

CONTENTS

PREFACE	5
I PRELIMINARIES	7
1 INTRODUCTION	9
1.1 Book Outline	11
1.2 Contributions	12
2 MACHINE LEARNING, DATA MINING, AND INFORMATION RETRIEVAL	14
2.1 Data Representation	17
2.1.1 Text Data	17
2.1.2 Time-Series Data	21
2.2 Distance and Similarity Measures	23
2.2.1 Minkowski Distances	24
2.2.2 Fractional Distances	24
2.2.3 Bray-Curtis and Normalized Euclidean Distance	24
2.2.4 Canberra Distance	25
2.2.5 Cosine Similarity	25
2.2.6 Jaccard Similarity	26
2.2.7 Dynamic Time Warping Distance	26
2.3 Classification	27
2.3.1 Algorithms	28
2.3.2 Using Binary Classifiers for Multi-Class Problems	36
2.3.3 Classifier Evaluation	38
2.3.4 Overfitting	44
2.3.5 Discussion	45
2.4 Semi-Supervised Learning	45
2.5 Clustering	47
2.5.1 Algorithms	47
2.5.2 Clustering Evaluation	52
2.5.3 Discussion	54
2.6 Outlier Detection	55

2.7	Information Retrieval	58
2.7.1	The Vector Space Model	59
2.7.2	Advanced Representations	59
2.7.3	Evaluation of IR Systems	60
2.8	Dimensionality Reduction	62
2.8.1	Feature Selection	63
2.8.2	Feature Extraction	67
2.9	Summary	72
 II METRICS		 73
3	THE CONCENTRATION PHENOMENON	75
3.1	Concentration of Distances	75
3.2	Concentration of Cosine Similarity	78
3.3	Proofs of Theorems 7 and 8	81
4	THE HUBNESS PHENOMENON	85
4.1	Related Work	86
4.2	Observing Hubness	87
4.3	Explaining Hubness	90
4.3.1	The Position of Hubs	90
4.3.2	Mechanisms Behind Hubness	90
4.4	Proof of Theorem 9	95
4.4.1	Distance Concentration Results	96
4.4.2	Distances in iid Normal Data	96
4.4.3	Asymptotic Equivalence	98
4.4.4	Expectation of the Noncentral Chi Distribution	99
4.4.5	Properties of the Generalized Laguerre Function	100
4.4.6	The Main Result	102
4.5	Discussion	104
4.5.1	Nearest-Neighbor Graph Structure	107
4.5.2	Rate of Convergence and the Role of Boundaries	109
5	HUBNESS AND MACHINE LEARNING	112
5.1	Related Work	112
5.2	Observing Hubness in Real Data	113
5.3	Explaining Hubness in Real Data	116
5.4	Hubs and Outliers	117
5.5	Hubness and Dimensionality Reduction	119
5.6	Impact of Hubness on Machine Learning	120
5.6.1	Supervised Learning	122

5.6.2	Semi-Supervised Learning	129
5.6.3	Unsupervised Learning	132
5.7	Summary and Future Work	136
6	HUBNESS AND TIME SERIES	138
6.1	Related Work	140
6.2	Observing Hubness in Time Series	141
6.3	Explaining Hubness in Time Series	141
6.4	Hubness and Dimensionality Reduction	144
6.5	Impact of Hubness on Time-Series Classification	146
6.5.1	“Good” and “Bad” k -Occurrences	146
6.5.2	A Framework for Categorizing Time-Series Data Sets	147
6.5.3	Weighting Scheme for k -NN Classification	150
6.6	Experimental Evaluation	151
6.6.1	The Experimental Setup	151
6.6.2	k -NN Classification Results	151
6.6.3	Other Distance Measures and Methods	153
6.7	Summary and Future Work	155
7	HUBNESS AND INFORMATION RETRIEVAL	156
7.1	Observing Hubness in Text Data	157
7.2	Explaining Hubness in Text Data	160
7.2.1	The Mechanism of Hub Formation	161
7.2.2	Hub Formation in Real Data	165
7.3	Hubness and Dimensionality Reduction	166
7.4	Impact of Hubness on Information Retrieval	167
7.4.1	Hubness and the Cluster Hypothesis	167
7.4.2	A Similarity Adjustment Scheme	168
7.4.3	Advanced Representations	171
7.5	Summary and Future Work	171
III	DOCUMENT REPRESENTATION AND FEATURE SELECTION	175
8	TERM WEIGHTING FOR TEXT CATEGORIZATION	177
8.1	Related Work	178
8.2	The Experimental Setup	178
8.2.1	Data Sets	179
8.2.2	Document Representations	180
8.2.3	Classifiers	180

8.3	Results	181
8.3.1	Effects of Stemming	183
8.3.2	Effects of Normalization	184
8.3.3	Effects of the logtf Transformation	186
8.3.4	Effects of the idf Transformation	187
8.3.5	Robustness	189
8.3.6	Training and Classification Speed	190
8.4	Summary and Future Work	191
9	TERM WEIGHTING AND FEATURE SELECTION	193
9.1	The Experimental Setup	194
9.1.1	Data Sets	195
9.1.2	Document Representations	195
9.1.3	Feature Selection	196
9.1.4	Classifiers	196
9.2	Results	196
9.2.1	Rankings of Feature-Selection Methods and Re- duction Rates	196
9.2.2	Interaction Between Bag-of-Words Transfor- mations and Feature Selection	198
9.3	Summary and Future Work	204
9.4	A Note on Hubness, Feature Selection, and Generation	206
10	CONCLUSION	209
A	TERM WEIGHTING IN THE BOW REPRESENTATION	213
A.1	Term Weighting Without Stemming	213
A.2	Term Weighting With Stemming	214
	BIBLIOGRAPHY	217
	ABOUT THE AUTHOR	235
	O AUTORU	236
	SAŽETAK	237