The Text Mining Handbook

Text mining is a new and exciting area of computer science research that tries to solve the crisis of information overload by combining techniques from data mining, machine learning, natural language processing, information retrieval, and knowledge management. Similarly, link detection – a rapidly evolving approach to the analysis of text that shares and builds on many of the key elements of text mining – also provides new tools for people to better leverage their burgeoning textual data resources. Link detection relies on a process of building up networks of interconnected objects through various relationships in order to discover patterns and trends. The main tasks of link detection are to extract, discover, and link together sparse evidence from vast amounts of data sources, to represent and evaluate the significance of the related evidence, and to learn patterns to guide the extraction, discovery, and linkage of entities.

The Text Mining Handbook presents a comprehensive discussion of the state of the art in text mining and link detection. In addition to providing an in-depth examination of core text mining and link detection algorithms and operations, the work examines advanced preprocessing techniques, knowledge representation considerations, and visualization approaches. Finally, the book explores current real-world, mission-critical applications of text mining and link detection in such varied fields as corporate finance business intelligence, genomics research, and counterterrorism activities.

Dr. Ronen Feldman is a Senior Lecturer in the Mathematics and Computer Science Department of Bar-Ilan University and Director of the Data and Text Mining Laboratory. Dr. Feldman is cofounder, Chief Scientist, and President of ClearForest, Ltd., a leader in developing next-generation text mining applications for corporate and government clients. He also recently served as an Adjunct Professor at New York University's Stern School of Business. A pioneer in the areas of machine learning, data mining, and unstructured data management, he has authored or coauthored more than 70 published articles and conference papers in these areas.

James Sanger is a venture capitalist, applied technologist, and recognized industry expert in the areas of commercial data solutions, Internet applications, and IT security products. He is a partner at ABS Ventures, an independent venture firm founded in 1982 and originally associated with technology banking leader Alex. Brown and Sons. Immediately before joining ABS Ventures, Mr. Sanger was a Managing Director in the New York offices of DB Capital Venture Partners, the global venture capital arm of Deutsche Bank. Mr. Sanger has been a board member of several thought-leading technology companies, including Inxight Software, Gomez Inc., and ClearForest, Inc.; he has also served as an official observer to the boards of AlphaBlox (acquired by IBM in 2004), Intralinks, and Imagine Software and as a member of the Technical Advisory Board of Qualys, Inc.

THE TEXT MINING HANDBOOK

Advanced Approaches in Analyzing Unstructured Data

Ronen Feldman
Bar-Ilan University, Israel

James Sanger
ABS Ventures, Waltham, Massachusetts

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK
Published in the United States of America by Cambridge University Press, New York
www.cambridge.org
Information on this title: www.cambridge.org/9780521836579

First published in print format 2006

ISBN-13 978-0-511-54691-4 OCeISBN

ISBN-13 978-0-521-83657-9 hardback
ISBN-10 0-521-83657-3 hardback

In loving memory of my father, Issac Feldman