

# Contents

---

<b>Preface</b>	<b>v</b>
<b>1 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?	6
1.3 Document Classification	7
1.4 Information Retrieval	8
1.5 Clustering and Organizing Documents	9
1.6 Information Extraction	10
1.7 Prediction and Evaluation	11
1.8 The Next Chapters	12
1.9 Historical and Bibliographical Remarks	13
<b>2 From Textual Information to Numerical Vectors</b>	<b>15</b>
2.1 Collecting Documents	15
2.2 Document Standardization	18
2.3 Tokenization	20
2.4 Lemmatization	21
2.4.1 Inflectional Stemming	21
2.4.2 Stemming to a Root	23
2.5 Vector Generation for Prediction	25
2.5.1 Multiword Features	32
2.5.2 Labels for the Right Answers	34
2.5.3 Feature Selection by Attribute Ranking	35
2.6 Sentence Boundary Determination	36

2.7	Part-Of-Speech Tagging	37
2.8	Word Sense Disambiguation	39
2.9	Phrase Recognition	39
2.10	Named Entity Recognition	40
2.11	Parsing	40
2.12	Feature Generation	42
2.13	Historical and Bibliographical Remarks	44
<b>3</b>	<b>Using Text for Prediction</b>	<b>47</b>
3.1	Recognizing that Documents Fit a Pattern	49
3.2	How Many Documents Are Enough?	51
3.3	Document Classification	52
3.4	Learning to Predict from Text	54
3.4.1	Similarity and Nearest-Neighbor Methods	55
3.4.2	Document Similarity	56
3.4.3	Decision Rules	58
3.4.3.1	How to Find the Best Decision Rules	64
3.4.4	Scoring by Probabilities	66
3.4.5	Linear Scoring Methods	69
3.4.5.1	How to Find the Best Scoring Model	71
3.5	Evaluation of Performance	77
3.5.1	Estimating Current and Future Performance	77
3.5.2	Getting the Most from a Learning Method	80
3.6	Applications	81
3.7	Historical and Bibliographical Remarks	82
<b>4</b>	<b>Information Retrieval and Text Mining</b>	<b>85</b>
4.1	Is Information Retrieval a Form of Text Mining?	85
4.2	Key Word Search	87
4.3	Nearest-Neighbor Methods	88
4.4	Measuring Similarity	89
4.4.1	Shared Word Count	89
4.4.2	Word Count and Bonus	90
4.4.3	Cosine Similarity	91
4.5	Web-Based Document Search	92
4.5.1	Link Analysis	93
4.6	Document Matching	97
4.7	Inverted Lists	98
4.8	Evaluation of Performance	100
4.9	Historical and Bibliographical Remarks	101

---

<b>5</b>	<b>Finding Structure in a Document Collection</b>	<b>103</b>
5.1	Clustering Documents by Similarity	106
5.2	Similarity of Composite Documents	107
5.2.1	$k$ -Means Clustering	109
5.2.1.1	Centroid Classifier	113
5.2.2	Hierarchical Clustering	114
5.2.3	The EM Algorithm	117
5.3	What Do a Cluster's Labels Mean?	120
5.4	Applications	122
5.5	Evaluation of Performance	123
5.6	Historical and Bibliographical Remarks	126
<b>6</b>	<b>Looking for Information in Documents</b>	<b>129</b>
6.1	Goals of Information Extraction	129
6.2	Finding Patterns and Entities from Text	132
6.2.1	Entity Extraction as Sequential Tagging	132
6.2.2	Tag Prediction as Classification	133
6.2.3	The Maximum Entropy Method	135
6.2.4	Linguistic Features and Encoding	140
6.2.5	Sequential Probability Model	143
6.3	Coreference and Relationship Extraction	145
6.3.1	Coreference Resolution	145
6.3.2	Relationship Extraction	148
6.4	Template Filling and Database Construction	149
6.5	Applications	151
6.5.1	Information Retrieval	151
6.5.2	Commercial Extraction Systems	151
6.5.3	Criminal Justice	152
6.5.4	Intelligence	153
6.6	Historical and Bibliographical Remarks	154
<b>7</b>	<b>Case Studies</b>	<b>157</b>
7.1	Market Intelligence from the Web	157
7.2	Lightweight Document Matching for Digital Libraries	163
7.3	Generating Model Cases for Help Desk Applications	167
7.4	Assigning Topics to News Articles	172
7.5	E-mail Filtering	178
7.6	Search Engines	182
7.7	Extracting Named Entities from Documents	186
7.8	Customized Newspapers	191
7.9	Historical and Bibliographical Remarks	194

---

<b>8</b>	<b>Emerging Directions</b>	<b>197</b>
8.1	Summarization	198
8.2	Active Learning	201
8.3	Learning with Unlabeled Data	202
8.4	Different Ways of Collecting Samples	203
8.4.1	Multiple Samples and Voting Methods	204
8.4.2	Online Learning	205
8.4.3	Cost-Sensitive Learning	206
8.4.4	Unbalanced Samples and Rare Events	207
8.5	Question Answering	208
8.6	Historical and Bibliographical Remarks	210
	<b>Appendix: Software Notes</b>	<b>213</b>
A.1	Summary of Software	213
A.2	Requirements	214
A.3	Download Instructions	215
	<b>References</b>	<b>217</b>
	<b>Author Index</b>	<b>229</b>
	<b>Subject Index</b>	<b>233</b>

Text Mining

Predictive Methods for Analyzing Unstructured  
Information

Weiss, S.M.; Indurkha, N.; Zhang, T.; Damerau, F.

2005, XII, 237 p., Hardcover

ISBN: 978-0-387-95433-2