

# Contents

<b>1</b>	<b>Overview of Text Mining</b> . . . . .	<b>1</b>
1.1	What's Special About Text Mining? . . . . .	1
1.1.1	Structured or Unstructured Data? . . . . .	2
1.1.2	Is Text Different from Numbers? . . . . .	3
1.2	What Types of Problems Can Be Solved? . . . . .	5
1.3	Document Classification . . . . .	6
1.4	Information Retrieval . . . . .	6
1.5	Clustering and Organizing Documents . . . . .	7
1.6	Information Extraction . . . . .	8
1.7	Prediction and Evaluation . . . . .	9
1.8	The Next Chapters . . . . .	10
1.9	Summary . . . . .	10
1.10	Historical and Bibliographical Remarks . . . . .	11
1.11	Questions and Exercises . . . . .	12
<b>2</b>	<b>From Textual Information to Numerical Vectors</b> . . . . .	<b>13</b>
2.1	Collecting Documents . . . . .	13
2.2	Document Standardization . . . . .	15
2.3	Tokenization . . . . .	16
2.4	Lemmatization . . . . .	17
2.4.1	Inflectional Stemming . . . . .	19
2.4.2	Stemming to a Root . . . . .	19
2.5	Vector Generation for Prediction . . . . .	21
2.5.1	Multiword Features . . . . .	26
2.5.2	Labels for the Right Answers . . . . .	28
2.5.3	Feature Selection by Attribute Ranking . . . . .	29
2.6	Sentence Boundary Determination . . . . .	29
2.7	Part-of-Speech Tagging . . . . .	31
2.8	Word Sense Disambiguation . . . . .	32
2.9	Phrase Recognition . . . . .	32
2.10	Named Entity Recognition . . . . .	33

2.11	Parsing . . . . .	33
2.12	Feature Generation . . . . .	35
2.13	Summary . . . . .	36
2.14	Historical and Bibliographical Remarks . . . . .	36
2.15	Questions and Exercises . . . . .	38
<b>3</b>	<b>Using Text for Prediction . . . . .</b>	<b>39</b>
3.1	Recognizing that Documents Fit a Pattern . . . . .	41
3.2	How Many Documents Are Enough? . . . . .	42
3.3	Document Classification . . . . .	43
3.4	Learning to Predict from Text . . . . .	44
3.4.1	Similarity and Nearest-Neighbor Methods . . . . .	45
3.4.2	Document Similarity . . . . .	46
3.4.3	Decision Rules . . . . .	48
3.4.4	Decision Trees . . . . .	54
3.4.5	Scoring by Probabilities . . . . .	55
3.4.6	Linear Scoring Methods . . . . .	58
3.5	Evaluation of Performance . . . . .	66
3.5.1	Estimating Current and Future Performance . . . . .	66
3.5.2	Getting the Most from a Learning Method . . . . .	69
3.6	Applications . . . . .	69
3.7	Summary . . . . .	70
3.8	Historical and Bibliographical Remarks . . . . .	70
3.9	Questions and Exercises . . . . .	72
<b>4</b>	<b>Information Retrieval and Text Mining . . . . .</b>	<b>75</b>
4.1	Is Information Retrieval a Form of Text Mining? . . . . .	75
4.2	Key Word Search . . . . .	76
4.3	Nearest-Neighbor Methods . . . . .	77
4.4	Measuring Similarity . . . . .	78
4.4.1	Shared Word Count . . . . .	78
4.4.2	Word Count and Bonus . . . . .	78
4.4.3	Cosine Similarity . . . . .	79
4.5	Web-based Document Search . . . . .	80
4.5.1	Link Analysis . . . . .	81
4.6	Document Matching . . . . .	85
4.7	Inverted Lists . . . . .	85
4.8	Evaluation of Performance . . . . .	87
4.9	Summary . . . . .	88
4.10	Historical and Bibliographical Remarks . . . . .	88
4.11	Questions and Exercises . . . . .	89
<b>5</b>	<b>Finding Structure in a Document Collection . . . . .</b>	<b>91</b>
5.1	Clustering Documents by Similarity . . . . .	93
5.2	Similarity of Composite Documents . . . . .	94
5.2.1	<i>k</i> -Means Clustering . . . . .	96

5.2.2	Hierarchical Clustering . . . . .	99
5.2.3	The EM Algorithm . . . . .	102
5.3	What Do a Cluster's Labels Mean? . . . . .	105
5.4	Applications . . . . .	107
5.5	Evaluation of Performance . . . . .	108
5.6	Summary . . . . .	110
5.7	Historical and Bibliographical Remarks . . . . .	110
5.8	Questions and Exercises . . . . .	111
<b>6</b>	<b>Looking for Information in Documents . . . . .</b>	<b>113</b>
6.1	Goals of Information Extraction . . . . .	113
6.2	Finding Patterns and Entities from Text . . . . .	115
6.2.1	Entity Extraction as Sequential Tagging . . . . .	116
6.2.2	Tag Prediction as Classification . . . . .	117
6.2.3	The Maximum Entropy Method . . . . .	118
6.2.4	Linguistic Features and Encoding . . . . .	123
6.2.5	Local Sequence Prediction Models . . . . .	124
6.2.6	Global Sequence Prediction Models . . . . .	128
6.3	Coreference and Relationship Extraction . . . . .	129
6.3.1	Coreference Resolution . . . . .	129
6.3.2	Relationship Extraction . . . . .	131
6.4	Template Filling and Database Construction . . . . .	132
6.5	Applications . . . . .	133
6.5.1	Information Retrieval . . . . .	133
6.5.2	Commercial Extraction Systems . . . . .	134
6.5.3	Criminal Justice . . . . .	135
6.5.4	Intelligence . . . . .	135
6.6	Summary . . . . .	136
6.7	Historical and Bibliographical Remarks . . . . .	137
6.8	Questions and Exercises . . . . .	138
<b>7</b>	<b>Data Sources for Prediction: Databases, Hybrid Data and the Web . . . . .</b>	<b>141</b>
7.1	Ideal Models of Data . . . . .	141
7.1.1	Ideal Data for Prediction . . . . .	141
7.1.2	Ideal Data for Text and Unstructured Data . . . . .	142
7.1.3	Hybrid and Mixed Data . . . . .	142
7.2	Practical Data Sourcing . . . . .	144
7.3	Prototypical Examples . . . . .	145
7.3.1	Web-based Spreadsheet Data . . . . .	146
7.3.2	Web-based XML Data . . . . .	146
7.3.3	Opinion Data and Sentiment Analysis . . . . .	148
7.4	Hybrid Example: Independent Sources of Numerical and Text Data . . . . .	151
7.5	Mixed Data in Standard Table Format . . . . .	152
7.6	Summary . . . . .	153
7.7	Historical and Bibliographical Remarks . . . . .	154
7.8	Questions and Exercises . . . . .	154

<b>8</b>	<b>Case Studies</b> . . . . .	157
8.1	Market Intelligence from the Web . . . . .	157
8.1.1	The Problem . . . . .	157
8.1.2	Solution Overview . . . . .	158
8.1.3	Methods and Procedures . . . . .	159
8.1.4	System Deployment . . . . .	160
8.2	Lightweight Document Matching for Digital Libraries . . . . .	161
8.2.1	The Problem . . . . .	161
8.2.2	Solution Overview . . . . .	162
8.2.3	Methods and Procedures . . . . .	163
8.2.4	System Deployment . . . . .	164
8.3	Generating Model Cases for Help Desk Applications . . . . .	165
8.3.1	The Problem . . . . .	165
8.3.2	Solution Overview . . . . .	165
8.3.3	Methods and Procedures . . . . .	166
8.3.4	System Deployment . . . . .	168
8.4	Assigning Topics to News Articles . . . . .	169
8.4.1	The Problem . . . . .	169
8.4.2	Solution Overview . . . . .	169
8.4.3	Methods and Procedures . . . . .	169
8.4.4	System Deployment . . . . .	173
8.5	E-mail Filtering . . . . .	174
8.5.1	The Problem . . . . .	174
8.5.2	Solution Overview . . . . .	174
8.5.3	Methods and Procedures . . . . .	175
8.5.4	System Deployment . . . . .	177
8.6	Search Engines . . . . .	177
8.6.1	The Problem . . . . .	177
8.6.2	Solution Overview . . . . .	177
8.6.3	Methods and Procedures . . . . .	178
8.6.4	System Deployment . . . . .	179
8.7	Extracting Named Entities from Documents . . . . .	181
8.7.1	The Problem . . . . .	181
8.7.2	Solution Overview . . . . .	181
8.7.3	Methods and Procedures . . . . .	182
8.7.4	System Deployment . . . . .	184
8.8	Customized Newspapers . . . . .	184
8.8.1	The Problem . . . . .	184
8.8.2	Solution Overview . . . . .	185
8.8.3	Methods and Procedures . . . . .	186
8.8.4	System Deployment . . . . .	187
8.9	Summary . . . . .	187
8.10	Historical and Bibliographical Remarks . . . . .	188
8.11	Questions and Exercises . . . . .	188

**9 Emerging Directions . . . . . 189**

9.1 Summarization . . . . . 189

9.2 Active Learning . . . . . 192

9.3 Learning with Unlabeled Data . . . . . 193

9.4 Different Ways of Collecting Samples . . . . . 194

9.4.1 Ensembles and Voting Methods . . . . . 194

9.4.2 Online Learning . . . . . 196

9.4.3 Cost-Sensitive Learning . . . . . 197

9.4.4 Unbalanced Samples and Rare Events . . . . . 198

9.5 Distributed Text Mining . . . . . 198

9.6 Learning to Rank . . . . . 200

9.7 Question Answering . . . . . 201

9.8 Summary . . . . . 202

9.9 Historical and Bibliographical Remarks . . . . . 203

9.10 Questions and Exercises . . . . . 204

**A Software Notes . . . . . 207**

A.1 Summary of Software . . . . . 207

A.2 Requirements . . . . . 208

A.3 Download Instructions . . . . . 208

**References . . . . . 211**

**Author Index . . . . . 219**

**Subject Index . . . . . 223**