

Contents

1	Introduction	1
1.1	Information Search in Times of Big Data	1
1.1.1	Text Mining to the Rescue	2
1.2	A Need for Efficient and Robust Text Analysis Pipelines	4
1.2.1	Basic Text Analysis Scenario	5
1.2.2	Shortcomings of Traditional Text Analysis Pipelines	6
1.2.3	Problems Approached in This Book	7
1.3	Towards Intelligent Pipeline Design and Execution	8
1.3.1	Central Research Question and Method	8
1.3.2	An Artificial Intelligence Approach	9
1.4	Contributions and Outline of This Book	12
1.4.1	New Findings in Ad-Hoc Large-Scale Text Mining	13
1.4.2	Contributions to the Concerned Research Fields	14
1.4.3	Structure of the Remaining Chapters	15
1.4.4	Published Research Within This Book	16
2	Text Analysis Pipelines	19
2.1	Foundations of Text Mining	20
2.1.1	Text Mining	20
2.1.2	Information Retrieval	21
2.1.3	Natural Language Processing	22
2.1.4	Data Mining	25
2.1.5	Development and Evaluation	30
2.2	Text Analysis Tasks, Processes, and Pipelines	34
2.2.1	Text Analysis Tasks	34
2.2.2	Text Analysis Processes	36
2.2.3	Text Analysis Pipelines	37
2.3	Case Studies in This Book	40
2.3.1	InfexBA – Information Extraction for Business Applications	40
2.3.2	ArguAna – Argumentation Analysis in Customer Opinions	42
2.3.3	Other Evaluated Text Analysis Tasks	43

2.4	State of the Art in Ad-Hoc Large-Scale Text Mining	44
2.4.1	Text Analysis Approaches	44
2.4.2	Design of Text Analysis Approaches	45
2.4.3	Efficiency of Text Analysis Approaches.	47
2.4.4	Robustness of Text Analysis Approaches	50
3	Pipeline Design	55
3.1	Ideal Construction and Execution for Ad-Hoc Text Mining	56
3.1.1	The Optimality of Text Analysis Pipelines	56
3.1.2	Paradigms of Designing Optimal Text Analysis Pipelines.	60
3.1.3	Case Study of Ideal Construction and Execution	64
3.1.4	Discussion of Ideal Construction and Execution	68
3.2	A Process-Oriented View of Text Analysis	68
3.2.1	Text Analysis as an Annotation Task.	69
3.2.2	Modeling the Information to Be Annotated.	70
3.2.3	Modeling the Quality to Be Achieved by the Annotation	71
3.2.4	Modeling the Analysis to Be Performed for Annotation	72
3.2.5	Defining an Annotation Task Ontology	74
3.2.6	Discussion of the Process-Oriented View	75
3.3	Ad-Hoc Construction via Partial Order Planning	76
3.3.1	Modeling Algorithm Selection as a Planning Problem	77
3.3.2	Selecting the Algorithms of a Partially Ordered Pipeline	78
3.3.3	Linearizing the Partially Ordered Pipeline.	80
3.3.4	Properties of the Proposed Approach	82
3.3.5	An Expert System for Ad-Hoc Construction	86
3.3.6	Evaluation of Ad-Hoc Construction	88
3.3.7	Discussion of Ad-Hoc Construction.	91
3.4	An Information-Oriented View of Text Analysis.	92
3.4.1	Text Analysis as a Filtering Task	92
3.4.2	Defining the Relevance of Portions of Text	95
3.4.3	Specifying a Degree of Filtering for Each Relation Type	97
3.4.4	Modeling Dependencies of the Relevant Information Types	98
3.4.5	Discussion of the Information-Oriented View	100
3.5	Optimal Execution via Truth Maintenance	101
3.5.1	Modeling Input Control as a Truth Maintenance Problem.	101
3.5.2	Filtering the Relevant Portions of Text.	104
3.5.3	Determining the Relevant Portions of Text	106
3.5.4	Properties of the Proposed Approach	107
3.5.5	A Software Framework for Optimal Execution	109
3.5.6	Evaluation of Optimal Execution.	111
3.5.7	Discussion of Optimal Execution.	116

3.6	Trading Efficiency for Effectiveness in Ad-Hoc Text Mining	117
3.6.1	Integration with Passage Retrieval	117
3.6.2	Integration with Text Filtering	118
3.6.3	Implications for Pipeline Efficiency	120
4	Pipeline Efficiency	123
4.1	Ideal Scheduling for Large-Scale Text Mining	124
4.1.1	The Efficiency Potential of Pipeline Scheduling	124
4.1.2	Computing Optimal Schedules with Dynamic Programming	126
4.1.3	Properties of the Proposed Solution	129
4.1.4	Case Study of Ideal Scheduling.	131
4.1.5	Discussion of Ideal Scheduling	134
4.2	The Impact of Relevant Information in Input Texts	134
4.2.1	Formal Specification of the Impact	135
4.2.2	Experimental Analysis of the Impact	136
4.2.3	Practical Relevance of the Impact	138
4.2.4	Implications of the Impact	140
4.3	Optimized Scheduling via Informed Search	141
4.3.1	Modeling Pipeline Scheduling as a Search Problem.	142
4.3.2	Scheduling Text Analysis Algorithms with k-best A* Search	144
4.3.3	Properties of the Proposed Approach	147
4.3.4	Evaluation of Optimized Scheduling	149
4.3.5	Discussion of Optimized Scheduling	155
4.4	The Impact of the Heterogeneity of Input Texts	156
4.4.1	Experimental Analysis of the Impact	156
4.4.2	Quantification of the Impact	159
4.4.3	Practical Relevance of the Impact	161
4.4.4	Implications of the Impact	163
4.5	Adaptive Scheduling via Self-supervised Online Learning	163
4.5.1	Modeling Pipeline Scheduling as a Classification Problem . . .	164
4.5.2	Learning to Predict Run-Times Self-supervised and Online. . .	165
4.5.3	Adapting a Pipeline's Schedule to the Input Text	166
4.5.4	Properties of the Proposed Approach	167
4.5.5	Evaluation of Adaptive Scheduling	169
4.5.6	Discussion of Adaptive Scheduling	175
4.6	Parallelizing Execution in Large-Scale Text Mining	177
4.6.1	Effects of Parallelizing Pipeline Execution	177
4.6.2	Parallelization of Text Analyses	179
4.6.3	Parallelization of Text Analysis Pipelines	180
4.6.4	Implications for Pipeline Robustness	182

5	Pipeline Robustness	183
5.1	Ideal Domain Independence for High-Quality Text Mining	184
5.1.1	The Domain Dependence Problem in Text Analysis	184
5.1.2	Requirements of Achieving Pipeline Domain Independence	186
5.1.3	Domain-Independent Features of Argumentative Texts	190
5.2	A Structure-Oriented View of Text Analysis	191
5.2.1	Text Analysis as a Structure Classification Task	191
5.2.2	Modeling the Argumentation and Content of a Text	192
5.2.3	Modeling the Argumentation Structure of a Text	193
5.2.4	Defining a Structure Classification Task Ontology	195
5.2.5	Discussion of the Structure-Oriented View	197
5.3	The Impact of the Overall Structure of Input Texts	198
5.3.1	Experimental Analysis of Content and Style Features	198
5.3.2	Statistical Analysis of the Impact of Task-Specific Structure	201
5.3.3	Statistical Analysis of the Impact of General Structure	204
5.3.4	Implications of the Invariance and Impact	206
5.4	Features for Domain Independence via Supervised Clustering	208
5.4.1	Approaching Classification as a Relatedness Problem	208
5.4.2	Learning Overall Structures with Supervised Clustering	209
5.4.3	Using the Overall Structures as Features for Classification	212
5.4.4	Properties of the Proposed Features	214
5.4.5	Evaluation of Features for Domain Independence	216
5.4.6	Discussion of Features for Domain Independence	221
5.5	Explaining Results in High-Quality Text Mining	223
5.5.1	Intelligible Text Analysis through Explanations	224
5.5.2	Explanation of Arbitrary Text Analysis Processes	224
5.5.3	Explanation of the Class of an Argumentative Text	227
5.5.4	Implications for Ad-Hoc Large-Scale Text Mining	229
6	Conclusion	231
6.1	Contributions and Open Problems	232
6.1.1	Enabling Ad-Hoc Text Analysis	232
6.1.2	Optimally Analyzing Text	233
6.1.3	Optimizing Analysis Efficiency	233
6.1.4	Robustly Classifying Text	234
6.2	Implications and Outlook	235
6.2.1	Towards Ad-Hoc Large-Scale Text Mining	235
6.2.2	Outside the Box	237
	Appendix A Text Analysis Algorithms	239
A.1	Analyses and Algorithms	239
A.1.1	Classification of Text	240
A.1.2	Entity Recognition	242
A.1.3	Normalization and Resolution	242

A.1.4 Parsing	243
A.1.5 Relation Extraction and Event Detection	244
A.1.6 Segmentation	246
A.1.7 Tagging	247
A.2 Evaluation Results	248
A.2.1 Efficiency Results	248
A.2.2 Effectiveness Results	249
Appendix B Software.	251
B.1 An Expert System for Ad-hoc Pipeline Construction	251
B.1.1 Getting Started.	252
B.1.2 Using the Expert System.	253
B.1.3 Exploring the Source Code of the System	254
B.2 A Software Framework for Optimal Pipeline Execution	256
B.2.1 Getting Started.	256
B.2.2 Using the Framework	256
B.2.3 Exploring the Source Code of the Framework	257
B.3 A Web Application for Sentiment Scoring and Explanation	258
B.3.1 Getting Started.	258
B.3.2 Using the Application.	259
B.3.3 Exploring the Source Code of the Application	260
B.3.4 Acknowledgments	261
B.4 Source Code of All Experiments and Case Studies	261
B.4.1 Software	261
B.4.2 Text Corpora	262
B.4.3 Experiments and Case Studies	262
Appendix C Text Corpora.	265
C.1 The Revenue Corpus	265
C.1.1 Compilation.	265
C.1.2 Annotation	267
C.1.3 Files	269
C.1.4 Acknowledgments	269
C.2 The ArguAna TripAdvisor Corpus	269
C.2.1 Compilation.	269
C.2.2 Annotation	271
C.2.3 Files	274
C.2.4 Acknowledgments	274
C.3 The LFA-11 Corpus	274
C.3.1 Compilation.	275
C.3.2 Annotation	275
C.3.3 Files	278
C.3.4 Acknowledgments	278

C.4 Used Existing Text corpora	278
C.4.1 CoNLL-2003 Dataset (English and German)	278
C.4.2 Sentiment Scale Dataset (and Related Datasets)	279
C.4.3 Brown Corpus	279
C.4.4 Wikipedia Sample	280
References.	281
Index	293

Text Analysis Pipelines

Towards Ad-hoc Large-Scale Text Mining

Wachsmuth, H.

2015, XX, 302 p. 74 illus. in color., Softcover

ISBN: 978-3-319-25740-2