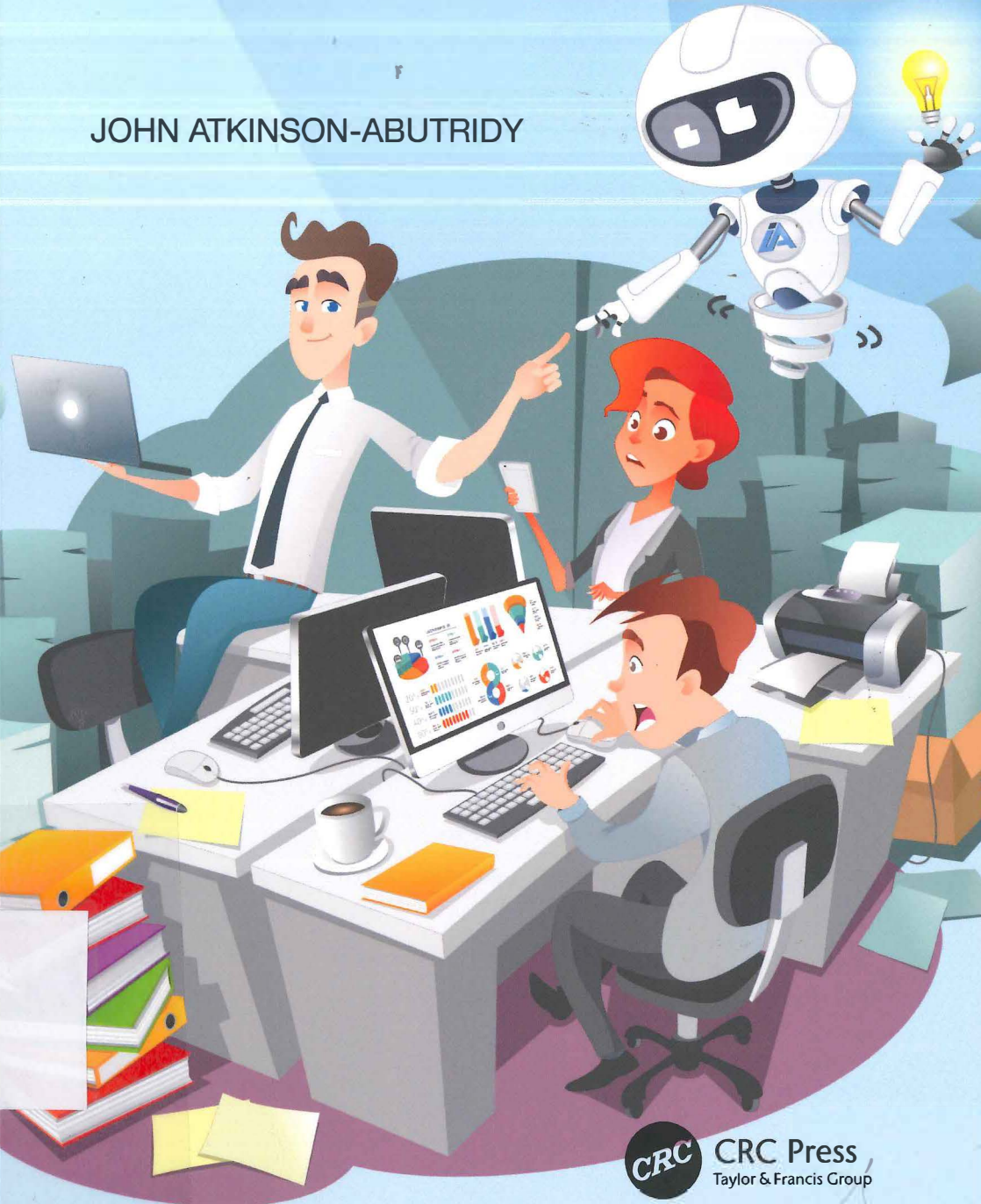


TEXT ANALYTICS

An Introduction to the Science and Applications
of Unstructured Information Analysis

JOHN ATKINSON-ABUTRIDY



CRC Press
Taylor & Francis Group

A CHAPMAN & HALL BOOK

๕/๖๖
1,๖๐๗.

๖/๖๖๘๘๘/๑
๑๑๕๙๑๖๙
๑๑๖๐๙๘๕

Text Analytics

An Introduction to the Science and Applications of Unstructured Information Analysis



John Atkinson-Abutridy



CRC Press is an imprint of the
Taylor & Francis Group, an informa business
A CHAPMAN & HALL BOOK

๑๑๖๐๙๘๕

Contents

List of Figures, xi

List of Tables, xv

Preface, xvii

Acknowledgments, xxv

Author, xxvii

CHAPTER 1 ■ Text Analytics	1
1.1 INTRODUCTION	1
1.2 TEXT MINING AND TEXT ANALYTICS	4
1.3 TASKS AND APPLICATIONS	7
1.4 THE TEXT ANALYTICS PROCESS	10
1.5 SUMMARY	12
1.6 QUESTIONS	13
CHAPTER 2 ■ Natural-Language Processing	15
2.1 INTRODUCTION	15
2.2 THE SCOPE OF NATURAL-LANGUAGE PROCESSING	16
2.3 NLP LEVELS AND TASKS	18
2.3.1 Phonology	20
2.3.2 Morphology	20
2.3.3 Lexicon	22
2.3.4 Syntax	29

2.3.5	Semantics	33
2.3.6	Reasoning and Pragmatics	38
2.4	SUMMARY	39
2.5	EXERCISES	39
2.5.1	Morphological Analysis	39
2.5.2	Lexical Analysis	44
2.5.3	Syntactic Analysis	45
CHAPTER 3 ■ Information Extraction		49
3.1	INTRODUCTION	49
3.2	RULE-BASED INFORMATION EXTRACTION	53
3.3	NAMED-ENTITY RECOGNITION	54
3.3.1	N-Gram Models	57
3.4	RELATION EXTRACTION	60
3.5	EVALUATION	64
3.6	SUMMARY	67
3.7	EXERCISES	67
3.7.1	Regular Expressions	68
3.7.2	Named-Entity Recognition	72
CHAPTER 4 ■ Document Representation		75
4.1	INTRODUCTION	75
4.2	DOCUMENT INDEXING	77
4.3	VECTOR SPACE MODELS	79
4.3.1	Boolean Representation Model	79
4.3.2	Term Frequency Model	80
4.3.3	Inverse Document Frequency Model	82
4.4	SUMMARY	84
4.5	EXERCISES	84
4.5.1	TFxIDF Representation Model	84

CHAPTER 5 ■ Association Rules Mining	91
5.1 INTRODUCTION	91
5.2 ASSOCIATION PATTERNS	92
5.3 EVALUATION	94
5.3.1 Support	94
5.3.2 Confidence	95
5.3.3 Lift	95
5.4 ASSOCIATION RULES GENERATION	96
5.5 SUMMARY	101
5.6 EXERCISES	101
5.6.1 Extraction of Association Rules	101
CHAPTER 6 ■ Corpus-Based Semantic Analysis	105
6.1 INTRODUCTION	105
6.2 CORPUS-BASED SEMANTIC ANALYSIS	107
6.3 LATENT SEMANTIC ANALYSIS	109
6.3.1 Creating Vectors with LSA	110
6.4 WORD2VEC	115
6.4.1 Embedding Learning	118
6.4.2 Prediction and Embeddings Interpretation	121
6.5 SUMMARY	123
6.6 EXERCISES	123
6.6.1 Latent Semantic Analysis	124
6.6.2 Word Embedding with Word2Vec	130
CHAPTER 7 ■ Document Clustering	137
7.1 INTRODUCTION	137
7.2 DOCUMENT CLUSTERING	139
7.3 K-MEANS CLUSTERING	145
7.4 SELF-ORGANIZING MAPS	149
7.4.1 Topological Maps Learning	150
7.5 SUMMARY	155
7.6 EXERCISES	155

7.6.1	K-means Clustering	155
7.6.2	Self-organizing Maps	162
CHAPTER 8 ■ Topic Modeling		165
8.1	INTRODUCTION	165
8.2	TOPIC MODELING	166
8.3	LATENT DIRICHLET ALLOCATION	169
8.4	EVALUATION	176
8.5	SUMMARY	179
8.6	EXERCISES	179
8.6.1	Modeling Topics with LDA	179
CHAPTER 9 ■ Document Categorization		185
9.1	INTRODUCTION	185
9.2	CATEGORIZATION MODELS	187
9.3	BAYESIAN TEXT CATEGORIZATION	191
9.3.1	Conditional Class Probability	192
9.3.2	<i>A Priori</i> Probability	193
9.3.3	Evidence	194
9.3.4	Classification	194
9.4	MAXIMUM ENTROPY CATEGORIZATION	195
9.5	EVALUATION	200
9.6	SUMMARY	203
9.7	EXERCISES	203
9.7.1	Naïve Bayes Categorization	203
9.7.2	MaxEnt Categorization	208
CONCLUDING REMARKS, 215		
BIBLIOGRAPHY, 221		
GLOSSARY, 225		
INDEX, 229		