Preface

Previously we authored "Text Mining: Predictive Methods for Analyzing Unstructured Information." That book was geared mostly to professional practitioners, but was adaptable to course work with some effort by the instructor. Many topics were evolving, and this was one of the earliest efforts to collect material for predictive text mining. Since then, the book has seen extensive use in education, by ourselves, and other instructors, with positive responses from students. With more data sourced from the Internet, the field has seen very rapid growth with many new techniques that would interest practitioners. Given the amount of supplementary new material we had begun using, a new edition was clearly needed. A year ago, our publisher asked us to update the book and to add material that would extend its use as a textbook. We have revised many sections, adding new material to reflect the increased use of the web. Exercises and summaries are also provided.

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. Large text collections, often readily available on the Internet, contain valuable information that can be mined with today’s tools instead of waiting for tomorrow’s linguistic techniques. While others search for the essence of language understanding, we can immediately look for recurring word patterns in large collections of digital documents.

Our main theme is a strictly empirical view of text mining and an application of well-known analytical methods. 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 and case studies.

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. A few sections discuss advanced concepts that require mathematical maturity for a proper understanding. In such sections, intuitive explanations are also provided that may suffice for the less advanced reader. Most parts of the book can be read and understood by anyone with a sufficient analytic bend. 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.

For Instructors

The material in this book has been successfully used for education in a variety of ways ranging from short intensive one-week courses to twelve-week full semester courses. In short courses, the mathematical material can be skipped. The exercises have undergone class-testing over several years. Each chapter has the following accompanying material:

- a chapter summary
- exercises.

In addition, numerous examples and figures are interlaced throughout the book. Slides, sample solutions to selected exercises and suggestions for using the book in courses are are available from the publisher's companion site for this book.

Optional Software

AI Data-Miner LLC 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. Linux scripts for many examples are also available for download. The software requires familiarity with running command-line programs and editing configuration files. See http://www.data-miner.com for details.

Second Edition Updates

The book has been thoroughly revised and updated to reflect developments in the field since the first edition was published. The following new sections have been added to the second edition:

- Deep Learning
- Graph Modeling
- Mining Social Media
- Errors and Pitfalls in Big Data Evaluation
- Twitter Sentiment Analysis
- Introduction to Dependency Parsing.

Acknowledgments

Fred Damerau, our colleague and mentor, was a co-author of our original book. He is no longer with us, and his contributions to our project, especially his expertise in linguistics, were immeasurable. Some of the case studies in Chap. 8 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. The exercises in the book evolved from our text-mining course conducted regularly at statistics.com. We thank our editor, Wayne Wheeler, and our previous editors Ann Kostant and Wayne Yuhasz, for their support.

New York, USA — Sholom M. Weiss
Sydney, Australia and CA, USA — Nitin Indurkhya
Piscataway, NJ, USA — Tong Zhang

Springer

http://www.springer.com/978-1-4471-6749-5

Fundamentals of Predictive Text Mining
Weiss, S.M.; Indurkhya, N.; Zhang, T.
2015, XIII, 239 p. 115 illus., Hardcover
ISBN: 978-1-4471-6749-5