1

What is Text Mining?

In this chapter, you will learn

- the basic definition of practical text mining
- why text mining is important to the modern enterprise
- examples of text mining used in enterprise
- the challenges facing text mining
- an example workflow for processing natural language in analytical contexts
- a simple text mining example
- when text mining is appropriate

Learning how to perform text mining should be an interesting and exciting journey throughout this book. A fun artifact of learning text mining is that you can use the methods in this book on your own social media or online exchanges. Beyond these everyday online applications to your personal interactions, this book provides business use cases in an effort to show how text mining can improve products, customer service, marketing or human resources.

1.1 What is it?

There are many technical definitions of text mining both on the Internet and in textbooks, but as the primary goal of text mining in this book is the extraction of an output that is useful such as a visualization or structured table of outputs to be used elsewhere; this is my definition:

> Text mining is the process of distilling actionable insights from text.

Text mining within the context of this book is a commitment to real world cases which impact business. Therefore, the definition and this book are aimed

Text Mining in Practice with R, First Edition. Ted Kwartler.

at meaningful distillation of text with the end goal to aid a decision-maker. While there may be some differences, the terms text mining and text analytics can be used interchangeably. Word choice is important; I use text mining because it more adequately describes the uncovering of insights and the use of specific algorithms beyond basic statistical analysis.

1.1.1 What is Text Mining in Practice?

In this book, text mining is more than an academic exercise. I hope to show that text mining has enterprise value and can contribute to various business units. Specifically, text mining can be used to identify actionable social media posts for a customer service organization. It can be used in human resources for various purposes such as understanding candidate perceptions of the organization or to match job descriptions with resumes. Text mining has marketing implications to measure campaign salience. It can even be used to identify brand evangelists and impact customer propensity modeling. Presently the state of text mining is somewhere between novelty and providing real actionable business intelligence. The book gives you not only the tools to perform text mining but also the case studies to help identify practical business applications to get your creative text mining efforts started.

1.1.2 Where Does Text Mining Fit?

Text mining fits within many disciplines. These include private and academic uses. For academics, text mining may aid in the analytical understanding of qualitatively collected transcripts or the study of language and sociology. For the private enterprise, text mining skills are often contained in a data science team. This is because text mining may yield interesting and important inputs for predictive modeling, and also because the text mining skillset has been highly technical. However, text mining can be applied beyond a data science modeling workflow. Business intelligence could benefit from the skill set by quickly reviewing internal documents such as customer satisfaction surveys. Competitive intelligence and marketers can review external text to provide insightful recommendations to the organization. As businesses are saving more textual data, they will need to break text-mining skills outside of a data science team. In the end, text mining could be used in any data driven decision where text naturally fits as an input.

1.2 Why We Care About Text Mining

We should care about textual information for a variety of reasons.

- Social media continues to evolve and affect an organization's public efforts.

- Online content from an organization, its competitors and outside sources, such as blogs, continues to grow.
- The digitization of formerly paper records is occurring in many legacy industries, such as healthcare.
- New technologies like automatic audio transcription are helping to capture customer touchpoints.
- As textual sources grow in quantity, complexity and number of sources, the concurrent advance in processing power and storage has translated to vast amounts of text being stored throughout an enterprise's data lake.

Yet today's successful technology companies largely rely on numeric and categorical inputs for information gains, machine learning algorithms or operational optimization. It is illogical for an organization to study only structured information yet still devote precious resources to recording unstructured natural language. Text represents an untapped input that can further increase competitive advantage. Lastly, enterprises are transitioning from an industrial age to an information age; one could argue that the most successful companies are transitioning again to a customer-centric age. These companies realize that taking a long term view of customer wellbeing ensures long term success and helps the company to remain salient. Large companies can no longer merely create a product and forcibly market it to end-users. In an age of increasing customer expectations customers want to be heard by corporations. As a result, to be truly customer centric in a hyper competitive environment, an organization should be listening to their constituents whenever possible. Yet the amount of textual information from these interactions can be immense, so text mining offers a way to extract insights quickly.

Text mining will make an analyst's or data scientist's efforts to understand vast amounts of text easier and help ensure credibility from internal decision-makers. The alternative to text mining may mean ignoring text sources or merely sampling and manually reviewing text.

1.2.1 What Are the Consequences of Ignoring Text?

There are numerous consequences of ignoring text.

- Ignoring text is not an adequate response of an analytical endeavor. Rigorous scientific and analytical exploration requires investigating sources of information that can explain phenomena.
- Not performing text mining may lead an analysis to a false outcome.
- Some problems are almost entirely text-based, so not using these methods would mean significant reduction in effectiveness or even not being able to perform the analysis.

Explicitly ignoring text may be a conscious analyst decision, but doing so ignores text's insightful possibilities. This is analogous to an ostrich that sticks

its head in the ground when confronted. If the aim is robust investigative quantitative analysis, then ignoring text is inappropriate. Of course, there are constraints to data science or business analysis, such as strict budgets or timelines. Therefore, it is not always appropriate to use text for analytics, but if the problem being investigated has a text component, and resource constraints do not forbid it, then ignoring text is not suitable.

Wisdom of Crowds

As an alternative, some organizations will sample text and manually review it. This may mean having a single assessor or panel of readers or even outsourcing analytical efforts to human-based services like `mturk` or `crowdflower`. Often communication theory does not support these methods as a sound way to score text, or to extract meaning. Setting aside sampling biases and logistical tabulation difficulties, communication theory states that the meaning of a message relies on the recipient. Therefore a single evaluator introduces biases in meaning or numerical scoring, e.g. sentiment as a numbered scale. Additionally, the idea behind a group of people scoring text relies on Sir Francis Galton's theory of "Vox Populi" or wisdom of crowds.

To exploit the wisdom of crowds four elements must be considered:

- Assessors need to exercise independent judgments.
- Assessors need to possess a diverse information understanding.
- Assessors need to rely on local knowledge.
- There has to be a way to tabulate the assessors' results.

Sir Francis Galton's experiment exploring the wisdom of crowds met these conditions with 800 participants. At an English country fair, people were asked to guess the weight of a single ox. Participants guessed separately from each other without sharing the guess. Participants were free to look at the cow themselves yet not receive expert consultation. In this case, contestants had a diverse background. For example, there were no prerequisites stating that they needed to be a certain age, demographic or profession. Lastly, guesses were recorded on paper for tabulation by Sir Francis to study. In the end, the experiment showed the merit of the wisdom of crowds. There was not an individual correct guess. However, the median average of the group was exactly right. It was even better than the individual farming experts who guessed the weight.

If these conditions are not met explicitly, then the results of the panel are suspect. This may seem easy to do, but in practice it is hard to ensure within an organization. For example a former colleague at a major technology company in California shared a story about the company's effort to create Internet-connected eyeglasses. The eyeglasses were shared with internal employees, and feedback was then solicited. The text feedback was sampled and scored by internal employees. At first blush this seems like a fair assessment of the product's features and

expected popularity. However, the conditions for the wisdom of crowds were not met. Most notably, the need for a decentralized understanding of the question was not met. As members of the same technology company, the respondents are already part of a self-selected group that understood the importance of the overall project within the company. Additionally, the panel had a similar assessment bias because they were from the same division that was working on the project. This assessing group did not satisfy the need for independent opinions when assessing the resulting surveys. Further, if a panel is creating summary text as the output of the reviews, then the effort is merely an information reduction effort similar to numerically taking an average. Thus it may not solve the problem of too much text in a reliable manner. Text mining solves all these problems. It will use all of the presented text and does so in a logical, repeatable and auditable way. There may be analyst or data scientist biases but they are documented in the effort and are therefore reviewable. In contrast, crowd-based reviewer assessments are usually not reviewable.

Despite the pitfalls of ignoring text or using a non-scientific sampling method, text mining offers benefits. Text mining technologies are evolving to meet the demands of the organization and provide benefits leading to data-driven decisions. Throughout this book, I will focus on benefits and applied applications of text mining in business.

1.2.2 What Are the Benefits of Text Mining?

There are many benefits of text mining including:

- Trust is engendered among stakeholders because little to no sampling is needed to extract information.
- The methodologies can be applied quickly.
- Using R allows for auditable and repeatable methods.
- Text mining identifies novel insights or reinforces existing perceptions based on all relevant information.

Interestingly, text mining first appears in the Gartner Hype Cycle in 2012. At that moment, it was listed in the "trough of disillusionment." In subsequent years, it has not been listed on the cycle at all, leading me to believe that text analysis is either at a steady enterprise use state or has been abandoned by enterprises as not useful. Despite not being listed, text mining is used across industries and in various manners. It may not have exceeded the over-hyped potential of 2012's Gartner Hype Cycle, but text is showing merit. Hospitals use text mining of doctors' notes to understand readmission characteristics of patients. Financial and insurance companies use text to identify compliance risks. Retailers use customer service notes to make operational changes when

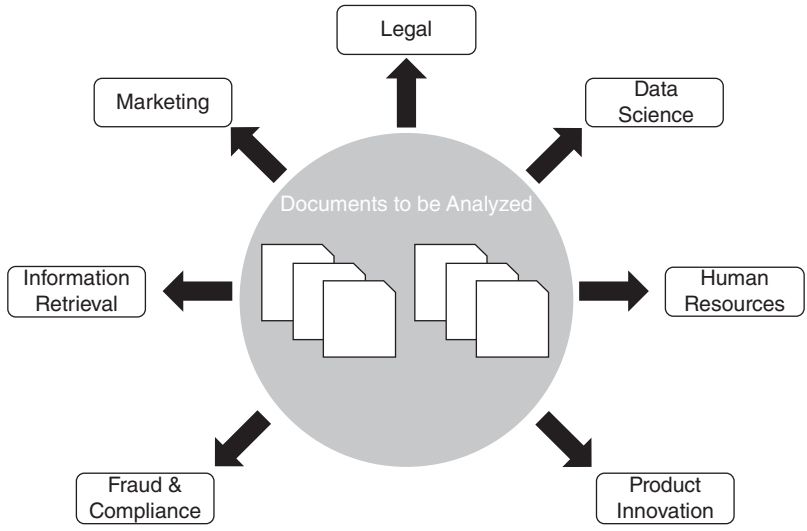

Figure 1.1 Possible enterprise uses of text min.

failing customer expectations. Technology product companies use text mining to seek out feature requests in online reviews. Marketing is a natural fit for text analysis. For example, marketing companies monitor social media to identify brand evangelists. Human resource analytics efforts focus on resume text to match to job description text. As described here, mastering text mining is a skill set sought out across verticals and is therefore a worthwhile professional endeavor. Figure 1.1 shows possible business units that can benefit from text mining in some form.

1.2.3 Setting Expectations: When Text Mining Should (and Should Not) Be Used

Since text is often a large part of a company's database, it is believed that text mining will lead to ground-breaking discoveries or significant optimization. As a result, senior leaders in an organization will devote resources to text mining, expecting to yield extensive results. Often specialists are hired, and resources are explicitly devoted to text mining. Outside of text mining software, in this case R, it is best to use text mining only in cases where it naturally fits the business objective and problem definition. For example, at a previous employer, I wondered how prospective employees viewed our organization compared to peer organizations. Since these candidates were outside the organization, capturing numerical or personal information such as age or company-related perspective scoring was difficult. However, there are forums and interview reviews anonymously shared online. These are shared as text so naturally text mining

was an appropriate tool. When using text mining, you should prioritize defining the problem and reviewing applicable data, not using an exotic text mining method. Text mining is not an end in itself and should be regarded as another tool in an analyst's or data scientist's toolkit.

Text mining cannot distill large amounts of text to gain an absolute view of the truth. Text mining is part art and part science. An analyst can mislead stakeholders by removing certain words or using only specific methods. Thus, it is important to be up front about the limitations of text mining. It does not reveal an absolute truth contained within the text. Just as an average reduces information for consumption of a large set of numbers, text mining will reduce information. Sometimes it confirms previously held beliefs and sometimes it provides novel insights. Similar to numeric dimension reduction techniques, text mining abridges outliers, low frequency phrases and important information. It is important to understand that language is more colorful and diverse in understanding than numerical or strict categorical data. This poses a significant problem for text miners. Stakeholders need to be wary of any text miner who knows a truth solely based on the algorithms in this book. Rather, the methods in this book can help with the narrative of the data and the problem at hand, or the outputs can even be used in supervised learning alongside numeric data to improve the predictive outcomes. If doing predictive modeling using text, a best practice when modeling alongside non-text data features is to model with and without the text in the attribute set. Text is so diverse that it may even add noise to predictive efforts. Table 1.1 refers to actual use cases where text mining may be appropriate.

Table 1.1 Example use cases and recommendations to use or not use text mining.

Example use case	Recommendation
Survey texts	Explore topics using various methods to gain a respondent's perspective.
Reviewing a small number of documents	Don't perform text mining on an extremely small corpus, as the results and conclusion can be skewed.
Human resource documents	Tread carefully; text mining may yield insights, but the data and legal barriers may make the analysis inappropriate.
Social media	Use text mining to collect (when allowed) from online sources and then apply preprocessing steps to extract information.
Data science predictive modeling	Text mining can yield structured inputs that could be useful in machine learning efforts.
Product/service reviews	Use text mining if the number of reviews is large.
Legal proceeding	Use text mining to identify individuals and specific information.

Another suggestion for effective text mining is to avoid over using a word cloud. Analysts armed with the knowledge of this book should not create a word cloud without a need for it. This is because word clouds are often used without need, and as a result they can actually diminish their impact. However, word clouds are popular and can be powerful in showing term frequency, among other things, such as the one in Figure 1.2, which runs over the text of this chapter. Throwing caution to the wind, it demonstrates a word cloud of terms in Chapter 1. It is not very insightful because, as expected, the terms text and mining are the most frequent and largest words in the cloud!

In fact, word clouds are so popular that an entire chapter is devoted to various types of word clouds that can be insightful. However, many people consider word clouds a cliché, so their impact is fading. Also, word clouds represent

Figure 1.2 A gratuitous word cloud for Chapter 1.

a relatively easy way to mislead consumers of an analysis. In the end, they should be used in conjunction with other methods to confirm the correctness of a conclusion.

1.3 A Basic Workflow – How the Process Works

Text represents unstructured data that must be preprocessed into a structured manner. Features need to be defined and then extracted from the larger body of organized text known as a corpus. These extracted features are then analyzed. The chevron arrows in Figure 1.3 represent structured predefined steps

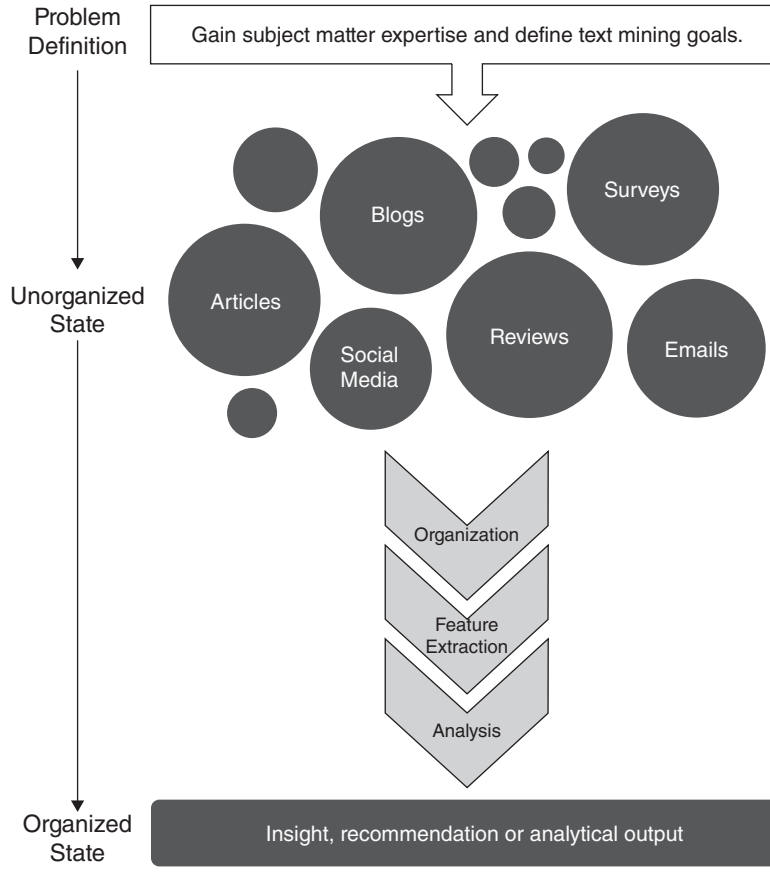

Figure 1.3 Text mining is the transition from an unstructured state to a structured understandable state.

that are applied to the unorganized text to reach the final output or conclusion. Overall Figure 1.3 is a high level workflow of a text mining project.

The steps for text mining include:

1) **Define the problem and specific goals.** As with other analytical endeavors, it is not prudent to start searching for answers. This will disappoint decision-makers and could lead to incorrect outputs. As the practitioner, you need to acquire subject matter expertise sufficient to define the problem and the outcome in an appropriate manner.
2) **Identify the text that needs to be collected.** Text can be from within the organization or outside. Word choice varies between mediums like Twitter and print so care must be taken to explicitly select text that is appropriate to the problem definition. Chapter 9 covers places to get text beyond reading in files. The sources covered include basic web scraping, APIs and R's specific API libraries, like "`twitteR`." Sources are covered later in the book so you can focus on the tools to text mine, without the additional burden of finding text to work on.
3) **Organize the text.** Once the appropriate text is identified, it is collected and organized into a corpus or collection of documents. Chapter 2 covers two types of text mining conceptually, and then demonstrates some preparation steps used in a "bag of words" text mining method.
4) **Extract features.** Creating features means preprocessing text for the specific analytical methodology being applied in the next step. Examples include making all text lowercase, or removing punctuation. The analytical technique in the next step and the problem definition dictate how the features are organized and used. Chapters 3 and 4 work on basic extraction to be used in visualizations or in a sentiment polarity score. These chapters are not performing heavy machine learning or technical analysis, but instead rely on simple information extraction such as word frequency.
5) **Analyze.** Apply the analytical technique to the prepared text. The goal of applying an analytical methodology is to gain an insight or a recommendation or to confirm existing knowledge about the problem. The analysis can be relatively simple, such as searching for a keyword, or it may be an extremely complex algorithm. Subsequent chapters require more in-depth analysis based on the prepared texts. A chapter is devoted to unsupervised machine learning to analyze possible topics. Another illustrates how to perform a supervised classification while another performs predictive modeling. Lastly you will switch from a "bag of words" method to syntactic parsing to find named entities such as people's names.
6) **Reach an insight or recommendation.** The end result of the analysis is to apply the output to the problem definition or expected goal. Sometimes this can be quite novel and unexpected, or it can confirm the previously held idea. If the output does not align to the defined problem or completely

satisfy the intended goal, then the process becomes repetitious and can be changed at various steps. By focusing on real case studies that I have encountered, I hope to instill a sense of practical purpose to text mining. To that end, the case studies, the use of non-academic texts and the exercises of this book are meant to lead you to an insight or narrative about the issue being investigated. As you use the tools of this book on your own, my hope is that you will remember to lead your audience to a conclusion.

The distinct steps are often specific to the particular problem definition or analytical technique being applied. For example, if one is analyzing tweets, then removing retweets may be useful but it may not be needed in other text mining exploration. Using R for text mining means the processing steps are repeatable and auditable. An analyst can customize the preprocessing steps outlined throughout the book to improve the final output. The end result is an insight, a recommendation or may be used in another analysis. The R scripts in this book follow this transition from an unorganized state to an organized state, so it is important to recall this mental map.

The rest of the book follows this workflow and adds more context and examples along the way. For example, Chapter 2 examines the two main approaches to text mining and how to organize a collection of documents into a clean corpus. From there you start to extract features of the text that are relevant to the defined problem. Subsequent chapters add visualizations, such as word clouds, so that a data scientist can tell the analytical narrative in a compelling way to stakeholders. As you progress through the book the types and methods of extracted features or information grow in complexity because the defined problems get more complex. You quickly divert to covering sentiment polarity so you can understand Airbnb reviews. Using this information you will build compelling visualizations and know what qualities are part of a good Airbnb review. Then in Chapter 5 you learn topic modeling using machine learning. Topic modeling provides a means to understand the smaller topics associated within a collection of documents without reading the documents themselves. It can be useful for tagging documents relating to a subject. The next subject, document classification, is used often. You may be familiar with document classification because it is used in email inboxes to identify spam versus legitimate emails. In this book's example you are searching for "clickbait" from online headlines. Later you examine text as it relates to patient records to model how a hospital identifies diabetic readmission. Using this method, some hospitals use text to improve patient outcomes. In the same chapter you even examine movie reviews to predict box office success. In a subsequent chapter you switch from the basic bag of words methodology to syntactic parsing using the OpenNLP library. You will identify named entities, such as people, organizations and locations within Hillary Clinton's emails. This can be useful in legal proceedings in which the volume of documentation is large and the deadlines

are tight. Marketers also use named entity recognition to understand what influencers are discussing. The remaining chapters refocus your attention back to some more basic principles at the top of the workflow, namely where to get text and how to read it into R. This will let you use the scripts in this book with text that is thought provoking to your own interests.

1.4 What Tools Do I Need to Get Started with This?

To get started in text mining you need a few tools. You should have access to a laptop or workstation with at least 4GB of RAM. All of the examples in this book have been tested on a Microsoft's Windows operating systems. RAM is important because R's processing is done "in memory." This means that the objects being analyzed must be contained in the RAM memory. Also, having a high speed internet connection will aid in downloading the scripts, R library packages and example text data and for gathering text from various webpages. Lastly, the computer needs to have an installation of R and R Studio. The operating system of the computer should not matter because R has an installation for Microsoft, Linux and Mac.

1.5 A Simple Example

Online customer reviews can be beneficial to understanding customer perspectives about a product or service. Further, reviewers can sometimes leave feedback anonymously, allowing authors to be candid and direct. While this may lead to accurate portrayals of a product it may lead to "keyboard courage" or extremely biased opinions. I consider it a form of selection bias, meaning that the people that leave feedback may have strong convictions not indicative of the overall product or service's public perception. Text mining allows an enterprise to benchmark their product reviews and develop a more accurate understanding of some public perceptions. Approaches like topic modeling and polarity (positive and negative scoring) which are covered later in this book may be applied in this context. Scoring methods can be normalized across different mediums such as forums or print, and when done against a competing product, the results can be compelling.

Suppose you are a Nike employee and you want to know about how consumers are viewing the Nike Men's Roshe Run Shoes. The text mining steps to follow are:

1) **Define the problem and specific goals.** Using online reviews, identify overall positive or negative views. For negative reviews, identify a consistent cause of the poor review to be shared with the product manager and manufacturing personnel.

2) **Identify the text that needs to be collected.** There are running websites providing expert reviews, but since the shoes are mass market, a larger collection of general use reviews would be preferable. New additions come out annually, so old reviews may not be relevant to the current release. Thus, a shopping website like Amazon could provide hundreds of reviews, and since there is a timestamp on each review, the text can be limited to a particular timeframe.
3) **Organize the text.** Even though Amazon reviewers rate products with a number of stars, reviews with three or fewer stars may yield opportunities to improve. Web scraping all reviews into a simple csv with a review per row and the corresponding timestamp and number of stars in the next columns will allow the analysis to subset the corpus by these added dimensions.
4) **Extract features.** Reviews will need to be cleaned so that text features can be analyzed. For this simple example, this may mean removing common words with little benefit like "shoe" or "nike," running a spellcheck and making all text lowercase.
5) **Analyze.** A very simple way to analyze clean text, discussed in an early chapter, is to scan for a specific group of keywords. The text-mining analyst may want to scan for words given their subject matter expertise. Since the analysis is about shoe problems one could scan for "fit," "rip" or "tear," "narrow," "wide," "sole," or any other possible quality problem from reviews. Then summing each could provide an indication of the most problematic feature. Keep in mind that this is an extremely simple example and the chapters build in complexity and analytical rigor beyond this illustration.
6) **Reach an insight or recommendation.** Armed with this frequency analysis, a text miner could present findings to the product manager and manufacturing personnel that the top consumer issue could be "narrow" and "fit." In practical application, it is best to offer more methodologies beyond keyword frequency, as support for a finding.

1.6 A Real World Use Case

It is regularly the case that marketers learn best practices from each other. Unlike in other professions many marketing efforts are available outside of the enterprise, and competitors can see the efforts easily. As a result, competitive intelligence in this space is rampant. It is also another reason why novel ideas are often copied and reused, and then the novel idea quickly loses salience with its intended audience. Text mining offers a quick way to understand the basics of a competitor's text-based public efforts.

When I worked at amazon.com, creating the social customer service team, we were obsessed with how others were doing it. We regularly read and reviewed

other companies' replies and learned from their missteps. This was early 2012, so customer service in social media was considered an emerging practice, let alone being at one of the largest retailers in the world. At the time, the belief was that it was fraught with risk. Amazon's legal counsel, channel marketers in charge of branding and even customer service leadership were weary of publically acknowledging any shortcomings or service issues. The legal department was involved to understand if we were going to set undeliverable expectations or cause any tax implications on a state-by-state basis. Further, each brand owner, such as Amazon Prime, Amazon Mom, Amazon MP3, Amazon Video on Demand, and Amazon Kindle had cultivated their own style of communicating through their social media properties. Lastly, customer service leadership had made multiple promises that reached all the way to Jeff Bezos, the CEO, about flawless execution and servicing in this channel demonstrating customer centricity. The mandate was clear: proceed, but do so cautiously and do not expand faster than could be reasonably handled to maintain quality set by all these internal parties. The initial channels we covered were the two "Help" forums on the site, then retail and Kindle Facebook pages, and lastly, Twitter. We had our own missteps. I remember the email from Jeff that came down through the ranks with a simple "?" concerning an inappropriate briefly posted video to the Facebook wall. That told me our efforts were constantly under review and that we had to be as good as or better than other companies.

Text mining proved to be an important part of the research that was done to understand how others were doing social media customer service. We had to grasp simple items like length of a reply by channel, basic language used, typical agent workload, and if adding similar links repeatedly made sense. My initial thought was that it was redundant to repeatedly post the same link, for example to our "contact us", form. Further, we didn't know what types of help links were best to post. Should they be informative pages or forms or links to outside resources? We did not even know how many people should be on the team and what an average workload for a customer service representative was.

In short, the questions basic text mining can help with are

1) What is the average length of a social customer service reply?
2) What links were referenced most often?
3) How many people should be on the team? How many social replies is reasonable for a customer service representative to handle?

Channel by channel we would find text of some companies already providing public support. We would identify and analyze attributes that would help us answer these questions. In the next chapter, covering basic text mining, we will actually answer these questions on real customer service tweets and go through the six-step process to do so.

Looking back, the answers to these questions seem common sense, but that is after running that team for a year. Now social media customer service has

expanded to be the norm. In 2012, we were creating something new at a Fortune 50 fast growing company with many opinions on the matter, including "do not bother!" At the time, I considered Wal-Mart, Dell and Delta Airlines to be best in class social customer service. Basic text mining allowed me to review their respective replies in an automated fashion. We spoke with peers at Expedia but it proved more helpful to perform basic text mining and read a small sample of replies to help answer our questions.

1.7 Summary

In this chapter you learned

- the basic definition of practical text mining
- why text mining is important to the modern enterprise
- examples of text mining used in enterprise
- the challenges facing text mining
- an example workflow for processing natural language in analytical contexts
- a simple text mining example
- when text mining is appropriate