

THE TEXT MINING HANDBOOK

Advanced Approaches in
Analyzing Unstructured Data

The screenshot shows a software interface for text mining. At the top, there is a list of news headlines. Below this is a table with three columns: 'Select Objects', 'Objects Selected', and 'Query Results'. The 'Select Objects' column lists various countries. The 'Objects Selected' column shows 'uk' and 'usa' are selected. The 'Query Results' column shows counts for various terms like 'acq', 'money_fx', 'trade', etc.

OK
Cancel
Help
Run Query
Graph

G-7 ISSUES STATEMENT AFTER MEETING
U.S. SAID UNFAIRLY PROTECTING DEFENSE INDUSTRY
BRITISH FARM MINISTER ATTACKS SUBSIDIES
GATT TO DEBATE U.S. CHARGES OF AIRBUS SUBSIDIES
BAKER DENIES DOLLAR TARGET EXISTS
UK MAY REVOKE JAPANESE FINANCIAL LICENSES
U.S. WINE EXPORTS ROSE 15 PER CENT LAST YEAR
JAPAN ISOLATED, YEN RISES, WORLD FEELS CHEATED

	Select Objects	Objects Selected	Query Results
First Category	arab	uk	acq : 42
countries	argentina	usa	money_fx : 32
	aruba		trade : 24
Second Category	australia		corp_news : 19
topics	austria		dlr : 15
	bahamas		cbond : 10
	bahrain		loan : 9
	balladur		ebond : 8
	bangladesh		ven : 8

Clear Selection

RONEN FELDMAN
JAMES SANGER

CAMBRIDGE

Contents

Preface

page x

I. Introduction to Text Mining	1
I.1 Defining Text Mining	1
I.2 General Architecture of Text Mining Systems	13
II. Core Text Mining Operations	19
II.1 Core Text Mining Operations	19
II.2 Using Background Knowledge for Text Mining	41
II.3 Text Mining Query Languages	51
III. Text Mining Preprocessing Techniques	57
III.1 Task-Oriented Approaches	58
III.2 Further Reading	62
IV. Categorization	64
IV.1 Applications of Text Categorization	65
IV.2 Definition of the Problem	66
IV.3 Document Representation	68
IV.4 Knowledge Engineering Approach to TC	70
IV.5 Machine Learning Approach to TC	70
IV.6 Using Unlabeled Data to Improve Classification	78
IV.7 Evaluation of Text Classifiers	79
IV.8 Citations and Notes	80
V. Clustering	82
V.1 Clustering Tasks in Text Analysis	82
V.2 The General Clustering Problem	84
V.3 Clustering Algorithms	85
V.4 Clustering of Textual Data	88
V.5 Citations and Notes	92

VI. Information Extraction	94
VI.1 Introduction to Information Extraction	94
VI.2 Historical Evolution of IE: The Message Understanding Conferences and Tipster	96
VI.3 IE Examples	101
VI.4 Architecture of IE Systems	104
VI.5 Anaphora Resolution	109
VI.6 Inductive Algorithms for IE	119
VI.7 Structural IE	122
VI.8 Further Reading	129
VII. Probabilistic Models for Information Extraction	131
VII.1 Hidden Markov Models	131
VII.2 Stochastic Context-Free Grammars	137
VII.3 Maximal Entropy Modeling	138
VII.4 Maximal Entropy Markov Models	140
VII.5 Conditional Random Fields	142
VII.6 Further Reading	145
VIII. Preprocessing Applications Using Probabilistic and Hybrid Approaches	146
VIII.1 Applications of HMM to Textual Analysis	146
VIII.2 Using MEMM for Information Extraction	152
VIII.3 Applications of CRFs to Textual Analysis	153
VIII.4 TEG: Using SCFG Rules for Hybrid Statistical–Knowledge-Based IE	155
VIII.5 Bootstrapping	166
VIII.6 Further Reading	175
IX. Presentation-Layer Considerations for Browsing and Query Refinement	177
IX.1 Browsing	177
IX.2 Accessing Constraints and Simple Specification Filters at the Presentation Layer	185
IX.3 Accessing the Underlying Query Language	186
IX.4 Citations and Notes	187
X. Visualization Approaches	189
X.1 Introduction	189
X.2 Architectural Considerations	192
X.3 Common Visualization Approaches for Text Mining	194
X.4 Visualization Techniques in Link Analysis	225
X.5 Real-World Example: The Document Explorer System	235
XI. Link Analysis	244
XI.1 Preliminaries	244

XI.2 Automatic Layout of Networks	246
XI.3 Paths and Cycles in Graphs	250
XI.4 Centrality	251
XI.5 Partitioning of Networks	259
XI.6 Pattern Matching in Networks	272
XI.7 Software Packages for Link Analysis	273
XI.8 Citations and Notes	274
XII. Text Mining Applications	275
XII.1 General Considerations	276
XII.2 Corporate Finance: Mining Industry Literature for Business Intelligence	281
XII.3 A “Horizontal” Text Mining Application: Patent Analysis Solution Leveraging a Commercial Text Analytics Platform	297
XII.4 Life Sciences Research: Mining Biological Pathway Information with GeneWays	309
Appendix A: DIAL: A Dedicated Information Extraction Language for Text Mining	317
A.1 What Is the DIAL Language?	317
A.2 Information Extraction in the DIAL Environment	318
A.3 Text Tokenization	320
A.4 Concept and Rule Structure	320
A.5 Pattern Matching	322
A.6 Pattern Elements	323
A.7 Rule Constraints	327
A.8 Concept Guards	328
A.9 Complete DIAL Examples	329
<i>Bibliography</i>	337
<i>Index</i>	391