

# • Brief Contents •

---

Preface		xii
Acknowledgments		xv
About the Authors		xvi
<b>PART I</b>	<b>• DIGITAL TEXTS, DIGITAL SOCIAL SCIENCE</b>	<b>1</b>
Chapter 1	• Social Science and the Digital Text Revolution	2
Chapter 2	• Research Design Strategies	16
<b>PART II</b>	<b>• TEXT MINING FUNDAMENTALS</b>	<b>33</b>
Chapter 3	• Web Crawling and Scraping	34
Chapter 4	• Lexical Resources	42
Chapter 5	• Basic Text Processing	52
Chapter 6	• Supervised Learning	62
<b>PART III</b>	<b>• TEXT ANALYSIS METHODS FROM THE HUMANITIES AND SOCIAL SCIENCES</b>	<b>73</b>
Chapter 7	• Thematic Analysis, Qualitative Data Analysis Software, and Visualization	74
Chapter 8	• Narrative Analysis	88
Chapter 9	• Metaphor Analysis	96
<b>PART IV</b>	<b>• TEXT MINING METHODS FROM COMPUTER SCIENCE</b>	<b>105</b>
Chapter 10	• Word and Text Relatedness	106
Chapter 11	• Text Classification	116
Chapter 12	• Information Extraction	130
Chapter 13	• Information Retrieval	136

Chapter 14	• Sentiment Analysis	148
Chapter 15	• Topic Models	156
<b>PART V</b>	<b>• CONCLUSIONS</b>	<b>163</b>
Chapter 16	• Text Mining, Text Analysis, and the Future of Social Science	164
References		168
Index		183

# • Detailed Contents •

---

Preface	xii
Acknowledgments	xv
About the Authors	xvi

**PART I • DIGITAL TEXTS, DIGITAL SOCIAL SCIENCE**

**1**

**1. Social Science and the Digital Text Revolution**

**2**

History of Text Analysis	3
Risks and Rewards of Text Mining for the Social Sciences	5
Social Data From Digital Environments	6
Theory and Metatheory	10
Ethics of Text Mining	12
<i>Participant Consent, Privacy, and Anonymity</i>	12
<i>Prompted and Unprompted Data</i>	13
Organization of This Volume	13

**2. Research Design Strategies**

**16**

Levels of Analysis	18
<i>The Textual Level</i>	18
<i>The Contextual Level</i>	18
<i>The Sociological Level</i>	18
Strategies for Document Selection and Sampling	19
<i>Case Selection</i>	19
<i>Text Sampling</i>	20
Types of Inferential Logic	22
<i>Inductive Logic</i>	23
<i>Deductive Logic</i>	24
<i>Abductive Logic</i>	25
Approaches to Research Design	27
<i>Analysis of Discourse Positions</i>	27
<i>Conversation Analysis</i>	28
<i>Critical Discourse Analysis</i>	28

*Content Analysis* 29  
*Foucauldian Intertextuality* 30  
*Analysis of Texts as Social Information* 31

**PART II • TEXT MINING FUNDAMENTALS**

**33**

**3. Web Crawling and Scraping 34**

Web Statistics 36  
Web Crawling 37  
    *Process Steps in Crawling* 37  
    *Traversal Strategies* 38  
    *Crawler Politeness* 38  
Web Scraping 39  
Software for Web Crawling and Scraping 41

**4. Lexical Resources 42**

WordNet 43  
    *WordNet-Affect* 45  
*Roget's Thesaurus* 46  
Linguistic Inquiry and Word Count 46  
General Inquirer 48  
Wikipedia 48  
    *Wiktionary* 51  
Downloadable Lexical Resources  
and Application Program Interfaces 51

**5. Basic Text Processing 52**

Tokenization 54  
Stop Word Removal 55  
Stemming and Lemmatization 55  
Text Statistics 56  
Language Models 59  
Other Text Processing 60  
    *Part of Speech Tagging* 60  
    *Collocation Identification* 60  
    *Syntactic Parsing* 61  
    *Named Entity Tagging* 61  
    *Word Sense Disambiguation* 61  
Software for Text Processing 61

**6. Supervised Learning 62**

- Feature Representation and Weighting 65
  - Feature Weighting* 65
- Supervised Learning Algorithms 66
  - Decision Trees* 67
  - Instance-Based Learning* 68
  - Support Vector Machines* 69
- Evaluation of Supervised Learning 71
- Software for Supervised Learning 71

**PART III • TEXT ANALYSIS METHODS FROM  
THE HUMANITIES AND SOCIAL SCIENCES**

**73**

**7. Thematic Analysis, Qualitative  
Data Analysis Software, and Visualization 74**

- Thematic Analysis 75
- Qualitative Data Analysis Software 77
- Visualization Tools 83
  - Word Clouds* 84
  - Word Trees and Phrase Nets* 84
  - Matrices and Maps* 85
  - Key Word in Context* 86
- Software for Thematic Analysis, Qualitative Data Analysis,  
and Visualization 86

**8. Narrative Analysis 88**

- Conceptual Foundations 90
  - Structural Approaches to Narrative* 90
  - Functionalist Approaches to Narrative* 91
  - Sociological Approaches to Narrative* 92
- Mixed Methods of Narrative Analysis 92
- Automated Methods of Narrative Analysis 93
- Future Directions 93
- Software for Narrative Analysis 94

**9. Metaphor Analysis 96**

- Theoretical Foundations 98
- Qualitative Metaphor Analysis 99

<i>Anthropology</i>	99
<i>Educational Research</i>	99
<i>Political Science</i>	100
<i>Psychology</i>	100
<i>Sociology</i>	101
Mixed Methods of Metaphor Analysis	101
<i>Management Research</i>	101
<i>Psychology</i>	102
<i>Sociology</i>	102
Automated Metaphor Identification Methods	103
Software for Metaphor Analysis	103

## **PART IV • TEXT MINING METHODS FROM COMPUTER SCIENCE** **105**

### **10. Word and Text Relatedness 106**

Theoretical Foundations	107
Corpus-Based and Knowledge-Based Measures of Relatedness	108
<i>Corpus-Based Measures of Word Relatedness</i>	108
<i>Knowledge-Based Measures of Word Relatedness</i>	110
<i>Measures of Text Relatedness</i>	112
Software and Data Sets for Word and Text Relatedness	114

### **11. Text Classification 116**

A Brief History of Text Classification	118
Applications of Text Classification	119
<i>Topic Classification</i>	119
<i>E-Mail Spam Detection</i>	120
<i>Sentiment Analysis/Opinion Mining</i>	120
<i>Gender Classification</i>	120
<i>Deception Detection</i>	122
<i>Other Applications</i>	122
Representing Texts for Supervised Text Classification	122
<i>Feature Weighting and Selection</i>	123
Text Classification Algorithms	124
<i>Naive Bayes</i>	124
<i>Rocchio Classifier</i>	125
Bootstrapping in Text Classification	126
Evaluation of Text Classification	127
Software and Data Sets for Text Classification	127

<b>12. Information Extraction</b>	<b>130</b>
Entity Extraction	132
Relation Extraction	133
Web Information Extraction	134
Template Filling	135
Software and Data Sets for Information Extraction and Text Mining	135
<b>13. Information Retrieval</b>	<b>136</b>
Theoretical Foundations	138
Components of an Information Retrieval System	138
Information Retrieval Models	140
The Vector Space Model	142
Evaluation of Information Retrieval Models	144
Web-Based Information Retrieval	145
Software and Data Sets for Information Retrieval	147
<b>14. Sentiment Analysis</b>	<b>148</b>
Theoretical Foundations	150
Lexicons	151
Corpora	152
Tools	153
Software and Data Sets for Sentiment Analysis	154
<b>15. Topic Models</b>	<b>156</b>
Digital Humanities	160
Political Science	160
Sociology	161
Software for Topic Modeling	161
<b>PART V • CONCLUSIONS</b>	<b>163</b>
<b>16. Text Mining, Text Analysis,     and the Future of Social Science</b>	<b>164</b>
Social and Computer Science Collaboration	166
References	168
Index	183