CHAPTER 2

The Seven Practice Areas of Text Analytics

CONTENTS

PREAMBLE

Presently, text mining is in a loosely organized set of competing technologies that function as analytical "city-states" with no clear dominance among them. To further complicate matters, different areas of text mining are in different stages of maturity. Some technology is easily accessible by practitioners today via commercial software (some of which is included with this book), while other areas are only now emerging from academia into practical use.

We can relate these technologies to seven different *practice areas* in text mining that are covered in the chapters in this book. In summary, this book is strongest in the practice area of document classification, solid in concept extraction and document clustering, reasonably useful on web mining, light on information extraction and natural language processing, and almost silent on the (most popular) practice area of search and information retrieval.

The unifying theme behind each of these technologies is the need to "turn text into numbers" so that powerful analytical algorithms can be applied to large document databases. Converting text into

Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications. DOI: 10.1016/B978-0-12-386979-1.00002-5

a structured, numerical format and applying analytical algorithms both require knowing how to use and combine techniques for handling text, ranging from individual words to documents to entire document databases.

Next, we provide a decision tree to help you determine which practice area is appropriate to satisfy your needs. Finally, we provide tables to relate the practice areas to appropriate technologies and show which chapter in this book deals with that subject area. That is the most organization that we can impose on the current disordered state of text mining technology. Our goal in this book is to provide an introduction to each of the seven practice areas and cover in depth only those areas that are *accessible for nonexperts*. We will follow that theme in Part I of the book to provide you with the basics you need to perform the tutorials. Very quickly, you will be learning by doing.

WHAT IS TEXT MINING?

Text mining and *text analytics* are broad umbrella terms describing a range of technologies for analyzing and processing semistructured and unstructured text data. The unifying theme behind each of these technologies is the need to "turn text into numbers" so powerful algorithms can be applied to large document databases. Converting text into a structured, numerical format and applying analytical algorithms require knowing how to both use and combine techniques for handling text, ranging from individual words to documents to entire document databases.

To date, text mining has resisted a more comprehensive definition because the field is emerging out of a group of related but distinct disciplines, as described in Chapter 1. Figure 2.1 shows the six other major fields that intersect with text mining. Due to the breadth and disparity of the contributing disciplines, it can be difficult even for text mining experts to concisely characterize. Text mining is something of the "Wild West" of analytics, since there are a number of competing technologies with no clear dominance among them (but much braggadocio). To further complicate matters, different areas of text mining are in different stages of maturity.

Our goal in this chapter is to bring clarity to the field by providing a framework and vocabulary for discussing the seven different *practice areas* within text mining. Due to the breadth of text mining, no single book can hope to fully cover the field. Our target audience is nonspecialist text-mining practitioners—analysts who have the technical expertise to handle challenges involving text but have limited experience or background with text processing. Consequently, this book provides an introduction to each of the seven practice areas, but it covers in depth only those areas that are *accessible for nonexperts*, yet not ubiquitous. We reference other resources in areas that are less mature or require additional expertise or, in the case of search technology, are already very useable in their current incarnations.

There are seven different text mining practice areas—that is, seven very different things that a client, speaker, boss, or colleague could have in mind when talking about text mining. The seven practice areas are defined in Figure 2.1. This book is strongest in the practice area of *document classification*, solid in *concept extraction* and *document clustering*, reasonably useful on *web mining*, light on *information extraction* and *natural language processing*, and almost silent on the (most popular) practice area of *search and information retrieval*.

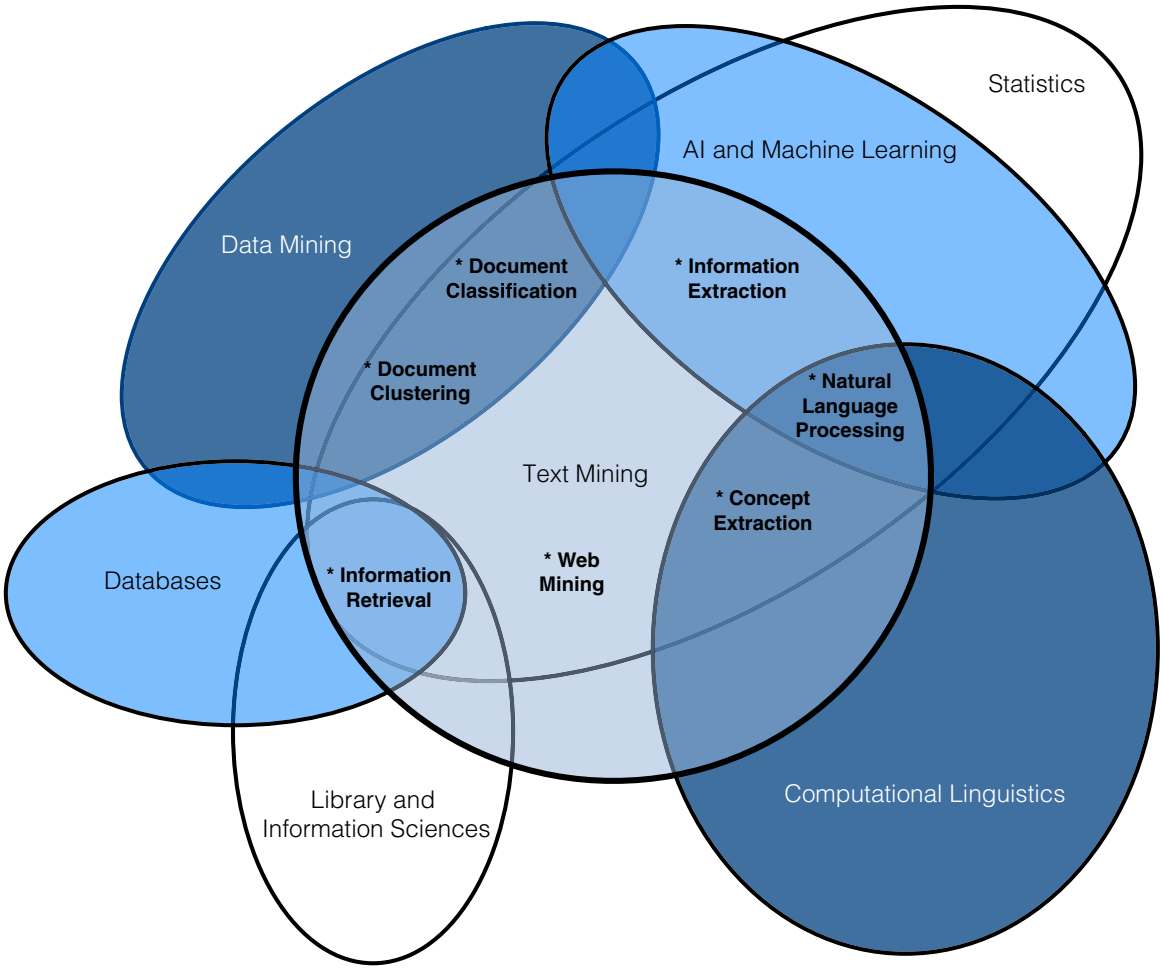

FIGURE 2.1

A Venn diagram of the intersection of text mining and six related fields (shown as ovals), such as data mining, statistics, and computational linguistics. The seven text mining practice areas exist at the major intersections of text mining with its six related fields.

THE SEVEN PRACTICE AREAS OF TEXT ANALYTICS

Text mining can be divided into seven practice areas, based on the unique characteristics of each area. Though distinct, these areas are highly interrelated; a typical text mining project will require techniques from multiple areas. This book views text mining through the eyes of practitioners. Instead of emphasizing the academic or technical differentiators between the practice areas, our focus is on guiding readers toward answers to the problem they are facing. We have inductively identified seven practice areas based on five resource and goal questions that text mining practitioners must answer

when facing a new problem. The five questions will be defined soon; meanwhile, the seven practice areas are as follows:

1. **Search and information retrieval (IR):** Storage and retrieval of text documents, including search engines and keyword search.
2. **Document clustering:** Grouping and categorizing terms, snippets, paragraphs, or documents, using data mining clustering methods.
3. **Document classification:** Grouping and categorizing snippets, paragraphs, or documents, using data mining classification methods, based on models trained on labeled examples.
4. **Web mining:** Data and text mining on the Internet, with a specific focus on the scale and interconnectedness of the web.
5. **Information extraction (IE):** Identification and extraction of relevant facts and relationships from unstructured text; the process of making structured data from unstructured and semistructured text.
6. **Natural language processing (NLP):** Low-level language processing and understanding tasks (e.g., tagging part of speech); often used synonymously with computational linguistics.
7. **Concept extraction:** Grouping of words and phrases into semantically similar groups.

These seven practice areas exist at the key intersections of text mining and the six major other fields that contribute to it. Figure 2.1 depicts, as a Venn diagram, the overlap of the seven fields of text mining, data mining, statistics, artificial intelligence and machine learning, computational linguistics, library and information sciences, and databases; it also locates the seven practice areas at their key intersections. For example, the practice area of text classification (the most thoroughly covered in this book) draws from the field of data mining, and the practice area of information retrieval (most popular, but least covered in this book) draws from the two fields of databases and library and information sciences. Tables 2.2 and 2.3 provide alternative methods for identifying the practice areas based on algorithms and desired products.

FIVE QUESTIONS FOR FINDING THE RIGHT PRACTICE AREA

Figure 2.2 is a decision tree depicting how answering a few straightforward questions can direct you to the appropriate text mining solution. Five questions—only two to four of which need to be answered, depending on your problem—best split the major branches of text mining. They identify the seven practice areas, which are depicted as the leaf nodes of the tree highlighted in blue in Figure 2.2. Rarely will a single pass through the tree solve any text mining problem. A text mining solution usually consists of multiple passes through the data at different levels of processing—starting with raw input documents and moving toward fully encoded text. At each step, a group of questions must be answered to determine the appropriate processing task. These questions are detailed in the following sections. In addition, Table 2.1 lists typical desired outcomes for text mining algorithms and their corresponding practice areas.

Question 1: Granularity

This question finds the desired granularity (level of detail of focus) of the text mining task. While documents and words are both integral to successful text mining, an algorithm virtually always emphasizes one or the other. Note that in this book we use the term *document* to describe the unit of text

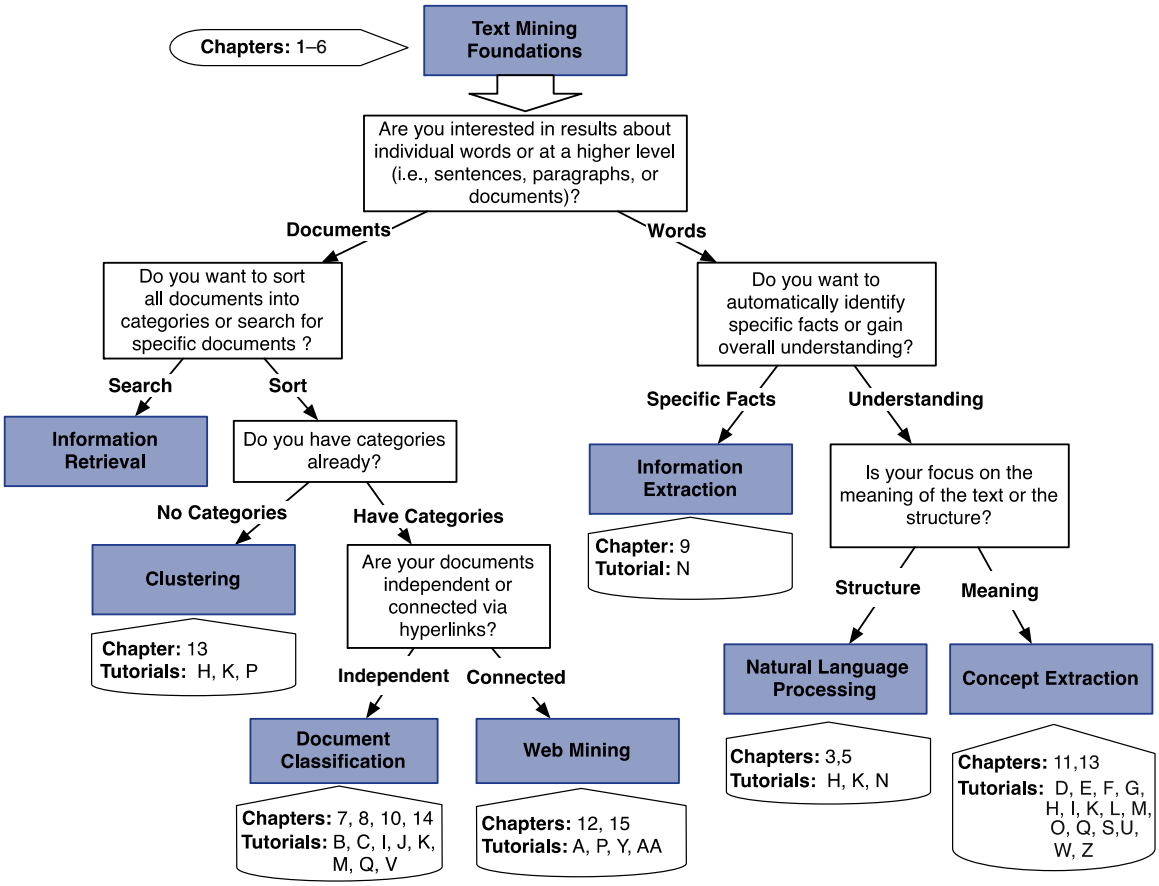

FIGURE 2.2

A decision tree for finding the right text mining practice area by answering 2 to 4 questions about your text resources and project goals.

under analysis. This is a broader definition than is usually employed. In practice, this could be mean typical documents, paragraphs, sentences, "tweets" on social media, or other defined sections of text.

To determine the granularity of your text mining problem, ask yourself about the desired outcome: Is it about characterizing or grouping together words or documents? This is the biggest division between classes of text mining algorithms.

Question 2: Focus

Whether you are interested in document or words, the next question in the decision tree of Figure 2.2 regards the focus of the algorithm: Are you interested in finding specific words and documents or characterizing the entire set? The two practice areas separated by this question—search and information extraction—both concentrate on identifying specific pieces of information within a document database, whereas the other solutions attempt to cluster or partition the space.

Table 2.1 Text Mining Topics and Related Practice Areas

Topic	Practice Area (Number)
Keyword search	Search and information retrieval (1)
Inverted index	Search and information retrieval (1)
Document clustering	Document Clustering (2)
Document similarity	Document Clustering (2)
Feature selection	Document classification (3)
Sentiment analysis	Document classification (3); Web mining (4)
Dimensionality reduction	Document classification (3)
eDiscovery	Document classification (3)
Web crawling	Web mining (4)
Link analytics	Web mining (4)
Entity extraction	Information extraction (5)
Link extraction	Information extraction (5)
Part of speech tagging	Natural language processing (6)
Tokenization	Natural language processing (6)
Question answering	Natural language processing (6), Search and information retrieval (1)
Topic modeling	Concept extraction (7)
Synonym identification	Concept extraction (7)

Table 2.2 Common Text Mining Algorithms and the Corresponding Practice Area

Algorithm	Area	Chapters	Tutorials
Naïve Bayes	Document classification	7, 15	F, X, Z
Conditional random fields	Information extraction	9	
Hidden Markov models	Information extraction	9	
k-means	Clustering	8, 13	F, H, L, O
Singular value decomposition (SVD)	Document classification, clustering	8, 10	K, L, O, Y
Logistic regression	Document classification	7, 8	Q
Decision trees	Document classification	7, 8	B, J, K
Neural network	Document classification	8	I
Support vector machines	Document classification	7	R, Z
MARSplines	Document classification		X, Y
Link analysis	Concept extraction	8	See*
k-nearest neighbors	Document classification	8	X, Z
Word clustering	Concept extraction	8, 13	D, E, G, M, P, Q, U
Regression	Classification		A

**See Tutorial Y in* Handbook of Statistical Analysis and Data Mining Applications, *by Nisbet, Elder, and Miner.*

Table 2.3 Finding a Practice Area Based on the Desired Product of Text Mining

Desired Product	Practice Area
Linguistic structure	Natural language processing
Topic/category assignment	Document classification
Documents that match keywords	Information retrieval
A structured database	Information extraction
"Needles in a haystack"	Document classification
List of synonyms	Concept extraction
Marked sentences	Natural language processing
Understanding of microblogs	Web mining
Similar documents	Clustering

Question 3: Available Information

If you are interested in documents, the next question regards the available information at the time of analysis. This is equivalent to the supervised/unsupervised question from data mining. A supervised algorithm requires training data with an answer (outcome label) for positive and negative examples of the classes you're trying to model (such as distinguishing "interesting versus not interesting" articles for an analyst studying a specialized topic). An unsupervised algorithm does not require any labeled data, and it can be applied to any data set without any available information at analysis time. Supervised learning is much more powerful when possible to use—that is, when enough example cases with target outcomes are known.

Question 4: Syntax or Semantics

If you are interested in words, the major question is about syntax or semantics. Syntax is about what the words "say," while semantics is about what the words "mean." Because natural language is so fluid and complex, semantics is the harder problem. However, there are text mining algorithms to address both areas.

Question 5: Web or Traditional Text

The rise of the Internet (including blogs, Twitter, and Facebook) is largely responsible for the prominence that text mining holds today by making available a vast number of previously unreachable text documents. The structure and style of web documents provide both unique opportunities and challenges when compared to nonweb documents. Though many of the algorithms are theoretically the same for web and traditional text, the scale of the web and its unique structural characteristics justify defining two different categories.

THE SEVEN PRACTICE AREAS IN DEPTH

We have categorized text mining into seven subdisciplines, based on the answers to the preceding questions:

1. Search and information retrieval
2. Document clustering

3. Document classification
4. Web mining
5. Information extraction
6. Natural language processing
7. Concept extraction

The following are brief descriptions of the problems faced in each practice area, a guide to the resources available in this book, and references to other resources if you wish to delve deeper into any of the areas.

Search and Information Retrieval

Search and information retrieval covers indexing, searching, and retrieving documents from large text databases with keyword queries. With the rise of powerful Internet search engines, including Google, Yahoo!, and Bing, search and information retrieval has become familiar to most people. Nearly every computer application from email to word processing includes a search function. Because search is so familiar and available to the practitioner, we have not covered it in this book. Instead, Table 2.4 lists three resources that you might find helpful in the area of search and information retrieval.

Document Clustering

Document clustering uses algorithms from data mining to group similar documents into clusters. Data mining has been a very active field for nearly two decades, and clustering algorithms preceded that, so clustering algorithms are widely available in many commercial data and text mining software packages. We explore document clustering in Chapter 13 and in tutorials G, H, K, P, and X.

For more background information on clustering, see our handbook on data mining: see *Handbook of Statistical Analysis and Data Mining Applications* by R. Nisbet, J. Elder, and G. Miner.

Document Classification

Document classification assigns a known set of labels to untagged documents, using a model of text learned from documents with known labels. Like document clustering, document classification draws from an enormous field of work in data mining, statistics, and machine learning. It is one of the most

Table 2.4 Additional Resources on Search and Information Retrieval

Resource	Emphasis
Search Engines: Information Retrieval in Practice, by Bruce Croft, Donald Metzler, and Trevor Strohman	Emphasis on the practical aspects of building a search engine, including an example search engine. Also includes an overview of the theory and technology behind search engines.
Introduction to Information Retrieval, by Christopher D. Manning, Prabhakar Raghavan, and Hinrich Schütze	Comprehensive coverage of information retrieval, with more of an emphasis on the theory and mathematical origins of the field.
Solr 1.4: Enterprise Search Server, by David Smiley and Eric Pugh	Solr is a widely used open source search engine package from the Apache Software Foundation. This book thoroughly covers implementing Solr.

prominent techniques used in text mining and is a major emphasis of this book. Document classification and related techniques are discussed in Chapters 7, 8, 10, and 14 and in tutorials B, C, G, H, I, J, K, M, P, Q, and X.

For more background information on the theory and practice of classification, see *Handbook of Statistical Analysis and Data Mining Applications*, by R. Nisbet, J. Elder, and G. Miner.

Web Mining

Web mining is its own practice area due to the unique structure and enormous volume of data appearing on the web. Web documents are typically presented in a structured text format with hyperlinks between pages. These differences from standard text present a few challenges and many opportunities. As the Internet becomes even more ingrained in our popular culture with the rise of Facebook, Twitter, and other social media channels, web mining will continue to increase in value. Though it is still an emerging area, web mining draws on mature technology in document classification and natural language understanding. Web mining is covered in Chapters 12 and 15 and in tutorials A, P, Y, and AA.

For more details about web mining, see *Mining the Web: Analysis of Hypertext and Semi Structured Data*, by Soumen Chakrabarti.

Information Extraction

The goal of information extraction is to construct (or *extract*) structured data from unstructured text. Information extraction is one of the more mature fields within text mining, but it is difficult for beginners to work in without considerable effort, since it requires specialized algorithms and software. Furthermore, the training and tuning of an information extraction system require a large amount of effort. There are a number of commercial products available for information extraction, but all of them require some customization to achieve high performance for a given document database. Information extraction is covered in Chapter 9 and in tutorial N.

For more information, see the proceedings of the Message Understanding Conferences (MUC).[1] The MUC were sponsored by the Defense Advanced Research Projects Administration (DARPA) for the express purpose of evaluating different systems on an information extraction task. They provide the earliest summary of the field. More recently, the Conference on Natural Language Learning (CoNLL)[2] has included a shared task for evaluating information extraction approaches in many languages.

Natural Language Processing

Natural language processing (NLP) has a relatively long history in both linguistics and computer science. Recently, the focus of NLP has moved further into the text mining realm by considering statistical approaches. NLP is a powerful tool for providing useful input variables for text mining such as part of speech tags and phrase boundaries. A few areas of NLP are discussed in Chapters 3 and 5 and in tutorials H, K, and N.

[1] http://www-nlpir.nist.gov/related_projects/muc/proceedings/muc_7_toc.html
[2] http://ifarm.nl/signll/conll/

For more thorough coverage, we heartily recommend *Foundations of Statistical Natural Language Processing,* by Chris Manning and Hinrich Schütze. This superb book is for both novice and expert readers.

Concept Extraction

Extracting concepts is, in some ways, both the easiest and the hardest of the practice areas to do. The meaning of text is notoriously hard for automated systems to "understand." However, some initial automated work combined with human understanding can lead to significant improvements over the performance of either a machine or a human alone. These techniques are discussed in Chapters 11 and 13 (on clustering) and tutorials D, E, F, G, H, I, K, L, M, O, Q, S, U, W, and Z.

INTERACTIONS BETWEEN THE PRACTICE AREAS

The seven practice areas overlap considerably, since many practical text mining tasks sit at the intersection of multiple practice areas. A visualization of this overlap between practice areas is shown as a Venn diagram in Figure 2.3. For example, entity extraction draws from the practice areas of

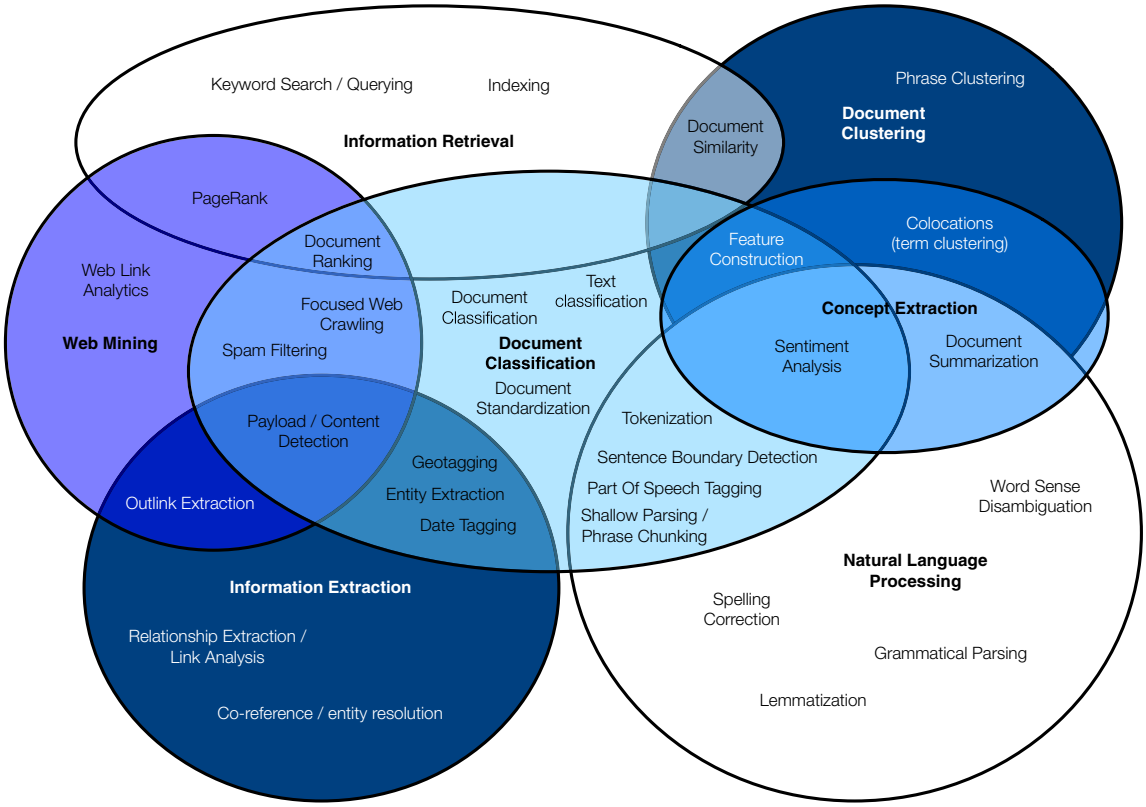

FIGURE 2.3

Visualizing the seven text mining practice areas (ovals) and how specific text mining tasks (labels within ovals) exist at their intersections.

information extraction and text classification, and document similarity measurement draws from the practice areas of document clustering and information retrieval.

SCOPE OF THIS BOOK

As we have just seen, text mining covers a diverse set of applications and algorithms. We have chosen to focus this book on techniques that are readily available for nonspecialists to apply immediately given the proper tools. Consequently, the areas of document classification, clustering, and concept extraction have the strongest representation in the book and the largest number of chapters and tutorials using these methods. This can be seen in Figure 2.2 with the list of the related chapters and tutorials. These three areas use techniques and algorithms that are drawn directly from data mining and are well represented in software for data mining and statistical analysis.

Web mining is a new and exciting application area for text mining practitioners. The Internet is rapidly changing with new information sources such as Facebook and Twitter. Because of this constant change, it has taken longer for a consensus to form over which methods perform best. We provide an introduction in the area of web mining with a limited number of chapters and tutorials. Interested readers are encouraged to explore the area on their own, and because of its high demand and rapid change, it may be possible to quickly become a leader in the field. Also, web mining borrows heavily from the areas of document classification, clustering, and concept extraction, allowing us to focus on those topics more.

Information extraction and natural language processing are becoming more accessible but still require significant amounts of domain expertise in linguistics to be successful. Of the seven areas, information extraction and natural language processing also are the most distinct technically, often requiring specialized software to achieve strong performance. Because of these challenges, we have chosen not to focus heavily on these two areas and instead provide an introduction and an avenue for exploration of these areas.

Finally, we provide minimal coverage of search and information retrieval. Since Google and other search engines have become such an integral part of our lives, search has become a key part of nearly every major software package. Consequently, search has become commoditized and is familiar enough to most users to skip the coverage of it here. If you are interested in building your own search engine, we have listed some excellent technical resources in Table 2.4.

An overall diagrammatic model that summarizes the scope of this book is presented in Figure 2.4. As shown in Figure 2.4, text mining draws upon many techniques in the broader field of text analytics. The central theme of this book is learning how to apply the diversity of powerful text mining models to solve practical problems in an organization. The tutorials in this book evolved out of the goal of driving you up the learning curve in text mining as efficiently as possible, using a learn-by-doing approach. That is the primary goal of this book.

SUMMARY

The term *text mining* can mean many different things to different authors, vendors, speakers, and clients. This chapter creates a rational taxonomy for the field, based on the perspective of a practitioner—a

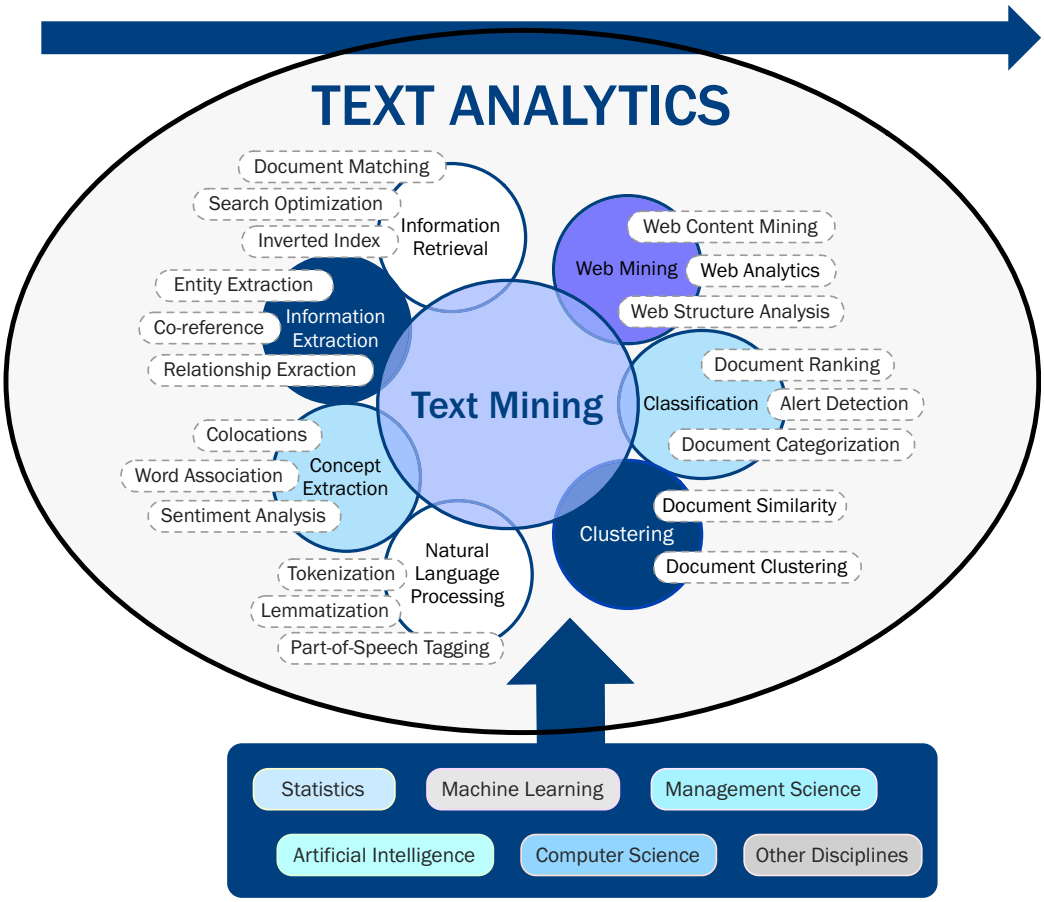

FIGURE 2.4

Text mining is the thematic center of this book, drawing upon contributions of many text analytical components and knowledge from many external disciplines (shown in blue at the bottom), which result in directional decisions affecting external results (shown by the blue arrow at the top).

person with some text data and an application goal. We define seven "practice areas" for text mining, based only on the practical distinctions in data and goal for an analyst trying to solve a given problem.

Chapter 1 described the history of text mining and how it is related to (borrows from and influences) six other fields. Figure 2.1 displays the overlap of those six fields with text mining and reveals the seven practice areas of text mining that are at the key intersections of the fields. An inductive model, in the form of a decision tree (Figure 2.2), asks the five key questions a practitioner needs to answer to be guided to the appropriate practice area for his or her text-based problem. The tree reveals not only the practice area most appropriate for a given text challenge but also the chapters and tutorials of this book that address that type of application. This allows the reader to jump right to the areas in the book that are most useful for his or her work. It further reveals where the book's coverage is strong and where it is light. The areas most covered in this book are those that have arrived just past the cutting edge of

research into development—that is, those that are within reach of a technical nonspecialist who is willing to learn and yet are not ubiquitous (like search is).

Finally, when the practice areas are themselves generalized to oval regions in a Venn diagram (Figure 2.3), individual text mining tasks, such as lemmatization, can be located at the intersection of the seven practice areas, further helping to focus a user on the appropriate resources to use for a task. Where the book's coverage is incomplete, recommended high-quality external resources are listed.

POSTSCRIPT

Text mining is proving to be extremely useful, and this taxonomy of the wide-ranging field is designed to help analysts hone in on the practice area and resources for that area that are most helpful to achieving high productivity on the particular text application challenge they are facing.

A common claim among data miners is that 80 to 90 percent of the project time is consumed by data preparation steps. The same is true for text mining. In contrast to data mining, where some of the data are in text format, *all* of the data for text mining are in text format. The initial challenge is to transform these text data into a numerical format for subsequent analysis. In the next chapter, you will be introduced to the steps necessary to preprocess text data to create data structures that can be analyzed numerically.

References

Chakrabarti, Soumen. Mining the Web: Analysis of Hypertext and Semi-Structured Data, Morgan Kaufmann, San Francisco, 2002.

Croft, Bruce, Donald Metzler, and Trevor Strohman. Search Engines: Information Retrieval in Practice, Addison-Wesley, Boston, MA, 2009.

Manning, Christopher D., Prabhakar Raghavan, and Hinrich Schütze. Introduction to Information Retrieval, Cambridge University Press, New York, 2008.

Manning, Chris, and Hinrich Schütze. Foundations of Statistical Natural Language Processing, Cambridge, MA: MIT Press, Cambridge, MA, 1999.

Nisbet, R., J. Elder, and G. Miner. (2009). Handbook of Statistical Analysis and Data Mining Applications, Elsevier, Burlington, MA.

Smiley, David, and Eric Pugh. Solr 1.4: Enterprise Search Server. Packt Publishing, Birmingham, England, UK, 2009.