

# Text Mining

## Applications and Theory

**Michael W. Berry**

*University of Tennessee, USA*

**Jacob Kogan**

*University of Maryland Baltimore County, USA*



A John Wiley and Sons, Ltd., Publication

# Contents

<b>List of Contributors</b>	<b>xi</b>
<b>Preface</b>	<b>xiii</b>
<b>PART I TEXT EXTRACTION, CLASSIFICATION, AND CLUSTERING</b>	<b>1</b>
<b>1 Automatic keyword extraction from individual documents</b>	<b>3</b>
1.1 Introduction	3
1.1.1 Keyword extraction methods	4
1.2 Rapid automatic keyword extraction	5
1.2.1 Candidate keywords	6
1.2.2 Keyword scores	7
1.2.3 Adjoining keywords	8
1.2.4 Extracted keywords	8
1.3 Benchmark evaluation	9
1.3.1 Evaluating precision and recall	9
1.3.2 Evaluating efficiency	10
1.4 Stoplist generation	11
1.5 Evaluation on news articles	15
1.5.1 The MPQA Corpus	15
1.5.2 Extracting keywords from news articles	15
1.6 Summary	18
1.7 Acknowledgements	19
References	19
<b>2 Algebraic techniques for multilingual document clustering</b>	<b>21</b>
2.1 Introduction	21
2.2 Background	22
2.3 Experimental setup	23
2.4 Multilingual LSA	25
2.5 Tucker1 method	27

2.6	PARAFAC2 method	28
2.7	LSA with term alignments	29
2.8	Latent morpho-semantic analysis (LMSA)	32
2.9	LMSA with term alignments	33
2.10	Discussion of results and techniques	33
2.11	Acknowledgements	35
	References	35
<b>3</b>	<b>Content-based spam email classification using machine-learning algorithms</b>	<b>37</b>
3.1	Introduction	37
3.2	Machine-learning algorithms	39
3.2.1	Naive Bayes	39
3.2.2	LogitBoost	40
3.2.3	Support vector machines	41
3.2.4	Augmented latent semantic indexing spaces	43
3.2.5	Radial basis function networks	44
3.3	Data preprocessing	45
3.3.1	Feature selection	45
3.3.2	Message representation	47
3.4	Evaluation of email classification	48
3.5	Experiments	49
3.5.1	Experiments with PU1	49
3.5.2	Experiments with ZH1	51
3.6	Characteristics of classifiers	53
3.7	Concluding remarks	54
3.8	Acknowledgements	55
	References	55
<b>4</b>	<b>Utilizing nonnegative matrix factorization for email classification problems</b>	<b>57</b>
4.1	Introduction	57
4.1.1	Related work	59
4.1.2	Synopsis	60
4.2	Background	60
4.2.1	Nonnegative matrix factorization	60
4.2.2	Algorithms for computing NMF	61
4.2.3	Datasets	63
4.2.4	Interpretation	64
4.3	NMF initialization based on feature ranking	65
4.3.1	Feature subset selection	66
4.3.2	FS initialization	66
4.4	NMF-based classification methods	70
4.4.1	Classification using basis features	70
4.4.2	Generalizing LSI based on NMF	72

4.5	Conclusions	78
4.6	Acknowledgements	79
	References	79
<b>5</b>	<b>Constrained clustering with <math>k</math>-means type algorithms</b>	<b>81</b>
5.1	Introduction	81
5.2	Notations and classical $k$ -means	82
5.3	Constrained $k$ -means with Bregman divergences	84
5.3.1	Quadratic $k$ -means with cannot-link constraints	84
5.3.2	Elimination of must-link constraints	87
5.3.3	Clustering with Bregman divergences	89
5.4	Constrained smoka type clustering	92
5.5	Constrained spherical $k$ -means	95
5.5.1	Spherical $k$ -means with cannot-link constraints only	96
5.5.2	Spherical $k$ -means with cannot-link and must-link constraints	98
5.6	Numerical experiments	99
5.6.1	Quadratic $k$ -means	100
5.6.2	Spherical $k$ -means	100
5.7	Conclusion	101
	References	102
<b>PART II ANOMALY AND TREND DETECTION</b>		<b>105</b>
<b>6</b>	<b>Survey of text visualization techniques</b>	<b>107</b>
6.1	Visualization in text analysis	107
6.2	Tag clouds	108
6.3	Authorship and change tracking	110
6.4	Data exploration and the search for novel patterns	111
6.5	Sentiment tracking	111
6.6	Visual analytics and FutureLens	113
6.7	Scenario discovery	114
6.7.1	Scenarios	115
6.7.2	Evaluating solutions	115
6.8	Earlier prototype	116
6.9	Features of FutureLens	117
6.10	Scenario discovery example: bioterrorism	119
6.11	Scenario discovery example: drug trafficking	121
6.12	Future work	123
	References	126
<b>7</b>	<b>Adaptive threshold setting for novelty mining</b>	<b>129</b>
7.1	Introduction	129
7.2	Adaptive threshold setting in novelty mining	131

7.2.1	Background	131
7.2.2	Motivation	132
7.2.3	Gaussian-based adaptive threshold setting	132
7.2.4	Implementation issues	137
7.3	Experimental study	138
7.3.1	Datasets	138
7.3.2	Working example	139
7.3.3	Experiments and results	142
7.4	Conclusion	146
	References	147
<b>8</b>	<b>Text mining and cybercrime</b>	<b>149</b>
8.1	Introduction	149
8.2	Current research in Internet predation and cyberbullying	151
8.2.1	Capturing IM and IRC chat	151
8.2.2	Current collections for use in analysis	152
8.2.3	Analysis of IM and IRC chat	153
8.2.4	Internet predation detection	153
8.2.5	Cyberbullying detection	158
8.2.6	Legal issues	159
8.3	Commercial software for monitoring chat	159
8.4	Conclusions and future directions	161
8.5	Acknowledgements	162
	References	162
	<b>PART III TEXT STREAMS</b>	<b>165</b>
<b>9</b>	<b>Events and trends in text streams</b>	<b>167</b>
9.1	Introduction	167
9.2	Text streams	169
9.3	Feature extraction and data reduction	170
9.4	Event detection	171
9.5	Trend detection	174
9.6	Event and trend descriptions	176
9.7	Discussion	180
9.8	Summary	181
9.9	Acknowledgements	181
	References	181
<b>10</b>	<b>Embedding semantics in LDA topic models</b>	<b>183</b>
10.1	Introduction	183
10.2	Background	184

10.2.1	Vector space modeling	184
10.2.2	Latent semantic analysis	185
10.2.3	Probabilistic latent semantic analysis	185
10.3	Latent Dirichlet allocation	186
10.3.1	Graphical model and generative process	187
10.3.2	Posterior inference	187
10.3.3	Online latent Dirichlet allocation (OLDA)	189
10.3.4	Illustrative example	191
10.4	Embedding external semantics from Wikipedia	193
10.4.1	Related Wikipedia articles	194
10.4.2	Wikipedia-influenced topic model	194
10.5	Data-driven semantic embedding	194
10.5.1	Generative process with data-driven semantic embedding	195
10.5.2	OLDA algorithm with data-driven semantic embedding	196
10.5.3	Experimental design	197
10.5.4	Experimental results	199
10.6	Related work	202
10.7	Conclusion and future work	202
	References	203
	<b>Index</b>	<b>205</b>