

Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications

Gary Miner

Tulsa, OK, USA

Dursun Delen

Tulsa, OK, USA

John Elder

Charlottesville, VA, USA

Andrew Fast

Charlottesville, VA, USA

Thomas Hill

Tulsa, OK, USA

Robert A. Nisbet

Santa Barbara, CA, USA

Major Guest Authors:

Jennifer Thompson

Woodward, OK, USA

Richard Foley

Raleigh, NC, USA

Angela Waner

Tulsa, OK, USA

Linda Winters-Miner

Tulsa, OK, USA

Karthik Balakrishnan

San Francisco, CA, USA



ELSEVIER

AMSTERDAM • BOSTON • HEIDELBERG • LONDON • NEW YORK • OXFORD
PARIS • SAN DIEGO • SAN FRANCISCO • SINGAPORE • SYDNEY • TOKYO

Academic Press is an imprint of Elsevier



Contents

ENDORSEMENTS FOR <i>PRACTICAL TEXT MINING & STATISTICAL ANALYSIS FOR NON-STRUCTURED TEXT DATA APPLICATIONS</i>	xi
FOREWORD 1	xv
FOREWORD 2	xvii
FOREWORD 3	xix
ACKNOWLEDGMENTS	xxi
PREFACE	xxiii
ABOUT THE AUTHORS	xxv
INTRODUCTION	xxxix
LIST OF TUTORIALS BY GUEST AUTHORS.....	xxxvii

Part I Basic Text Mining Principles

1. The History of Text Mining	3
2. The Seven Practice Areas of Text Analytics	29
3. Conceptual Foundations of Text Mining and Preprocessing Steps	43
4. Applications and Use Cases for Text Mining.....	53
5. Text Mining Methodology	73
6. Three Common Text Mining Software Tools	91

Part II Introduction to the Tutorial and Case Study Section of This Book

AA. CASE STUDY: Using the Social Share of Voice to Predict Events That Are about to Happen.....	127
BB. Mining Twitter for Airline Consumer Sentiment.....	133

A. Using <i>STATISTICA</i> Text Miner to Monitor and Predict Success of Marketing Campaigns Based on Social Media Data.....	151
B. Text Mining Improves Model Performance in Predicting Airplane Flight Accident Outcome.....	181
C. Insurance Industry: Text Analytics Adds “Lift” to Predictive Models with <i>STATISTICA</i> Text and Data Miner.....	203
D. Analysis of Survey Data for Establishing the “Best Medical Survey Instrument” Using Text Mining.....	233
E. Analysis of Survey Data for Establishing “Best Medical Survey Instrument” Using Text Mining: Central Asian (Russian Language) Study Tutorial 2: Potential for Constructing Instruments That Have Increased Validity.....	251
F. Using eBay Text for Predicting ATLAS Instrumental Learning.....	273
G. Text Mining for Patterns in Children’s Sleep Disorders Using <i>STATISTICA</i> Text Miner.....	357
H. Extracting Knowledge from Published Literature Using RapidMiner.....	375
I. Text Mining Speech Samples: Can the Speech of Individuals Diagnosed with Schizophrenia Differentiate Them from Unaffected Controls?.....	395
J. Text Mining Using STM™, CART®, and TreeNet® from Salford Systems: Analysis of 16,000 iPod Auctions on eBay.....	413
K. Predicting Micro Lending Loan Defaults Using SAS® Text Miner.....	417
L. Opera Lyrics: Text Analytics Compared by the Composer and the Century of Composition—Wagner versus Puccini.....	457
M. CASE STUDY: Sentiment-Based Text Analytics to Better Predict Customer Satisfaction and Net Promoter® Score Using IBM® SPSS® Modeler.....	509
N. CASE STUDY: Detecting Deception in Text with Freely Available Text and Data Mining Tools.....	533
O. Predicting Box Office Success of Motion Pictures with Text Mining.....	543
P. A Hands-On Tutorial of Text Mining in PASW: Clustering and Sentiment Analysis Using Tweets from Twitter.....	557
Q. A Hands-On Tutorial on Text Mining in SAS®: Analysis of Customer Comments for Clustering and Predictive Modeling.....	585

R. Scoring Retention and Success of Incoming College Freshmen Using Text Analytics	605
S. Searching for Relationships in Product Recall Data from the Consumer Product Safety Commission with <i>STATISTICA</i> Text Miner.....	645
T. Potential Problems That Can Arise in Text Mining: Example Using NALL Aviation Data.....	657
U. Exploring the Unabomber Manifesto Using Text Miner.....	681
V. Text Mining PubMed: Extracting Publications on Genes and Genetic Markers Associated with Migraine Headaches from PubMed Abstracts	703
W. CASE STUDY: The Problem with the Use of Medical Abbreviations by Physicians and Health Care Providers.....	751
X. Classifying Documents with Respect to “Earnings” and Then Making a Predictive Model for the Target Variable Using Decision Trees, MARSplines, Naïve Bayes Classifier, and K-Nearest Neighbors with <i>STATISTICA</i> Text Miner	773
Y. CASE STUDY: Predicting Exposure of Social Messages: The Bin Laden Live Tweeter	797
Z. The InFLUence Model: Web Crawling, Text Mining, and Predictive Analysis with 2010–2011 Influenza Guidelines—CDC, IDSA, WHO, and FMC	803

Part III Advanced Topics

7. Text Classification and Categorization.....	881
8. Prediction in Text Mining: The Data Mining Algorithms of Predictive Analytics	893
9. Entity Extraction.....	921
10. Feature Selection and Dimensionality Reduction.....	929
11. Singular Value Decomposition in Text Mining	935
12. Web Analytics and Web Mining.....	949
13. Clustering Words and Documents.....	959
14. Leveraging Text Mining in Property and Casualty Insurance.....	967
15. Focused Web Crawling.....	983

16. The Future of Text and Web Analytics.....	991
17. Summary.....	1007
GLOSSARY.....	1017
INDEX.....	1025
HOW TO USE THE DATA SETS AND THE TEXT MINING SOFTWARE ON THE DVD OR ON LINKS FOR PRACTICAL TEXT MINING	1047