Preface

Let us begin by telling you what this book is not about. It is not about the spoken word, in any way, shape, or form. When we refer to *Mining the Talk*, "the talk" refers to words on the page, or to be more precise, words on the electronic page, not words out of the mouth. The reason we call it "talk" is not to be cute, but to emphasize the informal nature of the data being mined. Most data that is mined, or searched, or graphed is meant to be used in this way. That is usually why the data was put there in the first place. The collection and the analysis of the data go hand in hand. Not so with the type of data we refer to as "talk"—talk is put on the earth simply to be read. It is casual, unstructured, unpredictable, and diverse. You never know what to expect from talk, and that is what makes it so endlessly fascinating.

This book is not about text-mining research—at least not in a general sense, and certainly not in the sense that text mining is defined in the literature. We will not survey all the approaches or discuss the pros and cons of various algorithms. We will not describe any other methods beyond those we are intimately familiar with—those we use on a daily basis to peer into obscure data sets and make sense out of them.

This book is not about text search. If text search is akin to finding the needle in a haystack, this book is like plunging into the haystack (and not with the purpose of getting stabbed with the needle, though you might regard it as one possible benefit of such an action). If you enter a query into a search engine and receive a message that reads "results 1–10 of 1,837,220,135," you might pause to wonder about the results you don't see. This book has little to say about results 1–10. It's much more about the other 1,837,220,125—an area in which, to our minds, far too little effort is generally spent, when compared to the relative potential rewards of discovery.

This book is not about natural language processing (NLP). The techniques employed are statistical in nature, meaning that basic counting of regularly recurring text features is the basis for reasoning. No higher-level syntax or grammar is recognized or utilized in the processing of the unstructured information.

This book is not about a black box approach that magically transforms streams of characters into actionable ideas. It is about using whatever time and attention you are willing to invest in your unstructured mining endeavor in an effective manner to achieve a positive benefit in a reasonable time.

This book is, quite simply, a description of a method and its application—a method we have devised for getting useful information out of large amounts of unstructured data. It is a method we created in order to deal with the fact that standard approaches that were readily available were not getting the job done. When we achieved success in one area, we started applying this same approach to more and more kinds of data, with equal success. As the years passed and we were able to convince ourselves and others that this method was valuable, it started to become clear that the method needed to be published—to share the capabilities with a much wider audience and invite closer comments and scrutiny of what we have created. We are convinced that the potential of this method is just beginning to be tapped. But only with a much wider exposure and application will we ever find out the limits of just what it can do.

The purpose of this book is to both explain our approach to *Mining the Talk* and also to expand awareness of how and where our techniques can be applied to business data. We believe that the potential of this technology is only beginning to be tapped, and that a greater awareness in the business and technical community of what knowledge can be gleaned from unstructured information will likely lead to an explosion of application areas.

We assume that you, the reader, are investing time in this book because you are faced with the same kind of problems that we were: lots of data, not much structure, and a certainty that if you only had a reasonable approach (something better than searching or reading it line by line), you could utilize what's in the data to improve whatever endeavor your organization is currently engaged in. Although we cannot give you any magic formula that will effortlessly transform your unstructured data into useful information, we can promise you a systematic way to efficiently turn effort spent on your data into understanding—as much understanding as the data and your effort will allow. Is it a perfect solution? No way. Is it better than reading each text example that you want to understand? Almost certainly.

At the end of reading this book, we hope you feel, as we do, a sense of wonder and excitement about the vast ocean of unexplored unstructured content that lies waiting to be charted. We also wish you to come away with a healthy respect for the dangers and a realistic appraisal of the costs of such a voyage. This book contains a very realistic, practical approach to mining unstructured information. There's no trick to getting the knowledge you need out of the available data; it just takes intense focus, hard work, thoughtful planning, and finally, an open mind.

Who Should Read This Book?

We have made a conscious effort in writing this book to reach out beyond the data mining community. We have strived to write this book for a general audience, including business executives, engineers, and students having an interest in studying this field as a potential career. The audience for this book is primarily business professionals who have data management or analysis responsibility or needs. This would include business consultants, managers and executives, IT professionals, knowledge workers, market analysts, and those involved in the management of intellectual property. The reader benefits from the book by seeing how data and text-mining techniques and processes can be employed to solve real-world business problems. The book does not assume a high level of familiarity with data mining or analytic concepts. It is primarily a qualitative description of our technique, though some quantitative supporting details are supplied. There is no prerequisite background required of the reader, though a mathematical or analytic background will certainly help. This book is also very relevant to students in data mining or machine learning, because it demonstrates some proven, practical approaches to real-world text-mining problems that will complement the techniques typically taught in academia.

How This Book Is Organized

Chapter 1 of this book is an introduction to our methodology, describing the history of how and why we developed the *Mining the Talk* method, and including a high-level description of what the methodology actually consists of.

Each of the next five chapters describes a different application area of *Mining the Talk*. These are the following:

- Chapter 2, "Mining Customer Interactions": Interactions between your business and its customers
- Chapter 3, "Mining the Voice of the Customer": Customers (and others) discussing your business and its products online

- Chapter 4, "Mining the Voice of the Employee": Internal organizational communication
- Chapter 5, "Mining to Improve Innovation": Public information on technical innovation for collaboration and partnership
- Chapter 6, "Mining to See the Future": Technology and market trends

Chapter 7, "Future Applications," discusses some potential future applications of *Mining the Talk* techniques and then concludes the book.

The Appendix, "The IBM Unstructured Information Modeler Users Manual," contains a detailed description of software that implements the *Mining the Talk* methodology.

A Clarification on Personal Pronouns

When two people author a book, they seldom actually sit down together and write it. Typically, one writes, the other edits/corrects/appends, and then they switch roles. The actual writing itself is generally a solitary activity. Therefore, occasionally, it seems more natural to use the singular personal pronoun when one is required, even though technically "I" probably should say "we." We hope the reader will not find this inconsistency confusing. Please don't take this to mean that one of the two of us is taking ownership of or claiming credit for whatever section *we* are writing about. In general, all of this work has been a team effort between the two of us, as well as many others.

Software Applications

It is my fervent hope that you will emerge from reading this book with an immediate desire to try out these techniques on your own data. If you wish to do this on your own, you have our best wishes. Of course, if you would like some help, we at IBM have some wonderful tools that will get you started as quickly as possible; in fact, a demo version of one of these tools is available for download.

Everything we talk about in this book has been implemented as software, and the reader will notice many screenshots taken from these tools. It turns out that none of these software applications is actually currently for sale as a shrink-wrapped product. Instead, they are all "research assets" used internally at IBM and by IBM consultants on customer service engagements. However, it seemed unfair to tell you about our method, explain it in great detail, and then not give you the tools you need to try it out. Therefore, the folks at IBM Alphaworks have graciously agreed to provide a free demonstration copy of one of our early text-mining tools, IBM Unstructured Information Modeler, as an adjunct to this book. Here is the URL for those who are interested:

http://www.alphaworks.ibm.com/tech/uimodeler.

The user manual for this tool is contained in the Appendix. We hope you find this useful, and that it whets your appetite to try out our more advanced applications.

On the other hand, if you want to take our methods and create your own software solution to sell as a product, while we applaud your initiative and enthusiasm, you really should first discuss this with suitable representatives from IBM business development. IBM has sole ownership of all the intellectual property described in this book, all of which is protected by U.S. patents, both granted and pending. All rights reserved, etc., etc.