

Contents

1	Overview of Text Mining	1
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.	11
1.10	Historical and Bibliographical Remarks.	11
1.11	Questions and Exercises	12
2	From Textual Information to Numerical Vectors	13
2.1	Collecting Documents	13
2.2	Document Standardization	15
2.3	Tokenization	17
2.4	Lemmatization	19
2.4.1	Inflectional Stemming.	19
2.4.2	Stemming to a Root	21
2.5	Vector Generation for Prediction	21
2.5.1	Multiword Features	26
2.5.2	Labels for the Right Answers	29
2.5.3	Feature Selection by Attribute Ranking.	29
2.6	Sentence Boundary Determination	30
2.7	Part-of-Speech Tagging	30
2.8	Word Sense Disambiguation	32
2.9	Phrase Recognition	33
2.10	Named Entity Recognition.	33

2.11	Parsing	34
2.12	Feature Generation	35
2.13	Summary.	37
2.14	Historical and Bibliographical Remarks.	37
2.15	Questions and Exercises	39
3	Using Text for Prediction	41
3.1	Recognizing that Documents Fit a Pattern	43
3.2	How Many Documents Are Enough?	44
3.3	Document Classification	45
3.4	Learning to Predict from Text	46
3.4.1	Similarity and Nearest-Neighbor Methods	47
3.4.2	Document Similarity.	48
3.4.3	Decision Rules.	50
3.4.4	Decision Trees.	56
3.4.5	Scoring by Probabilities	57
3.4.6	Linear Scoring Methods	60
3.5	Evaluation of Performance.	69
3.5.1	Estimating Current and Future Performance.	69
3.5.2	Getting the Most from a Learning Method	71
3.5.3	Errors and Pitfalls in Big Data Evaluation.	72
3.6	Applications	74
3.7	Graph Models for Social Networks.	74
3.8	Summary.	76
3.9	Historical and Bibliographical Remarks.	77
3.10	Questions and Exercises	79
4	Information Retrieval and Text Mining.	81
4.1	Is Information Retrieval a Form of Text Mining?	81
4.2	Key Word Search.	82
4.3	Nearest-Neighbor Methods	83
4.4	Measuring Similarity.	84
4.4.1	Shared Word Count	84
4.4.2	Word Count and Bonus	85
4.4.3	Cosine Similarity	86
4.5	Web-Based Document Search	87
4.5.1	Link Analysis	88
4.6	Document Matching	91
4.7	Inverted Lists.	92
4.8	Evaluation of Performance.	93
4.9	Summary.	94
4.10	Historical and Bibliographical Remarks.	95
4.11	Questions and Exercises	95

- 5 Finding Structure in a Document Collection** 97
 - 5.1 Clustering Documents by Similarity 99
 - 5.2 Similarity of Composite Documents 100
 - 5.2.1 *k*-Means Clustering 102
 - 5.2.2 Hierarchical Clustering 106
 - 5.2.3 The EM Algorithm. 108
 - 5.3 What Do a Cluster’s Labels Mean? 112
 - 5.4 Applications 113
 - 5.5 Evaluation of Performance. 114
 - 5.6 Summary. 116
 - 5.7 Historical and Bibliographical Remarks. 116
 - 5.8 Questions and Exercises 118

- 6 Looking for Information in Documents** 119
 - 6.1 Goals of Information Extraction 119
 - 6.2 Finding Patterns and Entities from Text 121
 - 6.2.1 Entity Extraction as Sequential Tagging 122
 - 6.2.2 Tag Prediction as Classification 123
 - 6.2.3 The Maximum Entropy Method. 124
 - 6.2.4 Linguistic Features and Encoding. 129
 - 6.2.5 Local Sequence Prediction Models. 130
 - 6.2.6 Global Sequence Prediction Models 134
 - 6.3 Coreference and Relationship Extraction 135
 - 6.3.1 Coreference Resolution 135
 - 6.3.2 Relationship Extraction 138
 - 6.4 Template Filling and Database Construction 139
 - 6.5 Applications 140
 - 6.5.1 Information Retrieval 140
 - 6.5.2 Commercial Extraction Systems. 140
 - 6.5.3 Criminal Justice. 141
 - 6.5.4 Intelligence 142
 - 6.6 Summary. 143
 - 6.7 Historical and Bibliographical Remarks. 143
 - 6.8 Questions and Exercises 145

- 7 Data Sources for Prediction: Databases, Hybrid Data and the Web** 147
 - 7.1 Ideal Models of Data 147
 - 7.1.1 Ideal Data for Prediction. 147
 - 7.1.2 Ideal Data for Text and Unstructured Data 148
 - 7.1.3 Hybrid and Mixed Data 148
 - 7.2 Practical Data Sourcing. 150

- 7.3 Prototypical Examples. 151
 - 7.3.1 Web-Based Spreadsheet Data. 152
 - 7.3.2 Web-Based XML Data 152
 - 7.3.3 Opinion Data and Sentiment Analysis. 153
- 7.4 Hybrid Example: Independent Sources of Numerical and Text Data 158
- 7.5 Mixed Data in Standard Table Format. 159
- 7.6 Summary. 160
- 7.7 Historical and Bibliographical Remarks. 162
- 7.8 Questions and Exercises 162

- 8 Case Studies 165**
 - 8.1 Market Intelligence from the Web 165
 - 8.1.1 The Problem 165
 - 8.1.2 Solution Overview 166
 - 8.1.3 Methods and Procedures 167
 - 8.1.4 System Deployment 168
 - 8.2 Lightweight Document Matching for Digital Libraries. 170
 - 8.2.1 The Problem 170
 - 8.2.2 Solution Overview 170
 - 8.2.3 Methods and Procedures 171
 - 8.2.4 System Deployment 173
 - 8.3 Generating Model Cases for Help Desk Applications 173
 - 8.3.1 The Problem 173
 - 8.3.2 Solution Overview 174
 - 8.3.3 Methods and Procedures 174
 - 8.3.4 System Deployment 176
 - 8.4 Assigning Topics to News Articles. 177
 - 8.4.1 The Problem 177
 - 8.4.2 Solution Overview 177
 - 8.4.3 Methods and Procedures 178
 - 8.4.4 System Deployment 182
 - 8.5 E-mail Filtering 182
 - 8.5.1 The Problem 182
 - 8.5.2 Solution Overview 183
 - 8.5.3 Methods and Procedures 184
 - 8.5.4 System Deployment 185
 - 8.6 Search Engines 186
 - 8.6.1 The Problem 186
 - 8.6.2 Solution Overview 186
 - 8.6.3 Methods and Procedures 187
 - 8.6.4 System Deployment 188

8.7	Extracting Named Entities from Documents	190
8.7.1	The Problem	190
8.7.2	Solution Overview	190
8.7.3	Methods and Procedures	191
8.7.4	System Deployment	193
8.8	Mining Social Media	194
8.8.1	The Problem	194
8.8.2	Solution Overview	195
8.8.3	Methods and Procedures	196
8.8.4	System Deployment	197
8.9	Customized Newspapers	197
8.9.1	The Problem	197
8.9.2	Solution Overview	198
8.9.3	Methods and Procedures	198
8.9.4	System Deployment	199
8.10	Summary.	200
8.11	Historical and Bibliographical Remarks.	200
8.12	Questions and Exercises	201
9	Emerging Directions.	203
9.1	Summarization	203
9.2	Active Learning	206
9.3	Learning with Unlabeled Data	207
9.4	Different Ways of Collecting Samples.	208
9.4.1	Ensembles and Voting Methods.	208
9.4.2	Online Learning.	210
9.4.3	Deep Learning.	211
9.4.4	Cost-Sensitive Learning	214
9.4.5	Unbalanced Samples and Rare Events.	214
9.5	Distributed Text Mining	215
9.6	Learning to Rank	217
9.7	Question Answering	218
9.8	Summary.	219
9.9	Historical and Bibliographical Remarks.	219
9.10	Questions and Exercises	222
	References.	223
	Author Index	231
	Subject Index	235



<http://www.springer.com/978-1-4471-6749-5>

Fundamentals of Predictive Text Mining

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

2015, XIII, 239 p. 115 illus., Hardcover

ISBN: 978-1-4471-6749-5