

Contents

1	Introduction	1
1.1	About Text Mining and MATLAB®	2
1.2	About this Book	3
1.3	A (Very) Brief Introduction to MATLAB®	6
1.4	Further Reading	11
	References	12

Part I Fundamentals

2	Handling Textual Data	15
2.1	Characters and Character Arrays	15
2.2	Handling Text with Cell Arrays	18
2.3	Handling Text with Structures	21
2.4	Some Useful Functions	24
2.5	Further Reading	29
2.6	Proposed Exercises	30
	References	32
3	Regular Expressions	33
3.1	Basic Operators for Matching Characters	33
3.2	Matching Sequences of Characters	36
3.3	Conditional Matching	40
3.4	Working with Tokens	42
3.5	Further Reading	44
3.6	Proposed Exercises	44
	References	48
4	Basic Operations with Strings	49
4.1	Searching and Comparing	49
4.2	Replacement and Insertion	57

4.3	Segmentation and Concatenation	60
4.4	Set Operations	66
4.5	Further Reading	72
4.6	Proposed Exercises	72
	References	75
5	Reading and Writing Files	77
5.1	Basic File Formats	77
5.2	Other Useful Formats	87
5.3	Handling Files and Directories	101
5.4	Further Reading	106
5.5	Proposed Exercises	107
	References	110
 Part II Mathematical Models		
6	Basic Corpus Statistics	113
6.1	Fundamental Properties	113
6.2	Word Co-Occurrences	126
6.3	Accounting for Order	134
6.4	Further Reading	138
6.5	Proposed Exercises	140
6.6	Short Projects	142
	References	143
7	Statistical Models	145
7.1	Basic n -Gram Models	145
7.2	Discounting	148
7.3	Model Interpolation	157
7.4	Statistical Bag-of-Words	161
7.5	Further Reading	168
7.6	Proposed Exercises	169
7.7	Short Projects	171
	References	173
8	Geometrical Models	175
8.1	The Term-Document Matrix	175
8.2	The Vector Space Model	183
8.3	Association Scores and Distances	192
8.4	Further Reading	199
8.5	Proposed Exercises	200
8.6	Short Projects	202
	References	203

9	Dimensionality Reduction	205
9.1	Vocabulary Pruning and Merging	205
9.2	The Linear Transformation Approach	211
9.3	Non-linear Projection Methods	222
9.4	Further Reading	229
9.5	Proposed Exercises	230
9.6	Short Projects	232
	References	233

Part III Methods and Applications

10	Document Categorization	237
10.1	Data Collection Preparation	237
10.2	Unsupervised Clustering	242
10.3	Supervised Classification in Vector Space	252
10.4	Supervised Classification in Probability Space	260
10.5	Further Reading	269
10.6	Proposed Exercises	270
10.7	Short Projects	274
	References	276
11	Document Search	277
11.1	Binary Search	277
11.2	Vector-Based Search	289
11.3	Cross-Language Search	296
11.4	Further Reading	307
11.5	Proposed Exercises	308
11.6	Short Projects	310
	References	311
12	Content Analysis	313
12.1	Dimensions of Analysis	313
12.2	Polarity Estimation	319
12.3	Property Extraction	329
12.4	Further Reading	341
12.5	Proposed Exercises	342
12.6	Short Projects	345
	References	347
	Index	353



<http://www.springer.com/978-1-4614-4150-2>

Text Mining with MATLAB®

Banchs, R.E.

2013, XII, 356 p., Hardcover

ISBN: 978-1-4614-4150-2