# Chapter 1
# Introduction


## 1.1 The Basic Concepts

Compared with generalized data mining technology, beyond analyzing various document formats (such as doc/docx files, PDF files, and HTML files), the greatest challenge in text data mining lies in the analysis and modeling of unstructured natural language text content. Two aspects need to be emphasized here: first, text content is almost always unstructured, unlike databases and data warehouses, which are structured; second, text content is described by natural language, not purely by data, and other non-text formats such as graphics and images are not considered. Of course, it is normal for a document to contain tables and figures, but the main body in such documents is text. Therefore, text data mining is de facto an integrated technology of natural language processing (NLP), pattern classification, and machine learning (ML).

The so-called mining usually has the meanings of "discovery, search, induction and refinement." Since discovery and refinement are necessary, the target results being sought are often not obvious but hidden and concealed in the text or cannot be found and summarized in a large range. The adjectives "hidden" and "concealed" noted here refer not only to computer systems but also human users. However, in either case, from the user's point of view, the hope is that the system can directly provide answers and conclusions to the questions of interest, instead of delivering numerous possible search results for the input keywords and leaving users to analyze and find the required answers themselves as in the traditional retrieval system. Roughly speaking, text mining can be classified into two types. In the first, the user's questions are very clear and specific, but they do not know the answer to the questions. For example, users want to determine what kind of relationship someone has with some organizations from many text sources. The other situation is when the user only knows the general aim but does not have specific and definite questions. For example, medical personnel may hope to determine the regularity of some diseases and the related factors from many case records. In this case, they may not be referring to a specific disease or specific factors, and the relevant data in their



C. Zong et al., *Text Data Mining*, https://doi.org/10.1007/978-981-16-0100-2_1

entirety need to be mined automatically by system. Certainly, there is sometimes no obvious boundary between the two types.

Text mining technology has very important applications in many fields, such as the national economy, social management, information services, and national security. The market demand is huge. For example, government departments and management can timely and accurately investigate the people's will and understand public opinions by analyzing and mining microblogs, WeChat, SMSs (short message services), and other network information for ordinary people. In the field of finance or commerce, through the in-depth excavation and analysis of extensive written material, such as news reports, financial reports, and online reviews, text mining can predict the economic situation and stock market trends for a certain period. Electronic product enterprises can acquire and evaluate their product users or market reactions at any time and capture data support for further improving product quality and providing personalized services. For national security and public security departments, text data mining technology is a useful tool for the timely discovery of social instability factors and effectively controlling the current situation. In the field of medicine and public health, many phenomena, regularities, and conclusions can be found by analyzing medical reports, cases, records, and relevant documents and materials.

Text mining, as a research field crossing multiple technologies, originated from single techniques such as text classification, text clustering, and automatic text summarization. In the 1950s, text classification and clustering emerged as an application of pattern recognition. At that time, research was mainly focused on the needs of books and on information classification, and classification and clustering are, of course, based on the topics and contents of texts. In 1958, H.P. Luhn proposed the concept of automatic summarization (Luhn 1958), which added new content to the field of text mining. In the late 1980s and early 1990s, with the rapid development and popularization of Internet technology, demand for new applications has promoted the continuous development and growth of this field. The US government has funded a series of research projects on information extraction, and in 1987, the US Defense Advanced Research Projects Agency (DARPA) initiated and organized the first Message Understanding Conference (MUC[1]) to evaluate the performance of this technology. In the subsequent 10 years, seven consecutive evaluations have made information extraction technology a research hot spot in this field. Next, a series of social media-oriented text processing technologies, such as text sentiment analysis, opinion mining, and topic detection and tracking, emerged and developed rapidly. Today, this technical field is growing rapidly not only in theory and method but also in the form of system integration and applications.

[1] https://www-nlpir.nist.gov/related_projects/muc/.

## 1.2 Main Tasks of Text Data Mining

As mentioned above, text mining is a domain that crosses multiple technologies involving a wide range of content. In practical applications, it is usually necessary to combine several related technologies to complete an application task, and the execution of mining technology is usually hidden behind the application system. For example, a question and answering (Q&A) system often requires several links, such as question parsing, knowledge base search, inference and filtering of candidate answers, and answer generation. In the process of constructing a knowledge base, key technologies such as text clustering, classification, named entity recognition, relationship extraction, and disambiguation are indispensable. Therefore, text mining is not a single technology system but is usually an integrated application of several technologies. The following is a brief introduction to several typical text mining technologies.

**(1) Text Classification**

Text classification is a specific application of pattern classification technology. Its task is to divide a given text into predefined text types. For example, according to the Chinese Library Classification (5-th Edition),[2] all books are divided into 5 categories and 22 subcategories. On the first page of www.Sina.com,[3] the content is divided into the following categories: news, finance, sports, entertainment, cars, blog, video, house and property, etc. Automatically classifying a book or an article into a certain category according to its content is a challenging task.

Chapter 5 of this book introduces text classification techniques in detail.

**(2) Text Clustering**

The purpose of text clustering is to divide a given text set into different categories. Generally, different results can be clustered based on different perspectives. For example, based on the text content, the text set can be clustered into news, culture and entertainment, sports or finance, and so on, while based on the author's tendency, it can be grouped into positive categories (positive views with positive and supportive attitudes) and negative categories (negative views with negative and passive attitudes).

The basic difference between text clustering and text classification is that classification predefines the number of categories and the classification process automatically classifies each given text into a certain category and labels it with a category tag. Clustering, by contrast, does not predefine the number of categories, and a given document set is divided into categories that can be distinguished from each other based on certain standards and evaluation indices. Many similarities exist between text clustering and text classification, and the adopted algorithms and

[2] https://baike.baidu.com/item/ 中国图书馆图书分类法/1919634?fr=aladdin.

[3] https://www.sina.com.cn/.

models have intersections, such as models of text representation, distance functions, and K-means algorithms.

Chapter 6 of this book introduces text clustering techniques in detail.

**(3) Topic Model**

In general, every article has a topic and several subtopics, and the topic can be expressed by a group of words that have strong correlation and that basically share the same concepts and semantics. We can consider each word as being associated with a certain topic with a certain probability, and in turn, each topic selects a certain vocabulary with a certain probability. Therefore, we can give the following simple formula:

$$p(word_i|document_j) = \sum_k p(word_i|topic_k) \times p(topic_k|document_j) \qquad (1.1)$$

Thus, the probability of each word appearing in the document can be calculated.

To mine the topics and concepts hidden behind words in text, people have proposed a series of statistical models called topic models.

Chapter 7 of this book introduces the topic model in detail.

**(4) Text Sentiment Analysis and Opinion Mining**

Text sentiment refers to the subjective information expressed by a text's author, that is, the author's viewpoint and attitude. Therefore, the main tasks of text sentiment analysis, which is also called text orientation analysis or text opinion mining, include sentiment classification and attribute extraction. Sentiment classification can be regarded as a special type of text classification in which text is classified based on subjective information such as views and attitudes expressed in the text or judgments of its positive or negative polarity. For example, after a special event (such as the loss of communication with Malaysia Airlines MH370, UN President Ban Ki-moon's participation in China's military parade commemorating the 70th anniversary of the victory of the Anti-Fascist War or talks between Korean and North Korean leaders), there is a high number of news reports and user comments on the Internet. How can we automatically capture and understand the various views (opinions) expressed in these news reports and comments? After a company releases a new product, it needs a timely understanding of users' evaluations and opinions (tendentiousness) and data on users' age range, sex ratio, and geographical distribution from their online comments to help inform the next decisions. These are all tasks that can be completed by text sentiment analysis.

Chapter 8 of this book introduces text sentiment analysis and opinion mining techniques.

**(5) Topic Detection and Tracking**

Topic detection usually refers to the mining and screening of text topics from numerous news reports and comments. Those topics that most people care about, pay attention to, and track are called *hot topics*. Hot topic discovery, detection,

and tracking are important technological abilities in public opinion analysis, social media computing, and personalized information services. The form of their application varies, for example, *Hot Topics Today* is a report on what is most attracting readers' attention from all the news events on that day, while *Hot Topics 2018* lists the top news items that attracted the most attention from all the news events throughout 2018 (this could also be from January 1, 2018, to a different specified date).

Chapter 9 of this book introduces techniques for topic detection and tracking.

**(6) Information Extraction**

Information extraction refers to the extraction of factual information such as entities, entity attributes, relationships between entities, and events from unstructured and semistructured natural language text (such as web news, academic documents, and social media), which it forms into structured data output (Sarawagi 2008). Typical information extraction tasks include named entity recognition, entity disambiguation, relationship extraction, and event extraction.

In recent years, biomedical/medical text mining has attracted extensive attention. Biomedical/medical text mining refers to the analysis, discovery, and extraction of text in the fields of biology and medicine, for example, research from the biomedical literature to identify the factors or causes related to a certain disease, analysis of a range of cases recorded by doctors to find the cause of certain diseases or the relationship between a certain disease and other diseases, and other similar uses. Compared with text mining in other fields, text mining in the biomedical/medical field faces many special problems, such as a multitude of technical terms and medical terminology in the text, including idioms and jargon used clinically, or proteins named by laboratories. In addition, text formats vary greatly based on their different source, such as medical records, laboratory tests, research papers, public health guidelines, or manuals. Unique problems faced in this field are how to express and utilize common knowledge and how to obtain a large-scale annotation corpus.

Text mining technology has also been a hot topic in the financial field in recent years. For example, from the perspective of ordinary users or regulatory authorities, the operational status and social reputation of a financial enterprise are analyzed through available materials such as financial reports, public reports, and user comments on social networks; from the perspective of an enterprise, forewarnings of possible risks may be found through the analysis of various internal reports, and credit risks can be controlled through analysis of customer data.

It should be noted that the relation in information extraction usually refers to some semantic relation between two or more concepts, and relation extraction automatically discovers and mines the semantic relation between concepts. Event extraction is commonly used to extract the elements that make up the pairs of events in a specific domain. The "event" mentioned here has a different meaning from that used in daily life. In daily life, how people describe events is consistent with their understanding of events: they refer to when, where, and what happened. The thing that happened is often a complete story, including detailed descriptions of causes, processes, and results. By contrast, in even extraction, the "event" usually

refers to a specific behavior or state expressed by a certain predicate framework. For example, "John meets Mary" is an event triggered by the predicate "meet." The event understood by ordinary people is a story, while the "event" in event extraction is just an action or state.

Chapter 10 of this book introduces information extraction techniques.

**(7) Automatic Text Summarization**

Automatic text summarization or automatic summarization, in brief, refers to a technology that automatically generates summaries using natural language processing methods. Today, when information is excessively saturated, automatic summarization technology has very broad applications. For example, an information service department needs to automatically classify many news reports, form summaries of some (individual) event reports (report), and recommend these reports to users who may be interested. Some companies or supervisory departments want to know roughly the main content of statements (SMS, microblog, WeChat, etc.) published by some user groups. Automatic summarization technology is used in these situations.

Chapter 11 of this book introduces automatic text summarization techniques.

## 1.3 Existing Challenges in Text Data Mining

Study of the techniques of text mining is a challenging task. First, the theoretical system of natural language processing has not yet been fully established. At present, text analysis is to a large extent only in the "processing" stage and is far from reaching the level of deep semantic understanding achieved by human beings. In addition, natural language is the most important tool used by human beings to express emotions, feelings, and thoughts, and thus they often use euphemism, disguise, or even metaphor, irony, and other rhetoric means in text. This phenomenon is obvious, especially in Chinese texts, which presents many special difficulties for text mining. Many machine learning methods that can achieve better results in other fields, such as image segmentation and speech recognition, are often difficult to use in natural language processing. The main difficulties confronted in text mining include the following aspects.

**(1) Noise or ill-formed expressions present great challenges to NLP**

Natural language processing is usually the first step in text mining. The main data source for text mining processing is the Internet, but when compared with formal publications (such as all kinds of newspapers, literary works, political and academic publications, and formal news articles broadcast by national and local government television and radio stations), online text content includes large ill-formed expressions. According to a random sampling survey of Internet news texts conducted by Zong (2013), the average length of Chinese words on the Internet is approximately 1.68 Chinese characters, and the average length of sentences is

47.3 Chinese characters, which are both shorter than the word length and sentence length in the normal written text. Relatively speaking, colloquial and even ill-formed expressions are widely used in online texts. This phenomenon is common, especially in online chatting, where phrases such as "up the wall," "raining cats and dog," and so on can be found. The following example is a typical microblog message:

> *//@XXXX://@YYYYY: Congratulations to the first prospective members of the Class of 2023 offered admission today under Stanford's restrictive early action program.* https://stanford.io/2E7cfGF#Stanford2023

The above microblog message contains some special expressions. Existing noise and ill-formed language phenomena greatly reduce the performance of natural language processing systems. For example, a Chinese word segmentation (CWS) system trained on a corpora of normal texts such as the *People's Daily* and the *Xinhua Daily* can usually achieve an accuracy rate of more than 95%, even as high as 98%, but its performance on online text immediately drops below 90%. According to the experimental results of (Zhang 2014), using the character-based Chinese word segmentation method based on the maximum entropy (ME)classifier, when the dictionary size is increased to more than 1.75 million (including common words and online terms), the performance of word segmentation on microblog text as measured by the $F_1$-measure metric can only reach approximately 90%. Usually, a Chinese parser can reach approximately 87% or more on normal text, but on online text, its performance decreases by an average of 13% points (Petrov and McDonald 2012). The online texts addressed by these data are texts on the Internet and do not include the texts of dialogues and chats in microblogs, Twitter, or WeChat.

**(2) Ambiguous expression and concealment of text semantics**

Ambiguous expressions are common phenomena in natural language texts, for example, the word "bank" may refer to a financial bank or a river bank. The word "Apple" may refer to the fruit or to a product such as an Apple iPhone or an Apple Computer, a Mac, or Macintosh. There also exist many phenomena of syntactic ambiguity. For example, the Chinese sentence "关于(guanyu, about)鲁迅(Lu Xun, a famous Chinese writer)的(de, auxiliary word)文章(wenzhang, articles)" can be understood as "关于【鲁迅的文章】 (about articles of Lu Xun)" or "【关于鲁迅】的文章 (articles about Lu Xun)." Similarly, the English sentence "I saw a boy with a telescope" may be understood as "I saw [a boy with a telescope]," meaning I saw a boy who had a telescope, or "[I saw a boy] with a telescope" meaning I saw a boy by using a telescope. The correct parsing of these ambiguous expressions has become a very challenging task in NLP. However, regrettably, there are no effective methods to address these problems, and a large number of intentionally created "special expressions/tokens" such as Chinese "words" "木有 (no)," "坑爹 (cheating)," and '奥特 (out/out-of-date)" and English words "L8er (later)," "Adorbs (adorable)," and "TL;DR (Too long, didn't read)" appear routinely in online dialogue texts.

Sometimes, to avoid directly identifying certain events or personages, the speaker will turn a sentence around deliberately, for example, asking "May I know the age of the ex-wife of X's father's son?".

Please look at the following news report:

> Mr. Smith, who had been a policeman for more than 20 years, had experienced a multitude of hardships, had numerous achievements and been praised as a hero of solitary courage. However, no one ever thought that such an steely hero, who had made addicted users frightened and filled them with trepidation, had gone on a perilous journey for a small profit and shot himself at home last night in hatred.

For most readers, it is easy to understand the incident reported by this news item without much consideration. However, if someone asks the following question to a text mining system based on this news *What kind of policeman is Mr. Smith?* and *Is he dead?* it will be difficult for any current system to give a correct answer. The news story never directly expresses what kind of policeman Mr. Smith is but uses *addicted users* to hint to readers that he is an antidrug policeman and uses "shot himself" to show that he has committed suicide. This kind of information hidden in the text can only be mined by technology with deep understanding and reasoning, which is very difficult to achieve.

**(3) Difficult collection and annotation of samples**

At present, the mainstream text mining methods are machine learning methods based on large-scale datasets, including the traditional statistical machine learning method and the deep learning (DL) method. These require a large-scale collection of labeled training samples, but it is generally very difficult to collect and annotate such large-scale samples. On the one hand, it is difficult to obtain much online content because of copyright or privacy issues, which prohibit publication opening and sharing. On the other hand, even when data are easy to obtain, processing these data is time-consuming and laborious because they often contain considerable noise and garbled messages, they lack a uniform format, and there is no standard criterion for data annotation. In addition, the data usually belong to a specific field, and help from experts in that specific domain is necessary for annotation. Without help from experts, it is impossible to provide high-quality annotation of the data. If the field changes, the work of data collection, processing, and annotation will have to start again, and many ill-formed language phenomena (including new online words, terms and ungrammatical expressions) vary with changing domains and over time, which greatly limits expansion of the data scale and affects the development of text mining technology.

**(4) Hard to express the purpose and requirements of text mining**

Text mining is unlike other theoretical problems, wherein objective functions are clearly established and then ideal answers obtained by optimizing functions and solving the extremum. In many cases, we do not know what the results of text mining will be or how to use mathematical models to describe the expected results and conditions clearly. For example, we can extract frequently used "hot" words from some text that can represent the themes and stories of these texts, but how to organize them into story outlines (summaries) expressed in fluent natural languages is not an easy task. As another example, we know that there are some regular patterns

and correlations hidden in many medical cases, but we do not know what regular patterns and correlations exist and how to describe them.

**(5) Unintelligent methods of semantic representation and computation model**

Effectively constructing semantic computing models is a fundamental challenge that has puzzled the fields adopting NLP for a long time. Since the emergence of deep learning methods, word vector representation and various computing methods based on word vectors have played an important role in NLP. However, semantics in natural language are different from pixels in images, which can be accurately represented by coordinates and grayscales. Linguists, computational linguists, and scholars engaged in artificial intelligence research have been paying close attention to the core issues of how to define and represent the semantic meanings of words and how to achieve combination computing from lexical semantics to phrase, sentence, and, ultimately, paragraph and discourse semantics. To date, there are no convincing, widely accepted and effective models or methods for semantic computing. At present, most semantic computing methods, including many methods for word sense disambiguation, word sense induction based on topic models, and word vector combinations, are statistical probability-based computational methods. In a sense, statistical methods are "gambling methods" that choose high probability events. In many cases, the events with the highest probability will become the final selected answer. In fact, this is somewhat arbitrary, subjective, or even wrong. Since the model for computing probability is based on samples taken by hand, the actual situation (test set) may not be completely consistent with the labeled samples, which inevitably means that some small probability events become "fishes escaping from the net." Therefore, the gambling method, which is always measured by probability, can solve most of the problems that are easy to count but cannot address events that occur with small probability, are hard to find, and occur with low frequency. Those small probability events are always difficult problems to solve, that is, they are the greatest "enemy" faced in text mining and NLP.

In summary, text mining is a comprehensive application technology that integrates the challenges in various fields, such as NLP, ML, and pattern classification, and is sometimes combined with technologies to process graphics, images, videos, and so on. The theoretical system in this field has not yet been established, its prospect for application is extremely broad, and time is passing: text mining will surely become a hot spot for research and will grow rapidly with the development of related technologies.

## 1.4 Overview and Organization of This Book

As mentioned in Sect. 1.1, text mining belongs to the research field combining NLP, pattern classification, ML, and other related technologies. Therefore, the use and development of technical methods in this field also change with the development and transition of related technologies.

Reviewing the history of development, which covers more than half a century, text mining methods can be roughly divided into two types: knowledge engineering-based methods and statistical learning methods. Before the 1980s, text mining was mainly based on knowledge engineering, which was consistent with the historical track of rule-based NLP and the mainstream application of expert systems dominated by syntactic pattern recognition and logical reasoning. The basic idea of this method is that experts in a domain collect and design logical rules manually for the given texts based on their empirical knowledge and common sense and then the given texts are analyzed and mined through inference algorithms using the designed rules. The advantage of this method is that it makes use of experts' experience and common sense, there is a clear basis for each inference step, and there is a good explanation for the result. However, the problem is that it requires extensive human resources to deduce and summarize knowledge based on experience, and the performance of the system is constrained by the expert knowledge base (rules, dictionaries, etc.). When the system needs to be transplanted to the new fields and tasks, much of the experience-based knowledge cannot be reused, so that usually, much time is needed to rebuild a system. Since the later 1980s, and particularly after 1990, with the rapid development and broad application of statistical machine learning methods, text mining methods based on statistical machine learning obtained obvious advantages in terms of accuracy and stability and do not need to consume the same level of human resources. Especially in the era of big data on the Internet, given massive texts, manual methods are obviously not comparable to statistical learning methods in terms of speed, scale, or coverage when processing data. Therefore, statistical machine learning methods are gradually becoming the mainstream in this field. Deep learning methods, or neural network-based ML methods, which have emerged in recent years, belong to the same class of methods, which can also be referred to as data-driven methods. However, statistical learning methods also have their own defects; for example, supervised machine learning methods require many manually annotated samples, while unsupervised models usually perform poorly, and the results from the system for both supervised and unsupervised learning methods lack adequate interpretability.

In general, knowledge engineering-based methods and statistical learning methods have their own advantages and disadvantages. Therefore, in practical application, system developers often combine the two methods, using the feature engineering method in some technical modules and the statistical learning method in the others to help the system achieve the strongest performance possible through the fusion of the two methods. Considering the maturity of technology, knowledge engineering-based methods are relatively mature, and their performance ceiling is predictable. For statistical learning methods, with the continuous improvement of existing models and the continuous introduction of new models, the performance of models and algorithms is gradually improving, but there is still great room for improvement, especially in large-scale data processing. Therefore, statistical learning methods are in the ascendant. These are the reasons this book focuses on statistical learning methods.

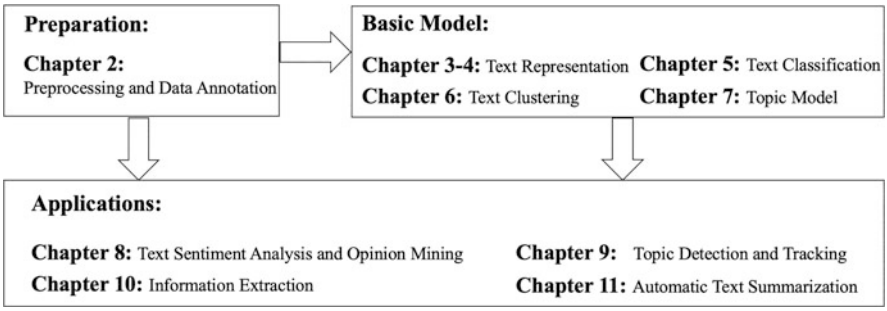


**Fig. 1.1** An overview of this book

This book mainly introduces the basic methods and concepts behind text mining but does not get involved in the implementation details of specific systems, nor does it overly elaborate the task requirements and special problems of specific applications. For example, text mining technology in the biomedical and financial fields has attracted much attention in recent years, and many related technologies and resources need to be developed for these fields, such as domain knowledge bases, annotation tools, and annotation samples for domain-related data. The authors hope that the basic methods and ideas introduced in this book have a certain universality and commonality. Once readers know these fundamental methods, they can expand them and implement the system oriented to their specific task requirements.

The remaining nine chapters are organized along the following lines, as shown in Fig. 1.1.

Chapter 2 introduces the methods for data preprocessing. Data preprocessing is the preparation stage before all subsequent models and algorithms are implemented; for example, word segmentation for the Chinese, Japanese, Vietnamese, and other possible languages requires word segmentation. In most online texts, there is much noise and many ill-formed expressions. If these data are not preprocessed well, the subsequent modules will be badly affected, and it will be difficult to achieve the expected final results; indeed, it may be that the model cannot even run. The text representation described in Chaps. 3 and 4 is the basis of the models used in the subsequent chapters. If the text cannot be accurately represented, it is impossible to obtain better results using any of the mathematical models and algorithms introduced in these chapters. The text classification methods introduced in Chap. 5, the text clustering algorithms introduced in Chap. 6, and the topic models introduced in Chap. 7 are the theoretical foundations of other text mining technologies, in a sense, because classification and clustering are the two most fundamental and core problems of pattern recognition and are the two most commonly used methods in machine learning and statistical natural language processing. Most of the models and methods introduced in the following chapters can be treated as classification and clustering problems or can be solved by adopting the concepts of classification or clustering. Therefore, Chaps. 5–7 can be regarded as the theoretical foundations

or basic models of the book. In addition, it should be noted that text classification, clustering, and topic models are sometimes used as the sole specific application in some tasks.

Chapters 8–11 can be regarded as an application technology for text mining. A specific task can be performed by one model of can be jointly carried out by several models and algorithms. In most practical applications, the latter method is adopted. For example, text mining tasks in the field of medicine usually involve the techniques of text automatic classification and clustering, topic modeling, information extraction, and automatic summarization, while public opinion analysis tasks for social networks may involve text classification, clustering, topic modeling, topic detection and tracking, sentiment analysis, and even automatic summarization.

With the rapid development and popularization of Internet and mobile communication technologies, requirements may emerge for new applications and new technologies to be applied to text mining. However, we believe that regardless of the application requirements and regardless of the technology behind the new name, new methods of text representation and category distance measurement, and new implementation methods and models (such as end-to-end neural network models), the basic ideas behind clustering and classification and their penetration and application in various tasks will not undergo fundamental changes. This belief is the so-called all things remain essentially the same.

## 1.5 Further Reading

The following chapters of this book will introduce text mining methods for different tasks and explain the objectives, solutions, and implementation methods. As the beginning of the book, this chapter mainly introduces the basic concepts and challenges of text mining. For a detailed explanation of the concept of data mining, readers can refer to the following literature: (Han et al. 2012; Cheng and Zhu 2010; Li et al. 2010b; Mao et al. 2007). Wu et al. (2008) introduced ten classical algorithms in the field of data mining. Aggarwal (2018) is a relatively comprehensive book introducing text mining technologies. By contrast, readers will find that text mining is regarded as a specific application of machine learning technology in Aggarwal's book, which focuses on discussing text information processing from the perspective of machine learning methods (especially statistical machine learning methods) without the involvement of deep learning and neural network-based methods. Moreover, only traditional statistical methods are used in various text mining tasks, such as text classification, sentiment analysis, and opinion mining, and few related works based on deep learning methods have been introduced in recent years. However, in this book, we regard text mining as the practical application of NLP technology because text is the most important mode of presentation for natural language. Since it is necessary to mine the information needed by users from text, NLP technology remains indispensable. Therefore, this book is driven by task requirements and illustrates the basic principles of text

mining models and algorithms through examples and descriptions of processes from the perspective of NLP. For example, in the text representation chapter, text representation and modeling methods based on deep learning are summarized based on the granularity of words, sentences, and documents. In the text mining tasks that follow, in addition to introducing traditional classical methods, deep learning methods, which are highly recommended in recent years, are given special attention.

If Zong (2013) is treated as a foundation or teaching material for the introduction of NLP technologies, then this book is an introduction to the application of NLP technologies. The former mainly introduces the basic concepts, theories, tools, and methods of NLP, while this book focuses on the implementation methods and classical models of NLP application systems.

Other books elaborate on specific technologies of text mining and have good reference value. For example, Liu (2011, 2012, 2015) introduce the concepts and technologies related to web data mining, sentiment analysis, and opinion mining in detail; Marcu (2000) and Inderjeet (2001) provide detailed descriptions of automatic summarization technology, especially the introduction of early summarization technology. Relevant recommendations will be introduced in the "Further Reading" section in each following chapter.

In addition, it should be noted that the authors of this book consider the readers as already having a foundation in pattern recognition and machine learning by default. Therefore, many basic theories and methods are not introduced in detail, their detailed derivation is omitted, and they are only cited as tools. If readers want to know the detailed derivation process, we recommend that they read the following books: (Li 2019; Yu 2017; Zhang 2016; Zhou 2016), etc.

## Exercises

**1.1** Please summarize the difference between KDD and text data mining.

**1.2** Please read an article on natural language processing and learn how to compute the metrics of precision, recall, and the $F_1$-measure to evaluate the performance of a natural language processing tool.

**1.3** Please make a comparison table to show the strengths and weaknesses of rule-based methods and statistical methods.

**1.4** Please give some examples for the use of text data mining techniques in real life. Specifically, explain the inputs (source text) and the outputs (what results should we expect to see?).