/aRxiv	Retrieves data from arXiv preprint repository	R
rcrossref https://github.com/ropensci/rcrossref	Retrieves data from Crossref	R
scholar https://cran.r-project.org/web/packages/scholar/index.html	Retrieves data from Google Scholar	R
rAltmetric https://cran.r-project.org/web/packages/rAltmetric/README.html	Retrieves data from Altmetric database	R
bibliometrix https://bibliometrix.org/	Shiny app that retrieves metadata from PubMed, Digital Science Dimensions, and Cochrane databases using their APIs and also perform various bibliometric methods of analysis	R

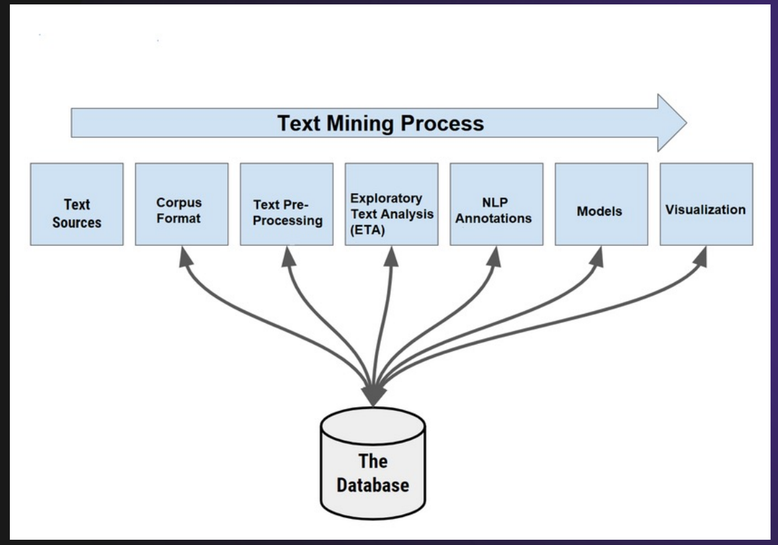


Chapters

- Chapter 1: The Computational Library
  - Case Study: Clustering of Documents using Two Different Tools DOI 10.5281/zenodo.5203303
- Chapter 2: Text Data and Where to Find Them?
- Chapter 3: Text Pre-Processing
  - Case Study: An Analysis of Tolkien's Books
- Chapter 4: Topic Modeling
  - Case Study: Topic Modeling of Documents using Three Different Tools DOI 10.5281/zenodo.5203494
- Chapter 5: Network Text Analysis
  - Case Study: Network Text Analysis of Documents using Two Different R Packages DOI 10.5281/zenodo.5203302
- Chapter 6: Burst Detection
  - Case Study: Burst Detection of Documents using Two Different Tools DOI 10.5281/zenodo.5203298
- Chapter 7: Sentiment Analysis
  - Case Study: Sentiment Analysis of Documents using Two Different Tools DOI 10.5281/zenodo.5203347
- Chapter 8: Predictive Modeling
  - Case Study: Predictive Modeling of Documents using RapidMiner DOI 10.5281/zenodo.5203567
- Chapter 9: Information Visualization
- Chapter 10: Tools and Techniques for Text Mining and Visualizations
- Chapter 11: Text Data and Mining Ethics
- Appendix A: Online Repositories Available for Text Mining DOI 10.5281/zenodo.5104488
- Appendix B: Language Corpora Available for Text Mining DOI 10.5281/zenodo.5104678
- Appendix C: Text Data and Mining Licensing Conditions DOI 10.5281/zenodo.5104740

!! BONUS -- Curated Datasets: This repository contains some of the additional datasets which are in open-access and can be used to practice or teach text mining. The goal of this repository is to act as a collection of textual data set to be used for training and practice in text mining/NLP.

Text Pre-Processing

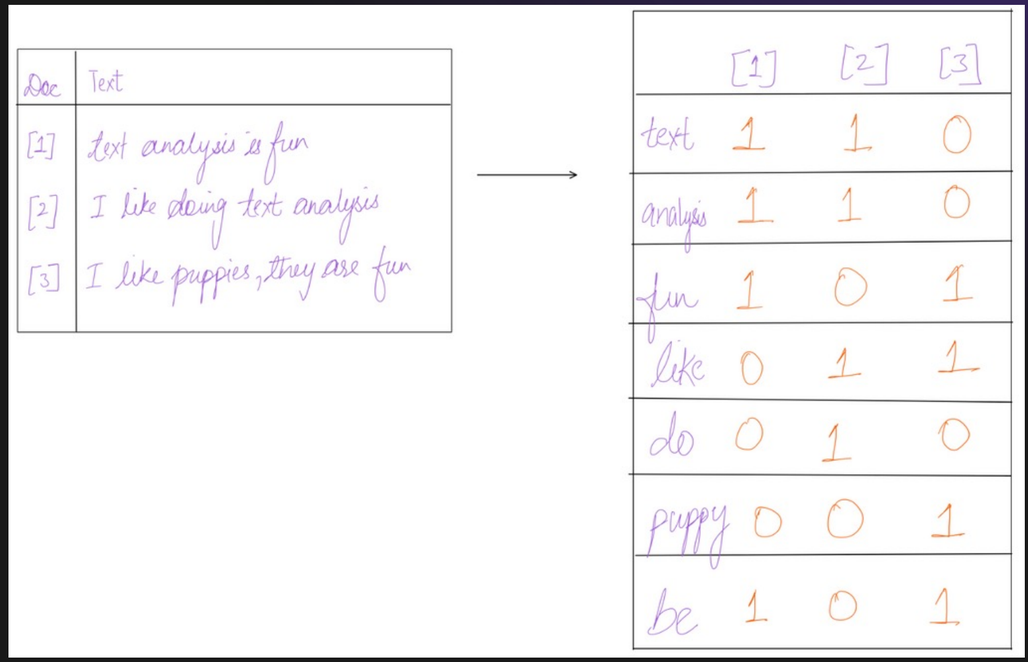


The initial presentation of data in a text mining process includes a combination of words, sentences, and paragraph

In text mining, textual data is required to be transformed into structured form (i.e. numerical data in a tabular format) for further analyses

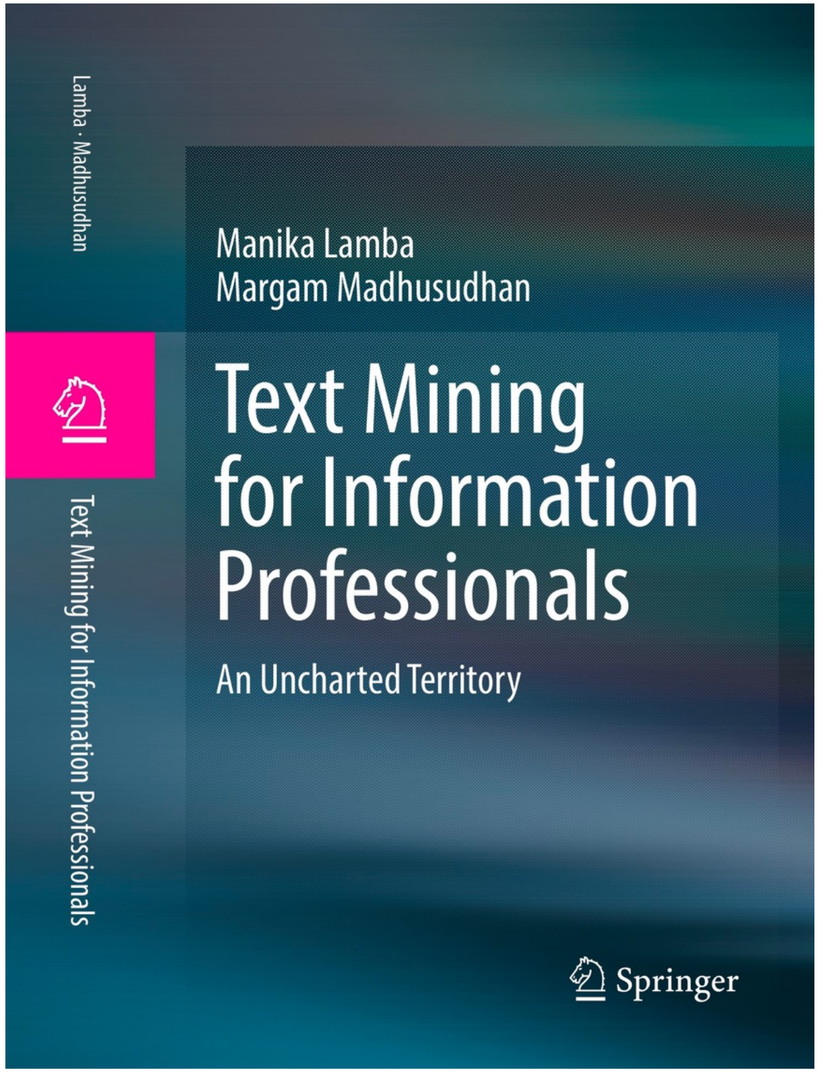
Row represents a document and column represents measurements taken to indicate the presence or absence of words for all the rows without understanding specific properties of text such as concepts of grammar or meaning of words

Not all text mining applications require the same level of text pre-processing but are domain- and task-specific



Doc	Text
[1]	text analysis is fun
[2]	I like doing text analysis
[3]	I like puppies, they are fun

	[1]	[2]	[3]
text	1	1	0
analysis	1	1	0
fun	1	0	1
like	0	1	1
do	0	1	0
puppy	0	0	1
be	1	0	1



Chapters

- Chapter 1: The Computational Library
  - Case Study: Clustering of Documents using Two Different Tools DOI 10.5281/zenodo.5203303
- Chapter 2: Text Data and Where to Find Them?
- Chapter 3: Text Pre-Processing
  - Case Study: An Analysis of Tolkien's Books
- Chapter 4: Topic Modeling
  - Case Study: Topic Modeling of Documents using Three Different Tools DOI 10.5281/zenodo.5203494
- Chapter 5: Network Text Analysis
  - Case Study: Network Text Analysis of Documents using Two Different R Packages DOI 10.5281/zenodo.5203302
- Chapter 6: Burst Detection
  - Case Study: Burst Detection of Documents using Two Different Tools DOI 10.5281/zenodo.5203298
- Chapter 7: Sentiment Analysis
  - Case Study: Sentiment Analysis of Documents using Two Different Tools DOI 10.5281/zenodo.5203347
- Chapter 8: Predictive Modeling
  - Case Study: Predictive Modeling of Documents using RapidMiner DOI 10.5281/zenodo.5203567
- Chapter 9: Information Visualization
- Chapter 10: Tools and Techniques for Text Mining and Visualizations
- Chapter 11: Text Data and Mining Ethics
- Appendix A: Online Repositories Available for Text Mining DOI 10.5281/zenodo.5104488
- Appendix B: Language Corpora Available for Text Mining DOI 10.5281/zenodo.5104678
- Appendix C: Text Data and Mining Licensing Conditions DOI 10.5281/zenodo.5104740

!! BONUS -- Curated Datasets: This repository contains some of the additional datasets which are in open-access and can be used to practice or teach text mining. The goal of this repository is to act as a collection of textual data set to be used for training and practice in text mining/NLP.

Topic Modeling

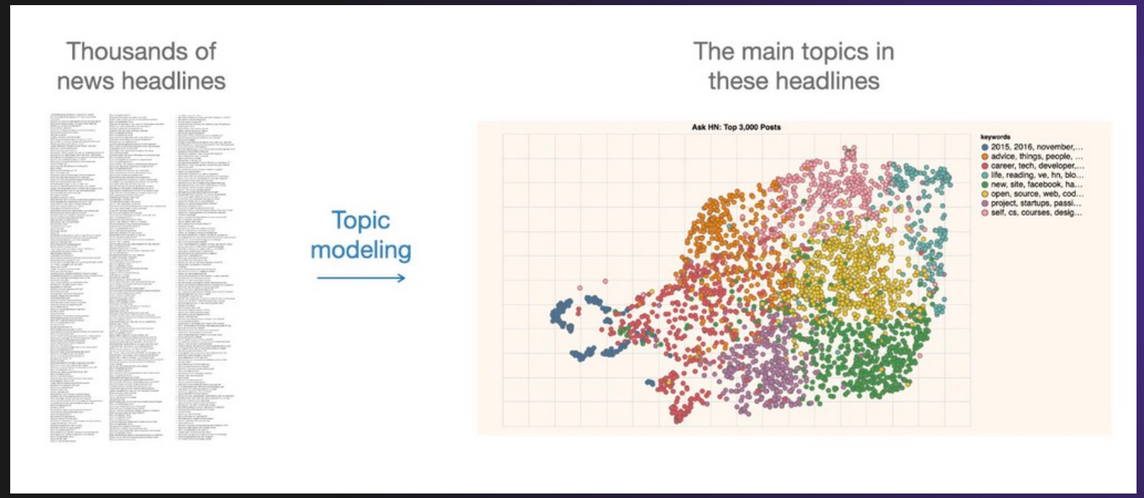
It is “a method for finding and tracing clusters of words (called *topics*) in large bodies of texts” (Brett)

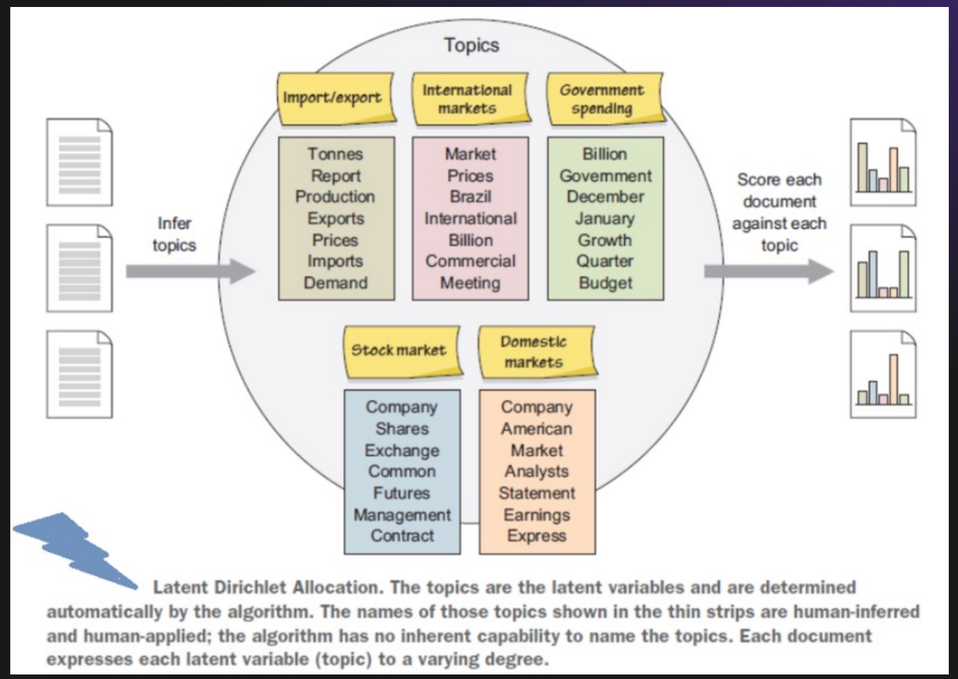
It performs *soft clustering*, where it presumes that every document is composed of a mixture of topics

It makes an excellent tool for discovery and helps to uncover evidence already present in the text

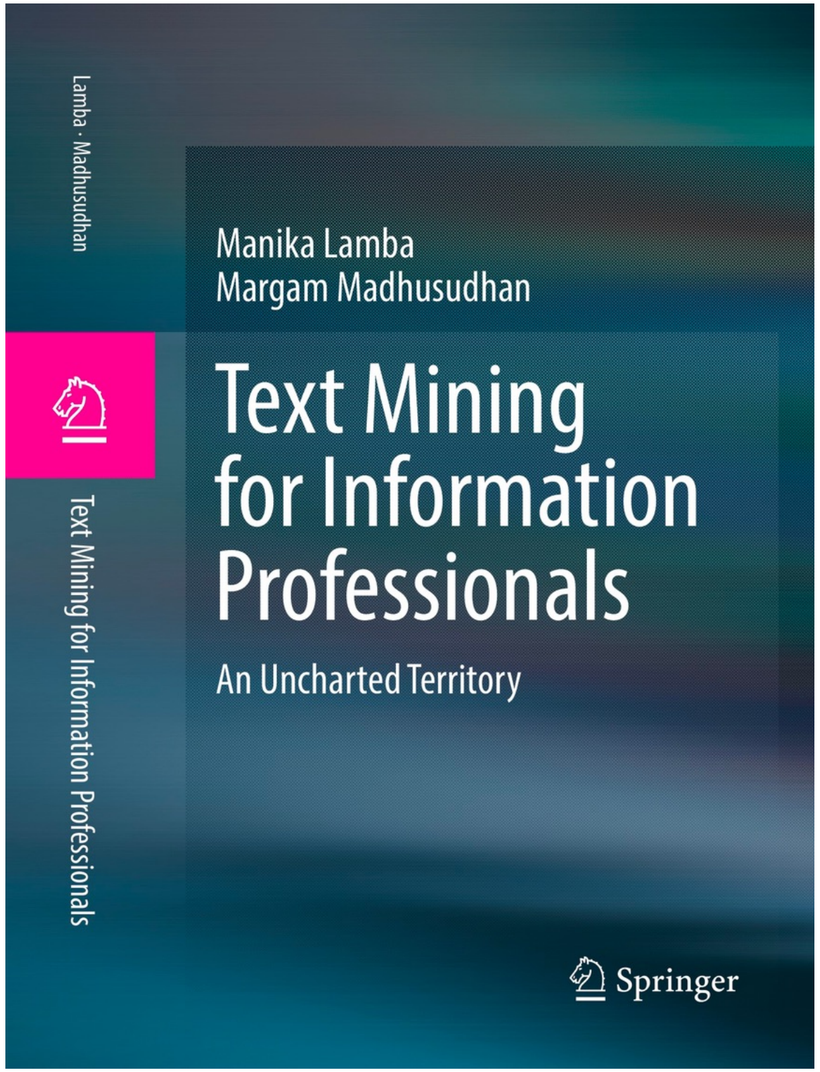
It has been called an act of reading tea leaves (Chang et al., 2009) or the process of highlighting words (Brett) based on their topics

Topics are simply groups of words from the collection of documents that represents the information in the collection in the best way





Latent Dirichlet Allocation. The topics are the latent variables and are determined automatically by the algorithm. The names of those topics shown in the thin strips are human-inferred and human-applied; the algorithm has no inherent capability to name the topics. Each document expresses each latent variable (topic) to a varying degree.



Chapters

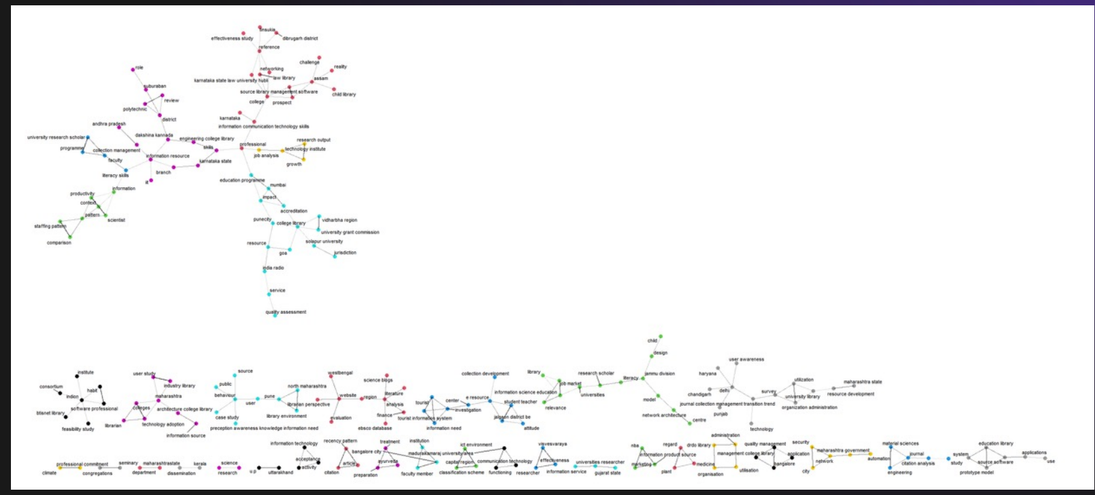
- Chapter 1: The Computational Library
  - Case Study: Clustering of Documents using Two Different Tools DOI 10.5281/zenodo.5203303
- Chapter 2: Text Data and Where to Find Them?
- Chapter 3: Text Pre-Processing
  - Case Study: An Analysis of Tolkien's Books
- Chapter 4: Topic Modeling
  - Case Study: Topic Modeling of Documents using Three Different Tools DOI 10.5281/zenodo.5203494
- Chapter 5: Network Text Analysis
  - Case Study: Network Text Analysis of Documents using Two Different R Packages DOI 10.5281/zenodo.5203302
- Chapter 6: Burst Detection
  - Case Study: Burst Detection of Documents using Two Different Tools DOI 10.5281/zenodo.5203298
- Chapter 7: Sentiment Analysis
  - Case Study: Sentiment Analysis of Documents using Two Different Tools DOI 10.5281/zenodo.5203347
- Chapter 8: Predictive Modeling
  - Case Study: Predictive Modeling of Documents using RapidMiner DOI 10.5281/zenodo.5203567
- Chapter 9: Information Visualization
- Chapter 10: Tools and Techniques for Text Mining and Visualizations
- Chapter 11: Text Data and Mining Ethics
- Appendix A: Online Repositories Available for Text Mining DOI 10.5281/zenodo.5104488
- Appendix B: Language Corpora Available for Text Mining DOI 10.5281/zenodo.5104678
- Appendix C: Text Data and Mining Licensing Conditions DOI 10.5281/zenodo.5104740

!! BONUS -- Curated Datasets: This repository contains some of the additional datasets which are in open-access and can be used to practice or teach text mining. The goal of this repository is to act as a collection of textual data set to be used for training and practice in text mining/NLP.

Network Text Analysis

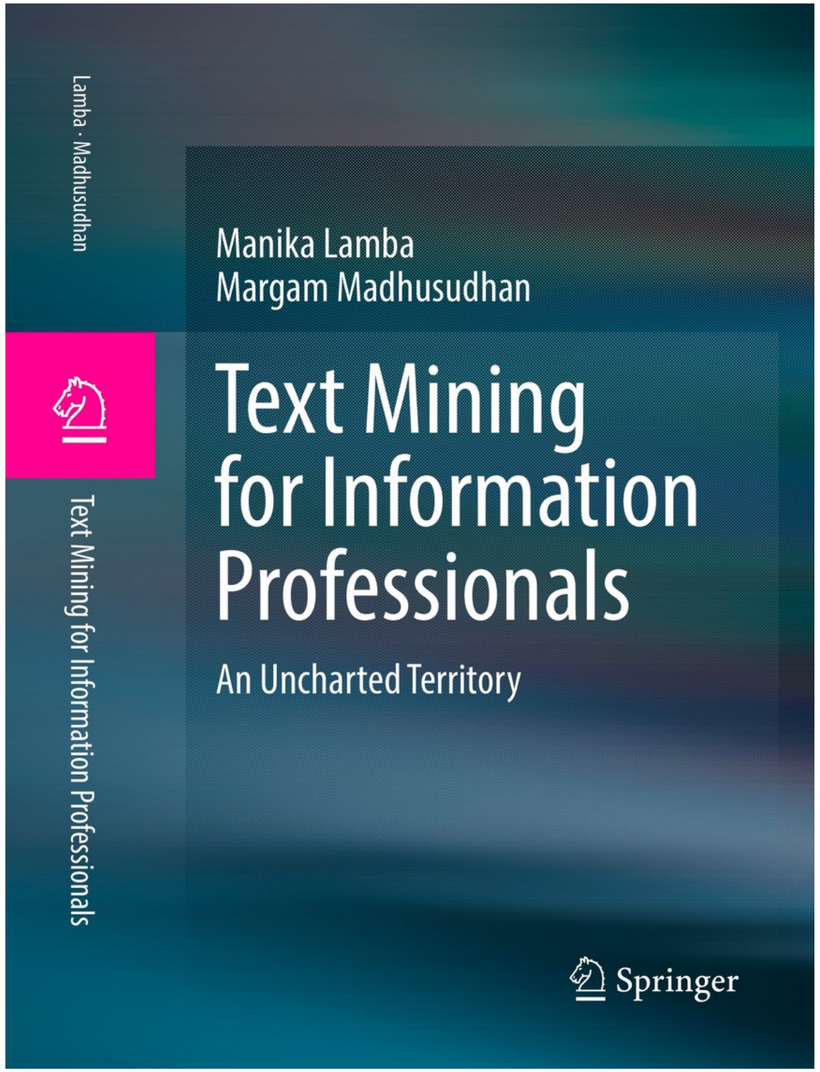
In 2005, Diesner and Carley coined the phrase network text analysis (NTA)

A collection of texts may be presented as a text network where every node represents a text, and the edges show the similarities between the words used in two or more documents



A text network can also be used to visualize individual words as nodes and their frequency with which they co-occur in the texts as the edges

Similar to topic modeling, you can cluster words in latent themes or topics in the form of word-based projections of the text network



Chapters

- Chapter 1: The Computational Library
  - Case Study: Clustering of Documents using Two Different Tools DOI 10.5281/zenodo.5203303
- Chapter 2: Text Data and Where to Find Them?
- Chapter 3: Text Pre-Processing
  - Case Study: An Analysis of Tolkien's Books
- Chapter 4: Topic Modeling
  - Case Study: Topic Modeling of Documents using Three Different Tools DOI 10.5281/zenodo.5203494
- Chapter 5: Network Text Analysis
  - Case Study: Network Text Analysis of Documents using Two Different R Packages DOI 10.5281/zenodo.5203302
- Chapter 6: Burst Detection
  - Case Study: Burst Detection of Documents using Two Different Tools DOI 10.5281/zenodo.5203298
- Chapter 7: Sentiment Analysis
  - Case Study: Sentiment Analysis of Documents using Two Different Tools DOI 10.5281/zenodo.5203347
- Chapter 8: Predictive Modeling
  - Case Study: Predictive Modeling of Documents using RapidMiner DOI 10.5281/zenodo.5203567
- Chapter 9: Information Visualization
- Chapter 10: Tools and Techniques for Text Mining and Visualizations
- Chapter 11: Text Data and Mining Ethics
- Appendix A: Online Repositories Available for Text Mining DOI 10.5281/zenodo.5104488
- Appendix B: Language Corpora Available for Text Mining DOI 10.5281/zenodo.5104678
- Appendix C: Text Data and Mining Licensing Conditions DOI 10.5281/zenodo.5104740

!! BONUS -- Curated Datasets: This repository contains some of the additional datasets which are in open-access and can be used to practice or teach text mining. The goal of this repository is to act as a collection of textual data set to be used for training and practice in text mining/NLP.

Burst Detection

Burst detection is a temporal analysis that aims to identify the nature of phenomena represented by a sequence of observations such as patterns, trends, seasonality, outliers, and bursts of activity

It can be used to (i) understand the temporal distance such as the most emerging or trending terms, growth of terms, and latency/peak or (ii) forecasting, that is, predicting, future values of the time-series variables

Topic modeling has difficulty in coherently linking topics together between subsequent time-steps. In contrast, burst detections first identify the bursty terms in a dataset, and then cluster them together into topics

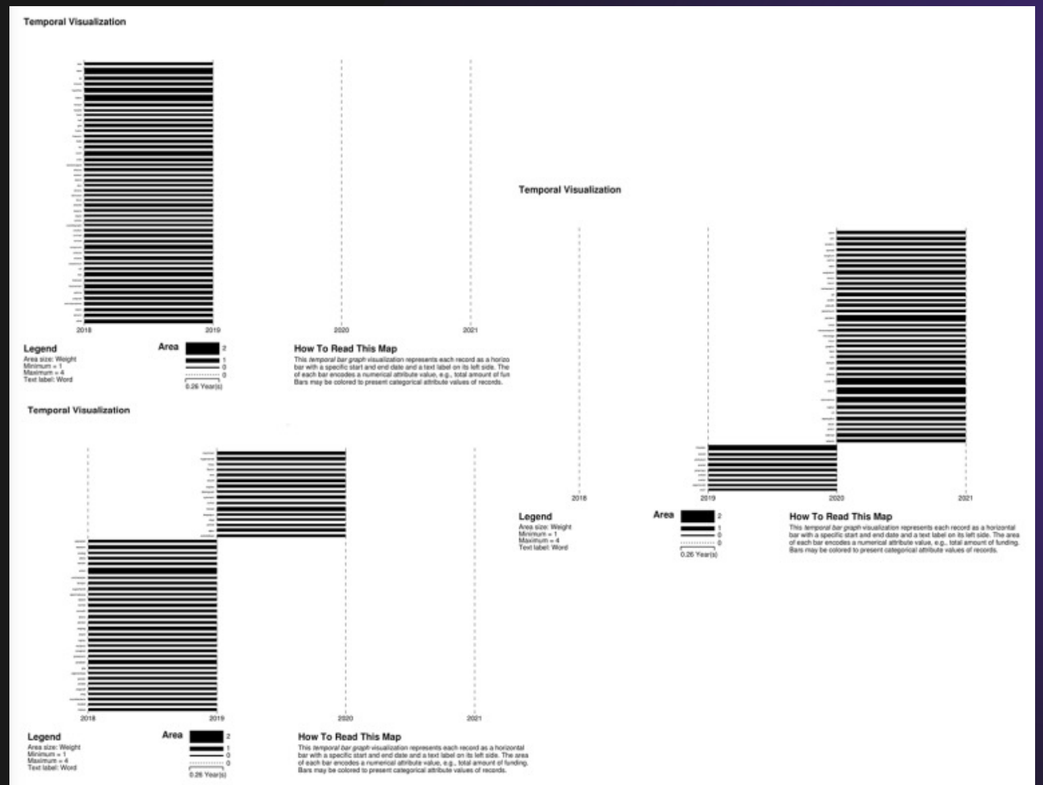
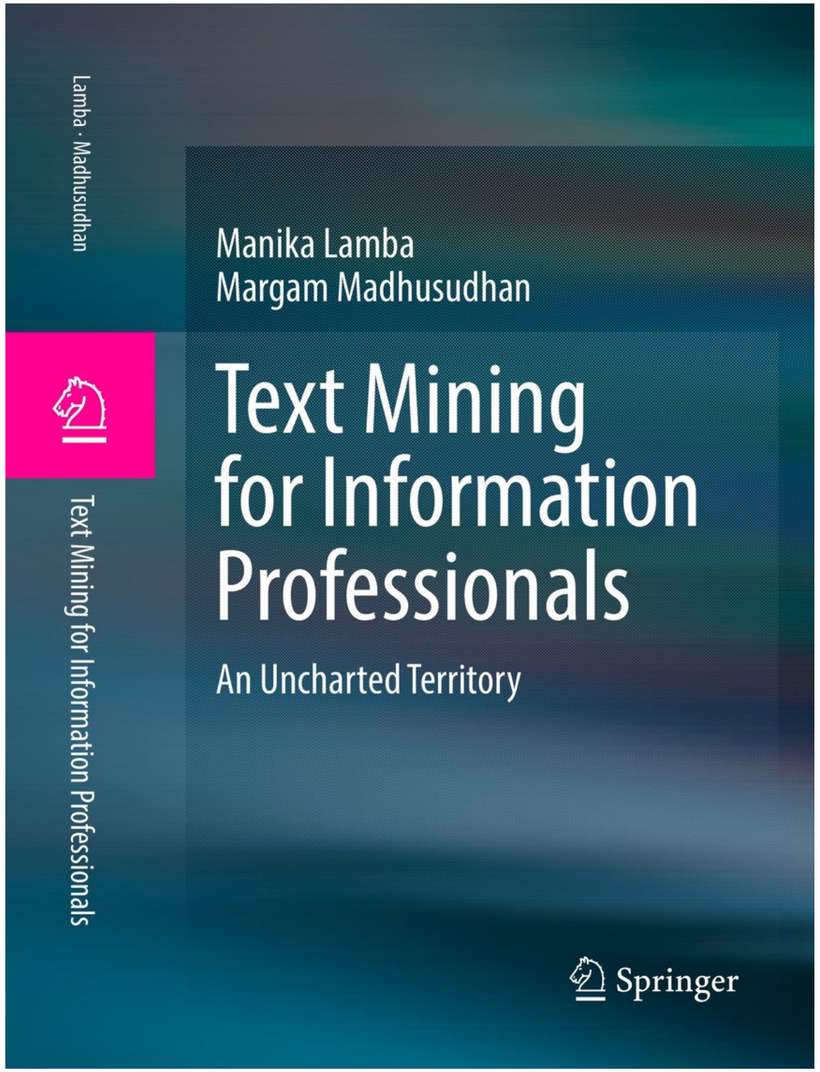


Fig. 6.10 Horizontal line graph

Table 6.1 Summarization of burst results

Level	Period/year	Top 5 words	Area	Topic
1	2018	Label, union, histon, event, hypothes	3.95, 3.93, 3.68, 2.84, 2.78	Histones
2	2019–2020	Threaten, breast, cytometri, machineri, elucid	3.26, 2.60, 2.10, 1.99, 1.95	Breast cancer
1	2020	Covid-19, cov-2, pandem, coronavirus, respiratori	4.23, 4.23, 3.96, 3.43	Coronavirus (COVID-19) pandemic



Chapters

- Chapter 1: The Computational Library
  - Case Study: Clustering of Documents using Two Different Tools DOI 10.5281/zenodo.5203303
- Chapter 2: Text Data and Where to Find Them?
- Chapter 3: Text Pre-Processing
  - Case Study: An Analysis of Tolkien's Books
- Chapter 4: Topic Modeling
  - Case Study: Topic Modeling of Documents using Three Different Tools DOI 10.5281/zenodo.5203494
- Chapter 5: Network Text Analysis
  - Case Study: Network Text Analysis of Documents using Two Different R Packages DOI 10.5281/zenodo.5203302
- Chapter 6: Burst Detection
  - Case Study: Burst Detection of Documents using Two Different Tools DOI 10.5281/zenodo.5203298
- Chapter 7: Sentiment Analysis
  - Case Study: Sentiment Analysis of Documents using Two Different Tools DOI 10.5281/zenodo.5203347
- Chapter 8: Predictive Modeling
  - Case Study: Predictive Modeling of Documents using RapidMiner DOI 10.5281/zenodo.5203567
- Chapter 9: Information Visualization
- Chapter 10: Tools and Techniques for Text Mining and Visualizations
- Chapter 11: Text Data and Mining Ethics
- Appendix A: Online Repositories Available for Text Mining DOI 10.5281/zenodo.5104488
- Appendix B: Language Corpora Available for Text Mining DOI 10.5281/zenodo.5104678
- Appendix C: Text Data and Mining Licensing Conditions DOI 10.5281/zenodo.5104740

!! BONUS -- Curated Datasets: This repository contains some of the additional datasets which are in open-access and can be used to practice or teach text mining. The goal of this repository is to act as a collection of textual data set to be used for training and practice in text mining/NLP.

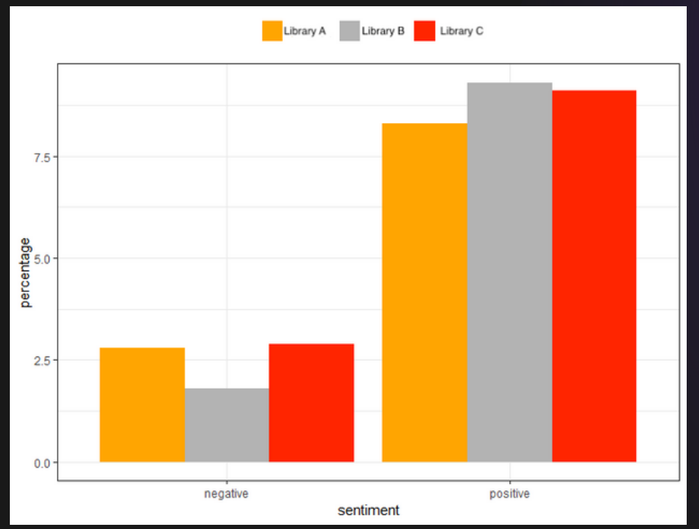
Sentiment Analysis

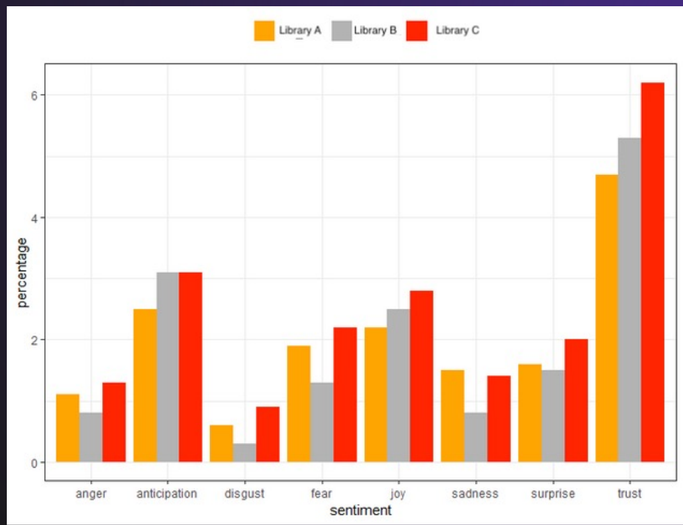
it is a natural language processing (NLP) technique that identifies important patterns of information and features from a large text corpus

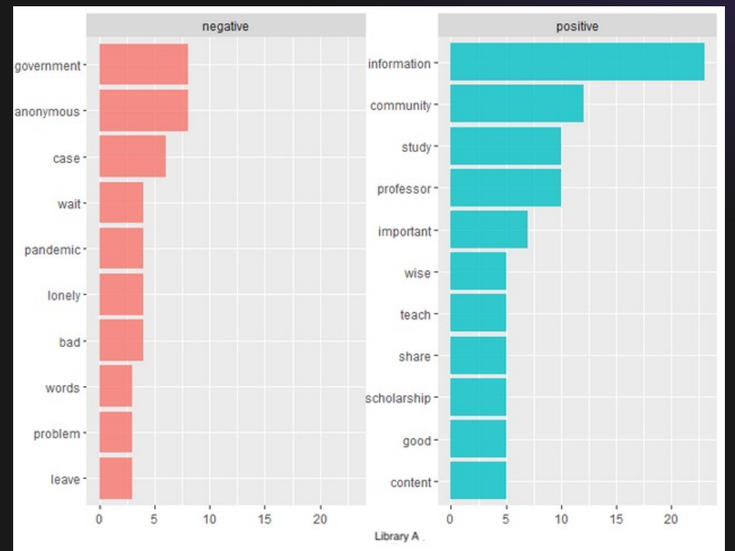
It analyzes thoughts, attitude, views, opinions, beliefs, comments, requests, questions, and preferences expressed by an author based on emotion rather than a reason in the form of text

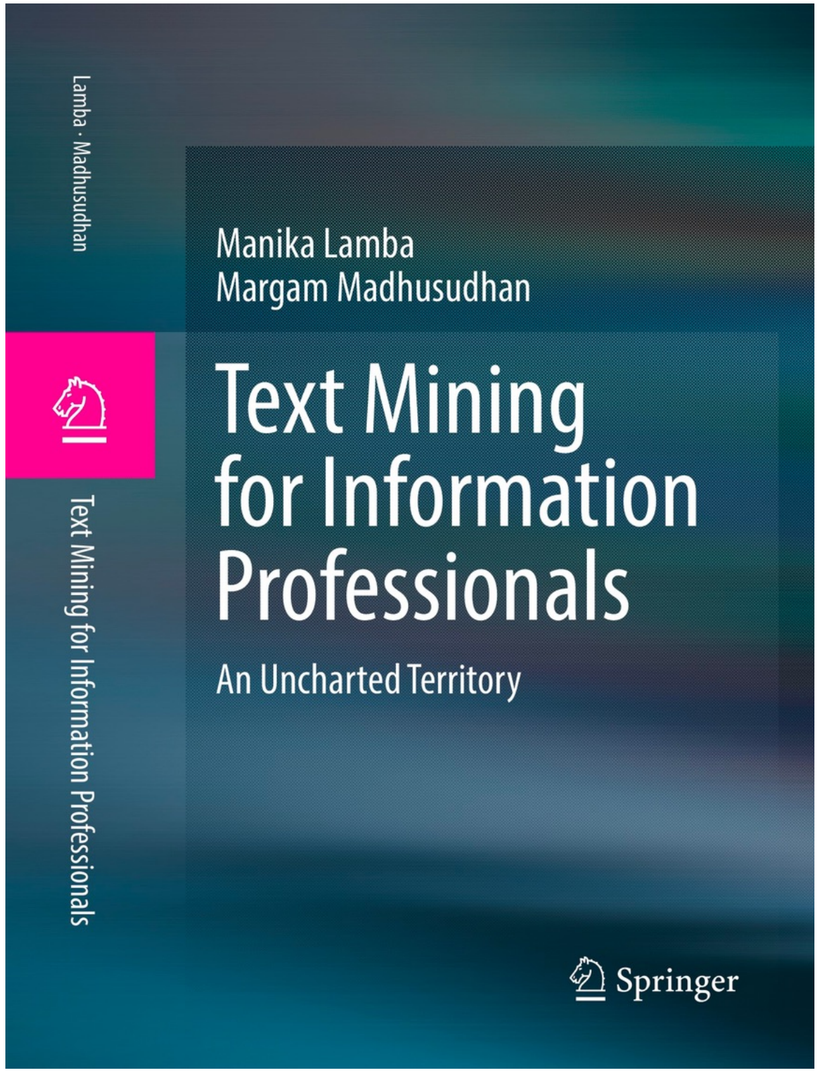
Text can be blog posts, product reviews, online forums, speech, database sources, social media data, and documents

It classifies the author's feelings into polarity (positive, negative, or neutral), subjectivity (objective or subjective), and emotions (angry, happy, surprised, sad, jealous, and mixed)









Chapters

- Chapter 1: The Computational Library
  - Case Study: Clustering of Documents using Two Different Tools DOI 10.5281/zenodo.5203303
- Chapter 2: Text Data and Where to Find Them?
- Chapter 3: Text Pre-Processing
  - Case Study: An Analysis of Tolkien's Books
- Chapter 4: Topic Modeling
  - Case Study: Topic Modeling of Documents using Three Different Tools DOI 10.5281/zenodo.5203494
- Chapter 5: Network Text Analysis
  - Case Study: Network Text Analysis of Documents using Two Different R Packages DOI 10.5281/zenodo.5203302
- Chapter 6: Burst Detection
  - Case Study: Burst Detection of Documents using Two Different Tools DOI 10.5281/zenodo.5203298
- Chapter 7: Sentiment Analysis
  - Case Study: Sentiment Analysis of Documents using Two Different Tools DOI 10.5281/zenodo.5203347
- Chapter 8: Predictive Modeling
  - Case Study: Predictive Modeling of Documents using RapidMiner DOI 10.5281/zenodo.5203567
- Chapter 9: Information Visualization
- Chapter 10: Tools and Techniques for Text Mining and Visualizations
- Chapter 11: Text Data and Mining Ethics
- Appendix A: Online Repositories Available for Text Mining DOI 10.5281/zenodo.5104488
- Appendix B: Language Corpora Available for Text Mining DOI 10.5281/zenodo.5104678
- Appendix C: Text Data and Mining Licensing Conditions DOI 10.5281/zenodo.5104740

!! BONUS -- Curated Datasets: This repository contains some of the additional datasets which are in open-access and can be used to practice or teach text mining. The goal of this repository is to act as a collection of textual data set to be used for training and practice in text mining/NLP.

Predictive Modeling

Traditionally, humans analyzed the data, but the volume of data surpasses their ability to make sense of it, which made them automate systems that can learn from the data

In most computing problems that need to be solved, a script is written manually that further includes a series of steps to solve that particular problem

However, for some problems like speech recognition, it is difficult to write a program manually

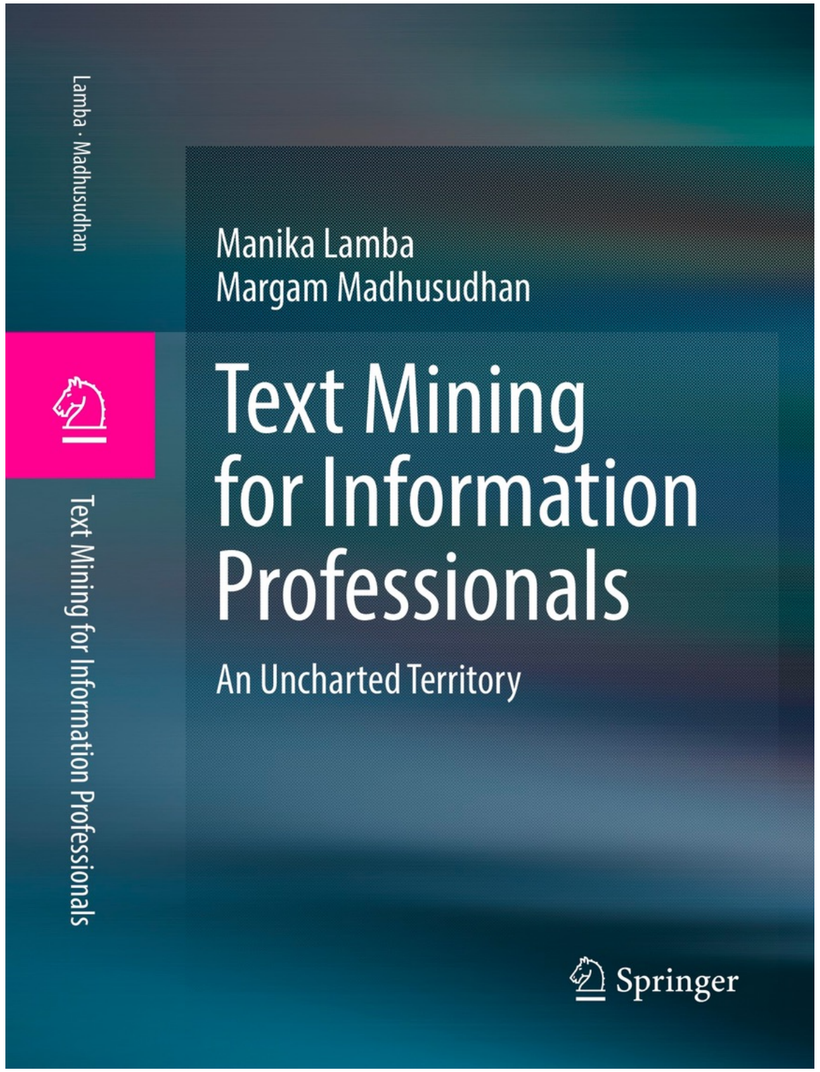
Machine Learning is a technique that helps to learn complex rules efficiently

It uses labeled instances called training data instead of programming all the rules manually

kappa: 0.997

	true Topic a	true Topic b	true Topic c	true Topic d	true Topic e	true Topic f	true Topic g	true Topic h	class precis...
pred. Topic a	50	0	0	0	0	0	0	0	100.00%
pred. Topic b	0	111	0	0	0	0	0	0	100.00%
pred. Topic c	0	0	45	0	1	0	0	0	97.83%
pred. Topic d	0	0	0	41	0	0	0	0	100.00%
pred. Topic e	0	0	0	0	58	0	0	0	100.00%
pred. Topic f	0	0	0	0	0	29	0	0	100.00%
pred. Topic g	0	0	0	0	0	0	65	0	100.00%
pred. Topic h	0	0	0	0	0	0	0	41	100.00%
class recall	100.00%	100.00%	100.00%	100.00%	98.31%	100.00%	100.00%	100.00%	

Fig. 8.12 Screenshot showing the evaluation results (©2020 Cadernos BAD, all rights reserved—reprinted under Creative Commons CC BY license, published in Lamba and Madhusudhan [4])



Chapters

- Chapter 1: The Computational Library
  - Case Study: Clustering of Documents using Two Different Tools DOI 10.5281/zenodo.5203303
- Chapter 2: Text Data and Where to Find Them?
- Chapter 3: Text Pre-Processing
  - Case Study: An Analysis of Tolkien's Books
- Chapter 4: Topic Modeling
  - Case Study: Topic Modeling of Documents using Three Different Tools DOI 10.5281/zenodo.5203494
- Chapter 5: Network Text Analysis
  - Case Study: Network Text Analysis of Documents using Two Different R Packages DOI 10.5281/zenodo.5203302
- Chapter 6: Burst Detection
  - Case Study: Burst Detection of Documents using Two Different Tools DOI 10.5281/zenodo.5203298
- Chapter 7: Sentiment Analysis
  - Case Study: Sentiment Analysis of Documents using Two Different Tools DOI 10.5281/zenodo.5203347
- Chapter 8: Predictive Modeling
  - Case Study: Predictive Modeling of Documents using RapidMiner DOI 10.5281/zenodo.5203567
- Chapter 9: Information Visualization
- Chapter 10: Tools and Techniques for Text Mining and Visualizations
- Chapter 11: Text Data and Mining Ethics
- Appendix A: Online Repositories Available for Text Mining DOI 10.5281/zenodo.5104488
- Appendix B: Language Corpora Available for Text Mining DOI 10.5281/zenodo.5104678
- Appendix C: Text Data and Mining Licensing Conditions DOI 10.5281/zenodo.5104740

!! BONUS -- Curated Datasets: This repository contains some of the additional datasets which are in open-access and can be used to practice or teach text mining. The goal of this repository is to act as a collection of textual data set to be used for training and practice in text mining/NLP.

Information Visualization

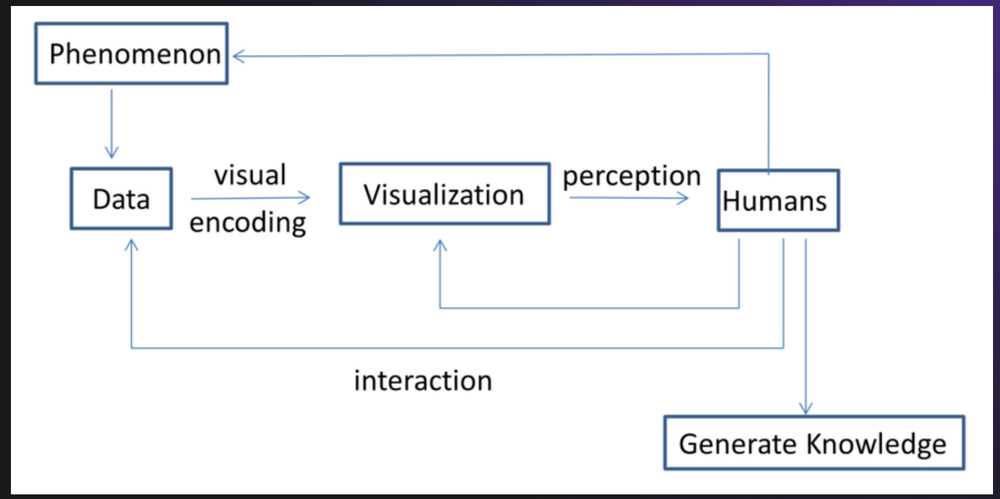
Visualization is a way to store information out of our minds and make it accessible through our eyes and manipulations with interactive systems

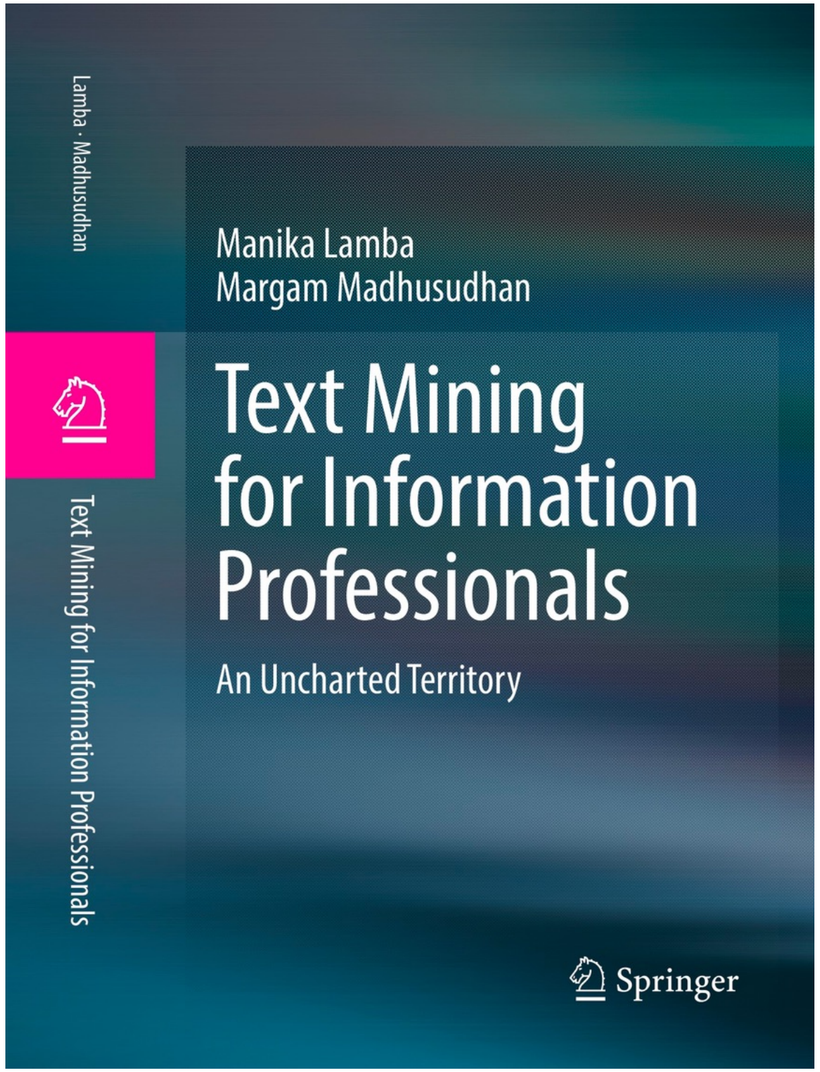
In contrast, information visualization helps to transform data into a visual representation

The basic idea of visualization is to transform data into something that enhances the understanding of the given data

It teaches how to design, evaluate, and develop interactive visualizations to help people generate insights and then communicate them to other people as effectively as possible

One can ask, why do we use data, or why do we visualize data? ---
It is not because we are interested in the data itself, but we do it because it is an abstract representation of some reality or some phenomenon we find interesting.





Chapters

- Chapter 1: The Computational Library
  - Case Study: Clustering of Documents using Two Different Tools DOI 10.5281/zenodo.5203303
- Chapter 2: Text Data and Where to Find Them?
- Chapter 3: Text Pre-Processing
  - Case Study: An Analysis of Tolkien's Books
- Chapter 4: Topic Modeling
  - Case Study: Topic Modeling of Documents using Three Different Tools DOI 10.5281/zenodo.5203494
- Chapter 5: Network Text Analysis
  - Case Study: Network Text Analysis of Documents using Two Different R Packages DOI 10.5281/zenodo.5203302
- Chapter 6: Burst Detection
  - Case Study: Burst Detection of Documents using Two Different Tools DOI 10.5281/zenodo.5203298
- Chapter 7: Sentiment Analysis
  - Case Study: Sentiment Analysis of Documents using Two Different Tools DOI 10.5281/zenodo.5203347
- Chapter 8: Predictive Modeling
  - Case Study: Predictive Modeling of Documents using RapidMiner DOI 10.5281/zenodo.5203567
- Chapter 9: Information Visualization
- Chapter 10: Tools and Techniques for Text Mining and Visualizations
- Chapter 11: Text Data and Mining Ethics
- Appendix A: Online Repositories Available for Text Mining DOI 10.5281/zenodo.5104488
- Appendix B: Language Corpora Available for Text Mining DOI 10.5281/zenodo.5104678
- Appendix C: Text Data and Mining Licensing Conditions DOI 10.5281/zenodo.5104740

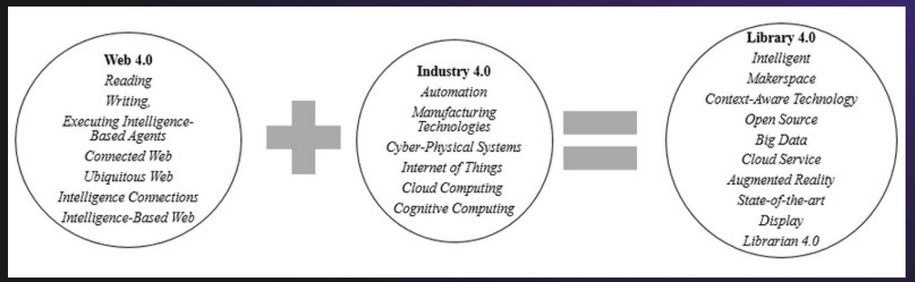
!! BONUS -- Curated Datasets: This repository contains some of the additional datasets which are in open-access and can be used to practice or teach text mining. The goal of this repository is to act as a collection of textual data set to be used for training and practice in text mining/NLP.

Tools & Techniques for Text Mining & Visualizations

A large number of machine learning (ML) and data mining (DM) tools have been created in the past 25 years, where the primary intent was to aid the cumbersome data analysis process

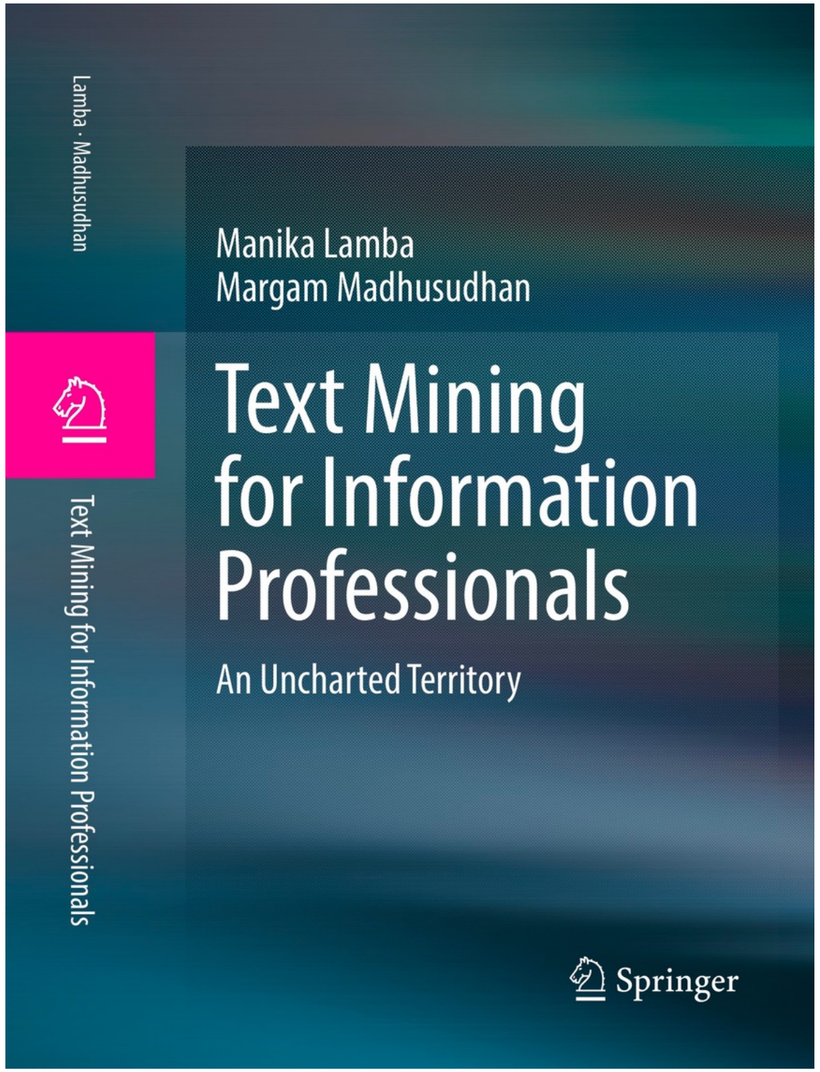
Industry 4.0 produces more software and platforms to provide tools for text and data mining (TDM), both Web 4.0 (Intelligent Web) and Industry 4.0 (Intelligent Industry) have amalgamated to give rise to what is called Library 4.0 (intelligent library)

An intelligent library (Library 4.0) evaluates significant data functions to analyze information and then uses the analyzed result for decision-making or to provide a value-based service to its customers



It is not practical to efficiently extract and analyze useful information manually. Software solutions in the form of automatic tools by Industry 4.0 are required to analyze, extract, and organize a large volume of relevant information from the textual data

With the increasing demands to visualize the results from the substantial amount of textual data available on the Web, information visualization is receiving considerable importance in research

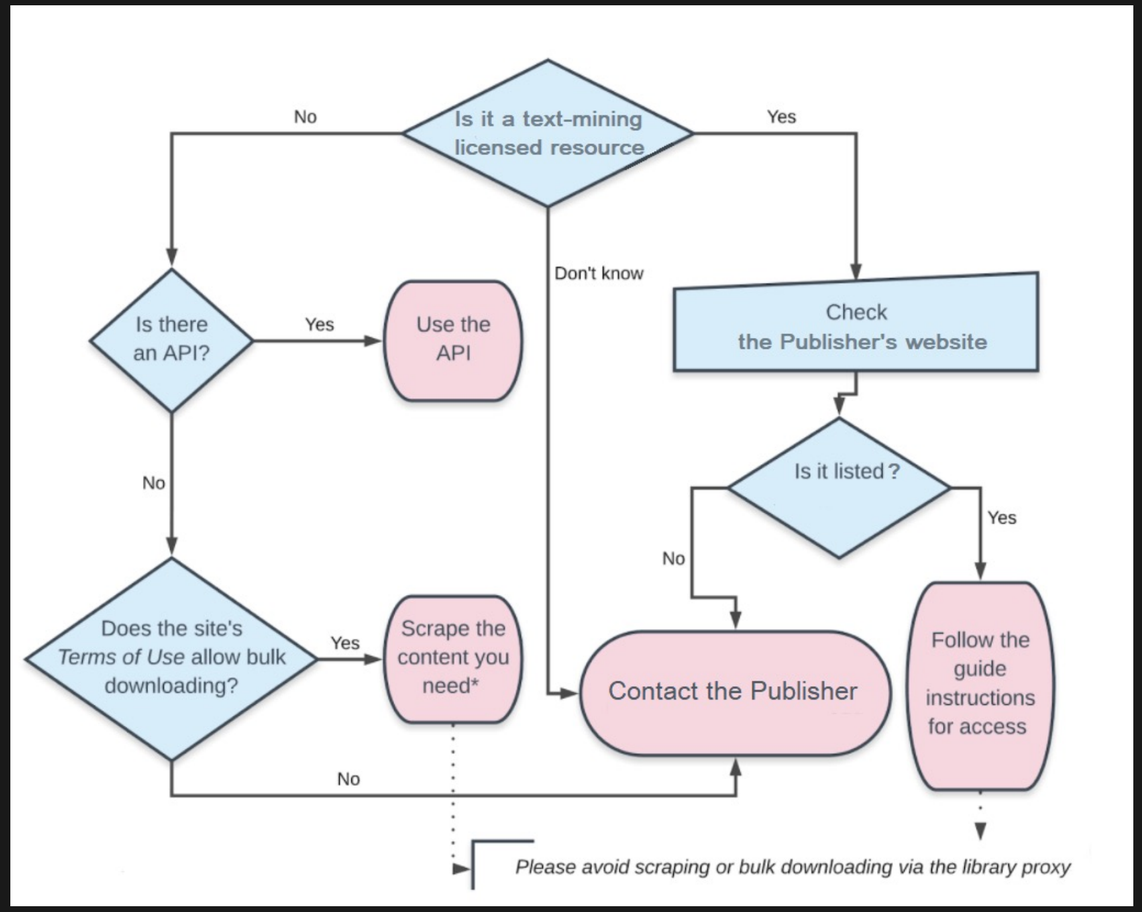


Chapters

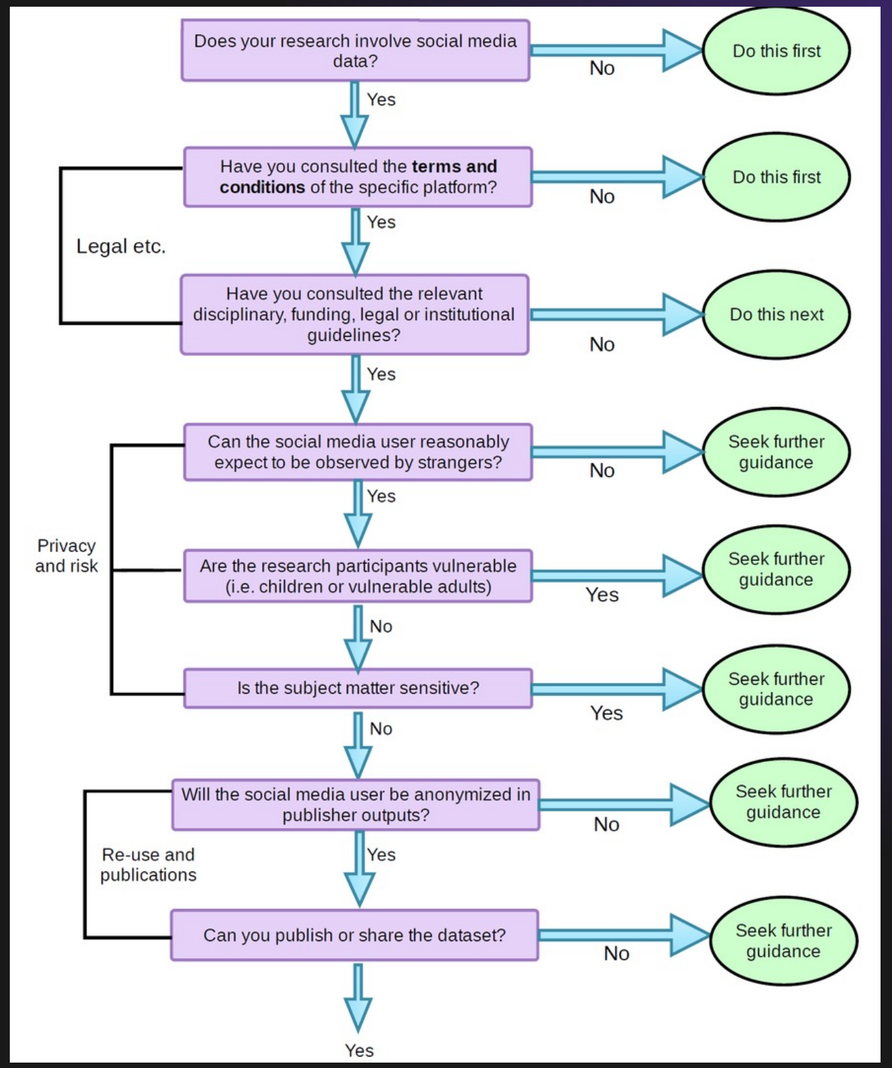
- Chapter 1: The Computational Library
  - Case Study: Clustering of Documents using Two Different Tools DOI 10.5281/zenodo.5203303
- Chapter 2: Text Data and Where to Find Them?
- Chapter 3: Text Pre-Processing
  - Case Study: An Analysis of Tolkien's Books
- Chapter 4: Topic Modeling
  - Case Study: Topic Modeling of Documents using Three Different Tools DOI 10.5281/zenodo.5203494
- Chapter 5: Network Text Analysis
  - Case Study: Network Text Analysis of Documents using Two Different R Packages DOI 10.5281/zenodo.5203302
- Chapter 6: Burst Detection
  - Case Study: Burst Detection of Documents using Two Different Tools DOI 10.5281/zenodo.5203298
- Chapter 7: Sentiment Analysis
  - Case Study: Sentiment Analysis of Documents using Two Different Tools DOI 10.5281/zenodo.5203347
- Chapter 8: Predictive Modeling
  - Case Study: Predictive Modeling of Documents using RapidMiner DOI 10.5281/zenodo.5203567
- Chapter 9: Information Visualization
- Chapter 10: Tools and Techniques for Text Mining and Visualizations
- Chapter 11: Text Data and Mining Ethics
- Appendix A: Online Repositories Available for Text Mining DOI 10.5281/zenodo.5104488
- Appendix B: Language Corpora Available for Text Mining DOI 10.5281/zenodo.5104678
- Appendix C: Text Data and Mining Licensing Conditions DOI 10.5281/zenodo.5104740

!! BONUS -- Curated Datasets: This repository contains some of the additional datasets which are in open-access and can be used to practice or teach text mining. The goal of this repository is to act as a collection of textual data set to be used for training and practice in text mining/NLP.

Text Data & Mining Ethics



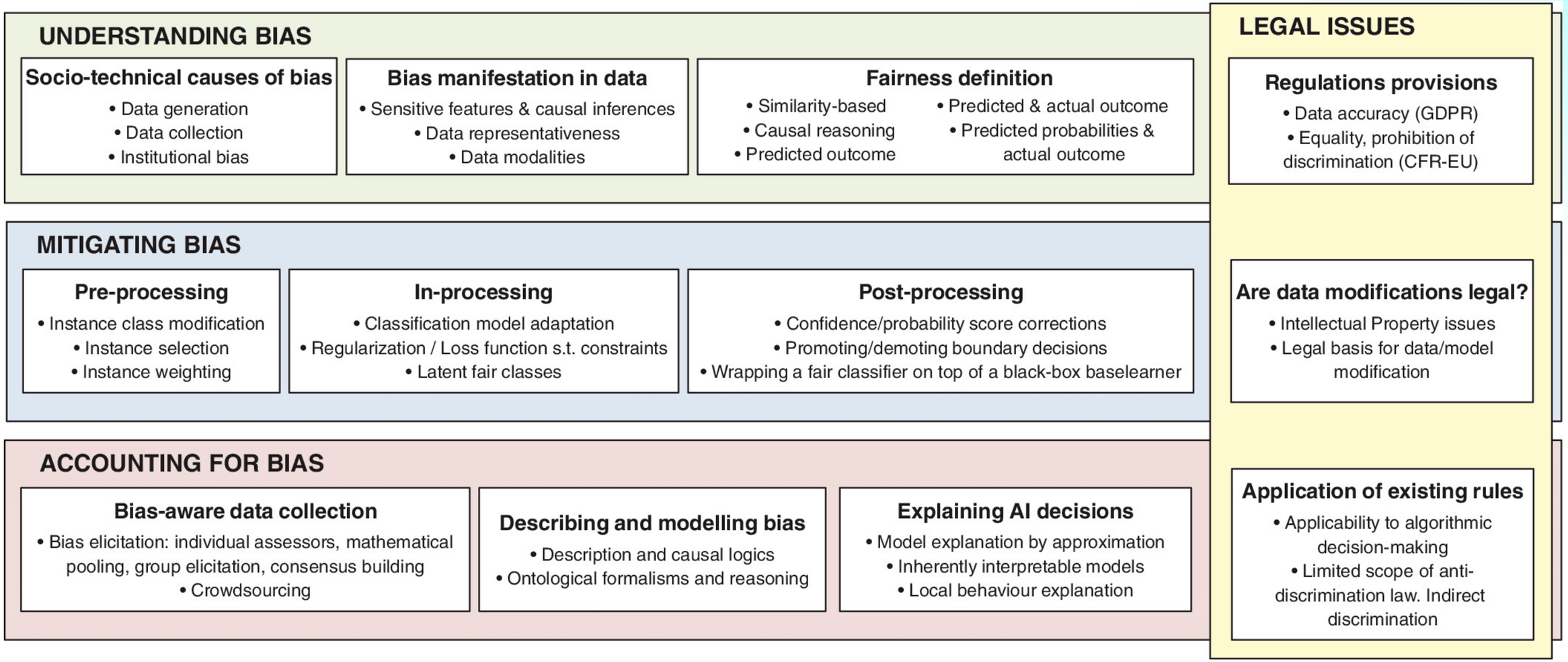
Ethical framework for text mining of databases



Ethical framework for social media data

Table 11.1 Do's and dont's of using text data for mining by librarians (adapted from McNeice et al. [2])

Do's	Dont's
Establish if you will use or mine data that contains sensitive data	Only think of data protection issues when you usually start the process of text mining
Assign a data protection officer if text mining is one of your organization's core activities, or if your organization does text mining regularly	Collect textual data and just assume it does not concern with personal, private, and sensitive data
Impact assessment (IA): establish what data you will use for what purposes, and who will have access to the data within and outside your organization, and whether your use of textual data brings any legal or ethical risks	Store and retain all data just because it may be helpful in the future
Check whether you have the legal grounds to collect and use the textual data	Randomly transfer or provide access to the data to third parties
Privacy by design: based on your impact assessment (IA), design your whole text mining project in a way that guarantees that you can safely and adequately use the textual data	Reuse textual data from one project to another, without making sure it is compatible with data protection rules, even though you had made sure that the use in the first project is compatible
Look into sector-specific regulation or self-regulation and codes of conduct within your domain, which may provide you more guidance and certainty on what kind of analyses and techniques you can employ on the textual data	Share any textual data with the public, without proper consultation
Anonymize data so you are not dealing with any personal, private, or sensitive data. Note that if you pseudonymize personal, private, or sensitive data, one can still be able to identify the anonymized data if additional information is used	Make decisions affecting the data subjects solely on automated processing of their personal data—it should be prohibited
	Ignore the request by the owner/publisher/funder to access, rectify, or erase data
	Transfer textual data to others without permission



Summary of different issues and solutions related to biasness

	Elsevier	Wiley	Springer	Emerald
Where is the licence?	http://www.elsevier.com/tdm/userlicense/1.0/	http://olabout.wiley.com/WileyCDA/Section/id-826542.html	http://www.springer.com/tdm	https://www.emeraldinsight.com/page/tdm
Do I need to check other licenses or documents?	**NO** The TDM Agreement supersedes any and all prior and contemporaneous agreements	**UNCLEAR** The click-through TDM agreements say that it supersedes all other prior and contemporaneous agreements, but all that it is superseded by any separate TDM agreement	**YES** The TDM clause may not have been included in existing SpringerLink subscription agreements but can be added by existing subscribers	**NO** Not mentioned in policy
Does this license affect my use of open access content?	**NO** Individual OA licenses supersede anything the contrary in the TDM Agreement	**NO** If more permissive licenses apply, you may use content in accordance with article-level permissions	**NO** TDM of OA Springer content is usually allowed without restrictions	**NO** Not mentioned in policy
Is TDM permitted?	**YES**	**YES**	**YES**	**YES**
Can I carry out TDM for any purposes?	**NO** You may not extract, develop or use the dataset for any direct or indirect commercial activity	**NO** You may only text and data mine Wiley content for non-commercial scholarly research related to specific projects; direct or indirect commercial purposes require prior written consent from Wiley	**NO** You may only access content for TDM for the purpose of non-commercial research	**NO** TDM rights are granted purely for internal non-commercial research purposes
Do I need to tell anyone what I am doing with TDM?	**MAYBE** You must provide TDM output and any related content to Elsevier on request	**NO** Not mentioned in policy	**NO** Not mentioned in policy	**NO** Not mentioned in policy
Are my TDM activities monitored?	**YES** You are required to use an API key; Elsevier maintains information about you which may be used in aggregate, and may be used to promote Elsevier offers to you	**YES** You are required to use an API key	**NO** No authentication is required when retrieving SpringerLink content for TDM	**NO** No authentication is required for CrossRef's TDM API

Note: Table B.1 is colored coded as **Ideal for TDM activities** | **Close to ideal for TDM activities** | **Some negative implications for TDM activities** | **Very restrictive for TDM activities**

Table C.1: Selected Prominent Publishers TDM Licensing Conditions (McNeice et al. (2017) FutureTDM: Reducing Barriers and Increasing Uptake of Text and Data Mining for Research Environments using a Collaborative Knowledge and Open Information Approach. https://project.futuretdm.eu/wp-content/uploads/2017/07/FutureTDM_D5.3-FutureTDM-practitioner-guidelines.pdf. Accessed 5 Nov 2020)

	Elsevier	Wiley	Springer	Emerald
Are there restrictions on how I can share new knowledge I generate as a result of TDM?	**YES** Results may be used by you and your company or institution, but may not be used in a way that would compete with existing Elsevier products; a specific proprietary notice must be used when sharing results externally	**YES** You may communicate TDM outputs as part of original non-commercial research, including in articles about that research	**NO** None mentioned in policy	**YES** There are no restrictions on where and how you can publish your research results, but you may not make results of TDM outputs available on any externally facing server or website
Am I required to share the outputs of my TDM research?	**YES** You must provide TDM output and any related content to Elsevier on request to ensure compliance with their agreement	**NO** None mentioned in policy	**NO** None mentioned in policy	**NO** None mentioned in policy
Can I support my results with experts from the content I have mined?	**YES** Limited to the query-dependent text of a maximum length of 200 characters surrounding the semantic entity matched, or bibliographic metadata; must include a DOI link to the original material	**YES** Limited to brief quotations as permitted under national copyright laws; must include a DOI link to the original material	**YES** Limited to quotations of up to 200 characters, 20 words, or one complete sentence; must include a DOI link to the original material	**YES** You can use snippets up to a maximum of 200 characters, provided these are referenced as you would reference a copyrighted work; you must contact Emerald if larger extracts are exceptionally required
Can I retain datasets for verifiability and reproducibility of my results?	**NO** You may not substantially retain the dataset; all Elsevier content stored for TDM must be permanently deleted on termination of the agreement	**NO** You must delete all Wiley content downloaded for TDM on completion of any specific TDM project, or on termination of the agreement with Wiley	**UNCLEAR** Not mentioned in policy	**NO** You may not substantially retain content; all copies of Emerald content that have been locally loaded for TDM must be destroyed on termination or expiry of this license
Do I have other responsibilities or obligations?	**YES** You are responsible for complying with data protection and relevant privacy laws when using or processing personal data	**YES** You must implement and maintain data security measures to protect Wiley content in line with international industry standards; you are responsible for complying with data protection and relevant privacy laws when using or processing personal data	**NO** None mentioned in policy	**NO** None mentioned in policy

Note: Table C.1 is colored coded as **Ideal for TDM activities** | **Close to ideal for TDM activities** | **Some negative implications for TDM activities** | **Very restrictive for TDM activities**

Conclusion

The rise of machine learning and artificial intelligence has ushered in a new era across various disciplines, including the field of library and information science

Text mining is a versatile tool that can be applied across various types of libraries, from schools to universities and special libraries, making it accessible to a wide range of information professionals

As we continue to embrace the power of artificial intelligence and data analytics, text mining emerges as a vital tool in the modern librarian's arsenal, unlocking new possibilities for the future of libraries and information science

Website: *https://manika-lamba.github.io/*

Email: *manika@illinois.edu*

Thank You

THANK YOU FOR ATTENDING

Submit a webinar proposal at https://www.asist.org/meetings-events/webinars/

webinars@asist.org

A copy of the recording and a follow-up survey will be emailed within 24 hours.

asis&t

Association for Information Science and Technology