Preface

Data mining is a mature technology. The prediction problem, looking for predictive patterns in data, has been widely studied. Strong methods are available to the practitioner. These methods process structured numerical information, where uniform measurements are taken over a sample of data. Text is often described as unstructured information. So, it would seem, text and numerical data are different, requiring different methods. Or are they? In our view, a prediction problem can be solved by the same methods, whether the data are structured numerical measurements or unstructured text. Text and documents can be transformed into measured values, such as the presence or absence of words, and the same methods that have proven successful for predictive data mining can be applied to text. Yet, there are key differences. Evaluation techniques must be adapted to the chronological order of publication and to alternative measures of error. Because the data are documents, more specialized analytical methods may be preferred for text. Moreover, the methods must be modified to accommodate very high dimensions: tens of thousands of words and documents. Still, the central themes are similar.

Our view of text mining allows us to unify the concepts of different fields. No longer is "natural language processing" the sole domain of linguists and their allied computer specialists. No longer is search engine technology distinct from other forms of machine learning. Ours is an open view. We welcome you to try your hand at learning from data, whether numerical or text. You need not have a Ph.D. in linguistics to work in this area.

Not everyone will agree with our perspective. The natural language specialist may argue that ours is a shallow view of text that will solve some problems, but the bigger problems, such as answering questions

posed by a user, can only be solved with a deeper understanding of language.

There is room for both viewpoints to coexist. Large text collections contain valuable information that can be mined with today's tools instead of waiting for tomorrow's techniques. While others search for the essence of language understanding, we can immediately look for recurring word patterns in large collections of digital documents.

Some parts of the book may seem simple to the advanced student or professional. Other parts may appear mathematical. They all fit our common theme of a strictly empirical view of text mining and an application of well-known analytical methods. We provide examples and software. Our presentation has a pragmatic bent with numerous references in the research literature for you to follow when so inclined. We want to be practical, yet inclusive of the wide community that might be interested in applications of text mining. We concentrate on predictive learning methods but also look at information retrieval and search engines, as well as clustering methods. We illustrate by examples, case studies, and the accompanying downloadable software.

While some analytical methods may be highly developed, predictive text mining is an emerging area of application. We have tried to summarize our experiences and provide the tools and techniques for your own experiments.

Audience

Our book is aimed at IT professionals and managers as well as advanced undergraduate computer science students and beginning graduate students. Some background in data mining is beneficial but is not essential. If you are looking to do research in the area, the material in this book will provide direction in expanding your horizons. If you want to be a practitioner of text mining, you can read about our recommended methods and our descriptions of case studies.

Supplementary Web Software

Data-Miner Pty. Ltd. has provided a free software license for those who have purchased the book. The software, which implements many of the methods discussed in the book, can be downloaded from the data-miner.com Web site.

Acknowledgements

Some of the case studies in Chapter 7 are based on our prior publications. In those projects, we acknowledge the participation of Chidanand Apté, Radu Florian, Abraham Ittycheriah, Vijay Iyengar, Hongyan Jing, David Johnson, Frank Oles, Naval Verma, and Brian White. Arindam Banerjee made many helpful comments on a draft of our book. We thank our editors, Wayne Wheeler, Ann Kostant, and Wayne Yuhasz, for their support. Our experiences in writing this book were quite enjoyable. We worked mostly on our own time, some of us located in different time zones, sometimes distant from home and communicating over the Internet. The four of us, three computer scientists and one linguist, are all colleagues and collaborators. Yet, we have worked in different areas, with substantial overlap in our approaches to text mining.

Sholom Weiss, Tong Zhang, and Fred Damerau - New York
Nitin Indurkhya - Australia and Brasil

Northern Summer and Southern Winter, 2004

Text Mining
Predictive Methods for Analyzing Unstructured
Information
Weiss, S.M.; Indurkhya, N.; Zhang, T.; Damerau, F.
2005, XII, 237 p., Hardcover