Text Data Mining

Chengqing Zong • Rui Xia • Jiajun Zhang

Text Data Mining

Springer

Chengqing Zong
Institute of Automation
Chinese Academy of Sciences
Beijing, Beijing, China

Rui Xia
School of Computer Science & Engineering
Nanjing University of Science
and Technology
Nanjing, Jiangsu, China

Jiajun Zhang
Institute of Automation
Chinese Academy of Sciences
Beijing, Beijing, China

ISBN 978-981-16-0099-9 ISBN 978-981-16-0100-2 (eBook)
https://doi.org/10.1007/978-981-16-0100-2

Jointly published with Tsinghua University Press
The print edition is not for sale in China (Mainland). Customers from China (Mainland) please order the print book from: Tsinghua University Press.

This Springer imprint is published by the registered company Springer Nature Singapore Pte Ltd.
The registered company address is: 152 Beach Road, #21-01/04 Gateway East, Singapore 189721, Singapore

Foreword

We are living in the Big Data era. Over 80% of real-world data are unstructured in the form of natural language text, such as books, news reports, research articles, social media messages, and webpages. Although data mining and machine learning have been popular in data analysis, most data mining methods handle only structured or semistructured data. In comparison with mining structured data, mining unstructured text data is more challenging but will also play a more essential role in turning massive data into structured knowledge. It is no wonder we have witnessed such a dramatic upsurge in the research on text mining and natural language processing and their applications in recent years.

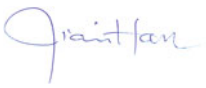
Text mining is a confluence of natural language processing, data mining, machine learning, and statistics used to mine knowledge from unstructured text. There have already been multiple textbooks dedicated to data mining, machine learning, statistics, and natural language processing. However, we seriously lack textbooks on text mining that systematically introduce important topics and up-to-date methods for text mining. This book, "Text Data Mining," bridges this gap nicely. It is the first textbook, and a brilliant one, on text data mining that not only introduces foundational issues but also offers comprehensive and state-of-the-art coverage of the important and ongoing research themes on text mining. The in-depth treatment of a wide spectrum of text mining themes and clear introduction to the state-of-the-art deep learning methods for text mining make the book unique, timely, and authoritative. It is a great textbook for graduate students as well as a valuable handbook for practitioners working on text mining, natural language processing, data mining, and machine learning and their applications. This book is written by three pioneering researchers and highly reputed experts in the fields of natural language processing and text mining. The first author has written an authoritative and popular textbook on natural language processing that has been adopted as a standard textbook for university undergraduate and first-year graduate students in China. However, this new text mining book has completely different coverage from his NLP textbook and offers new and complementary text mining themes. Both books can be studied independently, although I would strongly encourage students working on NLP and text mining to study both.

This text mining book starts with text preprocessing, including both English and Chinese text preprocessing, and proceeds to text representation, covering the vector space model and distributed representation of words, phrases, sentences, and documents, in both statistical modeling and deep learning models. It then introduces feature selection methods, statistical learning methods, and deep neural network methods, including multilayer feed-forward neural networks, convolutional neural networks, and recurrent neural networks, for document classification. It next proceeds to text clustering, covering sample and cluster similarities, various clustering methods, and clustering evaluation. After introducing the fundamental theories and methods of text mining, the book uses five chapters to cover a wide spectrum of text mining applications, including topic modeling (which is also treated as a fundamental issue from some viewpoints but can be used independently), sentiment analysis and opinion mining, topic detection and tracking, information extraction, and automated document summarization. These themes are active research frontiers in text mining and are covered comprehensively and thoroughly, with a good balance between classical methods and recent developments, including deep learning methods.

As a data mining researcher, I have recently been deeply involved in text mining due to the need to handle the large scale of real-world data. I could not find a good text mining textbook written in English or Chinese to learn and teach. It is exciting to see that this book provides such a comprehensive and cutting-edge introduction. I believe this book will benefit data science researchers, graduate students, and those who want to include text mining in practical applications. I loved reading this book and recommend it highly to everyone who wants to learn text mining!

ACM Fellow and IEEE Fellow
Michael Aiken Chair Professor
Department of Computer Science
University of Illinois at Urbana-Champaign
Champaign, IL, USA

Jiawei Han



Preface

With the rapid development and popularization of Internet and mobile communication technologies, text data mining has attracted much attention. In particular, with the wide use of new technologies such as cloud computing, big data, and deep learning, text mining has begun playing an increasingly important role in many application fields, such as opinion mining and medical and financial data analysis, showing broad application prospects.

Although I was supervising graduate students studying text classification and automatic summarization more than ten years ago, I did not have a clear understanding of the overall concept of text data mining and only regarded the research topics as specific applications of natural language processing. Professor Jiawei Han's book *Data Mining: Concepts and Technology*, published by Elsevier, Professor Bing Liu's *Web Data Mining*, published by Springer, and other books have greatly benefited me. Every time I listen to their talks and discuss these topics with them face to face, I have benefited immensely. I was inspired to write this book for the course Text Data Mining, which I was invited to teach to graduates of the University of Chinese Academy of Sciences. At the end of 2015, I accepted the invitation and began to prepare the content design and selection of materials for the course. I had to study a large number of related papers, books, and other materials and began to seriously think of the rich connotation and extension of the term Text Data Mining. After more than a year's study, I started to compile the courseware. With teaching practice, the outline of the concept has gradually formed.

Rui Xia and Jiajun Zhang, two talented young people, helped me materialize my original writing plan. Rui Xia received his master's degree in 2007 and was admitted to the Institute of Automation, Chinese Academy of Sciences, and studied for Ph.D. degree under my supervision. He was engaged in sentiment classification and took it as the research topic of his Ph.D. dissertation. After he received his Ph.D. degree in 2011, his interests extended to opinion mining, text clustering and classification, topic modeling, event detection and tracking, and other related topics. He has published a series of influential papers in the field of sentiment analysis and opinion mining. He received the ACL 2019 outstanding paper award, and his paper on ensemble learning for sentiment classification has been cited more than

600 times. Jiajun Zhang joined our institute after he graduated from university in 2006 and studied in my group in pursuit of his Ph.D. degree. He mainly engaged in machine translation research, but he performed well in many research topics, such as multilanguage automatic summarization, information extraction, and human–computer dialogue systems. Since 2016, he has been teaching some parts of the course on Natural Language Processing in cooperation with me, such as machine translation, automatic summarization, and text classification, at the University of Chinese Academy of Sciences; this course is very popular with students. With the solid theoretical foundation of these two talents and their keen scientific insights, I am gratified that many cutting-edge technical methods and research results could be verified and practiced and included in this book.

From early 2016 to June 2019, when the Chinese version of this book was published, it took more than three years. In these three years, most holidays, weekends, and other spare times of ours were devoted to the writing of this book. It was really suffering to endure the numerous modifications or even rewriting, but we were also very happy. We started to translate the Chinese version into English in the second half of 2019. Some more recent topics, including BERT (bidirectional encoder representations from transformers), have been added to the English version.

As a cross domain of natural language processing and machine learning, text data mining faces the double challenges of the two domains and has broad application to the Internet and equipment for mobile communication. The topics and techniques presented in this book are all the technical foundations needed to develop such practical systems and have attracted much attention in recent years. It is hoped that this book will provide a comprehensive understanding for students, professors, and researchers in related areas. However, I must admit that due to the limitation of the authors' ability and breadth of knowledge, as well as the lack of time and energy, there must be some omissions or mistakes in the book. We will be very grateful if readers provide criticism, corrections, and any suggestions.

Beijing, China
20 May 2020

Chengqing Zong

Acknowledgments

During the writing process and after the completion of the first draft of the Chinese version of this book, many experts from related fields reviewed selected chapters and gave us valuable comments and suggestions. They are (in alphabetical order) Xianpei Han, Yu Hong, Shoushan Li, Kang Liu, Xiaojun Wan, Kang Xu, Chengzhi Zhang, and Xin Zhao. In addition, we also received help from several graduate students and Ph.D. candidates (also in alphabetical order): Hongjie Cai, Zixiang Ding, Huihui He, Xiao Jin, Junjie Li, Mei Li, Yuchen Liu, Cong Ma, Liqun Ma, Xiangqing Shen, Jingyuan Sun, Fanfan Wang, Leyi Wang, Qain Wang, Weikang Wang, Yining Wang, Kaizhou Xuan, Shiliang Zheng, and Long Zhou. They helped us to double check and confirm English expressions, references, and web addresses and to redraw several charts in the book, which saved us much time. We would like to express our heartfelt thanks to all of them!

We also want to sincerely thank Professor Jiawei Han for his guidance and suggestions on this book. We are honored that he was willing to write the foreword to this book despite his busy schedule. Finally, we would like to recognize Ms. Hui Xue and Qian Wang, Tsinghua University Press, and Ms. Celine Chang, and Ms. Suraj Kumar, Springer, for their great help!

Beijing, China — Chengqing Zong
Nanjing, China — Rui Xia
Beijing, China — Jiajun Zhang
20 May 2020

Contents

About the Authors

Chengqing Zong is a Professor at the National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences (CASIA) and an adjunct professor in the School of Artificial Intelligence at University of Chinese Academy of Sciences (UCAS). He authored the book "Statistical Natural Language Processing" (which is in Chinese and sold more than 32K copies) and has published more than 200 papers on machine translation, natural language processing, and cognitive linguistics. He served as the chair for numerous prestigious conferences, such as ACL, COLING, AAAI, and IJCAI, and has served as an associate editor for journals, such as ACM TALLIP and ACTA Automatic Sinica, and as an editorial board member for journals, including IEEE Intelligent Systems, Journal of Computer Science and Technology, and Machine Translation. He is currently the President of the Asian Federation of Natural Language Processing (AFNLP) and a member of the International Committee on Computational Linguistics (ICCL).

Rui Xia is a Professor at the School of Computer Science and Engineering, Nanjing University of Science and Technology, China. He has published more than 50 papers in high-quality journals and top-tiered conferences in the field of natural language processing and text data mining. He serves as area chair and senior program committee member for several top conferences, such as EMNLP, COLING, IJCAI, and AAAI. He received the outstanding paper award of ACL 2019, and the Distinguished Young Scholar award from the Natural Science Foundation of Jiangsu Province, China in 2020.

Jiajun Zhang is a Professor at the NLPR, CASIA and an adjunct professor in the SAIU of UCAS. He has published more than 80 conference papers and journal articles on natural language processing and text mining and received 5 best paper awards. He served as the area chair or on the senior program committees for several top conferences, such as ACL, EMNLP, COLING, AAAI, and IJCAI. He is the

deputy director of China's Machine Translation Technical Committee of the Chinese Information Processing Society of China. He received the Qian Wei-Chang Science and Technology Award of Chinese Information Processing and the CIPS Hanvon Youth Innovation Award. He was supported by the Elite Scientists Sponsorship Program of China Association for Science and Technology (CAST).

Acronyms

ACE	Automatic content extraction
AMR	Abstract meaning representation
ATT	Adaptive topic tracking
AUC	Area under the ROC curve
Bagging	Bootstrap aggregating
BERT	Bidirectional encoder representations from transformer
BFGS	Broyden–Fletcher–Goldfarb–Shanno
Bi-LSTM	Bidirectional long short-term memory
BIO	Begin–inside–outside
BLEU	Bilingual evaluation understudy
BOW	Bag of words
BP	Back-propagation
BPTS	Back-propagation through structure
BPTT	Back-propagation through time
BRAE	Bilingually constrained recursive autoencoder
CBOW	Continuous bag-of-words
CFG	Context-free grammar
CNN	Convolutional neural network
CRF	Conditional random field
C&W	Collobert and Weston
CWS	Chinese word segmentation
DBSCAN	Density-based spatial clustering of applications with noise
DF	Document frequency
DL	Deep learning
DMN	Deep memory network
ELMo	Embeddings from language models
EM	Expectation maximization
EP	Expectation propagation
FAR	False alarm rate
FNN	Feed-forward neural network
GPT	Generative pretraining

GRU	Gated recurrent unit
HAC	Hierarchical agglomerative clustering
HMM	Hidden Markov model
HTD	Hierarchical topic detection
ICA	Independent component analysis
IDF	Inverse document frequency
IE	Information extraction
IG	Information gain
KBP	Knowledge base population
KDD	Knowledge discovery in databases
KKT	Karush–Kuhn–Tucker
K-L	Kullback–Leibler
L-BFGS	Limited-memory Broyden–Fletcher–Goldfarb–Shanno
LDA	Latent Dirichlet allocation
LM	Language model
LSA	Latent semantic analysis
LSI	Latent semantic indexing
LSTM	Long short-term memory
MCMC	Markov chain Monte Carlo
MDR	Missed detection rate
ME	Maximum entropy
MI	Mutual information
ML	Machine learning
MLE	Maximum likelihood estimation
MMR	Maximum marginal relevance
MUC	Message understanding conference
NB	Naïve Bayes
NCE	Noise contrastive estimation
NED	New event detection
NER	Named entity recognition
NLP	Natural language processing
NNLM	Neural network language model
PCA	Principal component analysis
PLSA	Probabilistic latent semantic analysis
PLSI	Probabilistic latent semantic indexing
PMI	Pointwise mutual information
P-R	Precision–recall
PU	Positive-unlabeled
PCFG	Probabilistic context-free grammar
PMI-IR	Pointwise mutual information—information retrieval
POS	Part of speech
PV-DBOW	Distributed bag-of-words version of the paragraph vector
PV-DM	Paragraph vector with sentence as distributed memory
Q&A	Question and answering
RAE	Recursive autoencoder

RecurNN	Recursive neural network
RG	Referent graph
RNN	Recurrent neural network
ROC	Receiver operating characteristic
ROUGE	Recall-oriented understudy for gisting evaluation
RTD	Retrospective topic detection
SCL	Structure correspondence learning
SCU	Summary content unit
SMO	Sequential minimal optimization
SO	Semantic orientation
SRL	Semantic role labeling
SS	Story segmentation
SST	Stanford sentiment treebank
STC	Suffix tree clustering
SVD	Singular value decomposition
SVM	Support vector machine
TAC	Text analysis conference
TD	Topic detection
TDT	Topic detection and tracking
TF	Term frequency
TF-IDF	Term frequency—inverted document frequency
UGC	User-generated context
UniLM	Unified pretraining language model
VBEM	Variational Bayes expectation maximization
VSM	Vector space model
WCSS	Within-cluster sum of squares
WSD	Word sense disambiguation