Contents

Preface xv

Acknowledgments xviii

PART I OVERVIEW AND BACKGROUND 1

Chapter 1 Introduction 3

1.1 Functions of Text Information Systems 7
1.2 Conceptual Framework for Text Information Systems 10
1.3 Organization of the Book 13
1.4 How to Use this Book 15
Bibliographic Notes and Further Reading 18

Chapter 2 Background 21

2.1 Basics of Probability and Statistics 21
2.2 Information Theory 31
2.3 Machine Learning 34
Bibliographic Notes and Further Reading 36
Exercises 37

Chapter 3 Text Data Understanding 39

3.1 History and State of the Art in NLP 42
3.2 NLP and Text Information Systems 43
3.3 Text Representation 46
3.4 Statistical Language Models 50
Bibliographic Notes and Further Reading 54
Exercises 55

Chapter 4 META: A Unified Toolkit for Text Data Management and Analysis 57
4.1 Design Philosophy 58
4.2 Setting up META 59
4.3 Architecture 60
4.4 Tokenization with META 61
4.5 Related Toolkits 64
Exercises 65

PART II TEXT DATA ACCESS 71

Chapter 5 Overview of Text Data Access 73
5.1 Access Mode: Pull vs. Push 73
5.2 Multimode Interactive Access 76
5.3 Text Retrieval 78
5.4 Text Retrieval vs. Database Retrieval 80
5.5 Document Selection vs. Document Ranking 82
Bibliographic Notes and Further Reading 84
Exercises 85

Chapter 6 Retrieval Models 87
6.1 Overview 87
6.2 Common Form of a Retrieval Function 88
6.3 Vector Space Retrieval Models 90
6.4 Probabilistic Retrieval Models 110
Bibliographic Notes and Further Reading 128
Exercises 129

Chapter 7 Feedback 133
7.1 Feedback in the Vector Space Model 135
7.2 Feedback in Language Models 138
Bibliographic Notes and Further Reading 144
Exercises 144

Chapter 8 Search Engine Implementation 147
8.1 Tokenizer 148
8.2 Indexer 150
8.3 Scorer 153

8.4 Feedback Implementation 157
8.5 Compression 158
8.6 Caching 162
Bibliographic Notes and Further Reading 165
Exercises 165

Chapter 9 Search Engine Evaluation 167
9.1 Introduction 167
9.2 Evaluation of Set Retrieval 170
9.3 Evaluation of a Ranked List 174
9.4 Evaluation with Multi-level Judgements 180
9.5 Practical Issues in Evaluation 183
Bibliographic Notes and Further Reading 187
Exercises 188

Chapter 10 Web Search 191
10.1 Web Crawling 192
10.2 Web Indexing 194
10.3 Link Analysis 200
10.4 Learning to Rank 208
10.5 The Future of Web Search 212
Bibliographic Notes and Further Reading 216
Exercises 216

Chapter 11 Recommender Systems 221
11.1 Content-based Recommendation 222
11.2 Collaborative Filtering 229
11.3 Evaluation of Recommender Systems 233
Bibliographic Notes and Further Reading 235
Exercises 235

PART III TEXT DATA ANALYSIS 239

Chapter 12 Overview of Text Data Analysis 241
12.1 Motivation: Applications of Text Data Analysis 242
12.2 Text vs. Non-text Data: Humans as Subjective Sensors 244
12.3 Landscape of text mining tasks 246

Chapter 13 Word Association Mining 251
13.1 General idea of word association mining 252
13.2 Discovery of paradigmatic relations 255
13.3 Discovery of Syntagmatic Relations 260
13.4 Evaluation of Word Association Mining 271
Bibliographic Notes and Further Reading 273
Exercises 273

Chapter 14 Text Clustering 275
14.1 Overview of Clustering Techniques 277
14.2 Document Clustering 279
14.3 Term Clustering 284
14.4 Evaluation of Text Clustering 294
Bibliographic Notes and Further Reading 296
Exercises 296

Chapter 15 Text Categorization 299
15.1 Introduction 299
15.2 Overview of Text Categorization Methods 300
15.3 Text Categorization Problem 302
15.4 Features for Text Categorization 304
15.5 Classification Algorithms 307
15.6 Evaluation of Text Categorization 313
Bibliographic Notes and Further Reading 315
Exercises 315

Chapter 16 Text Summarization 317
16.1 Overview of Text Summarization Techniques 318
16.2 Extractive Text Summarization 319
16.3 Abstractive Text Summarization 321
16.4 Evaluation of Text Summarization 324
16.5 Applications of Text Summarization 325
Bibliographic Notes and Further Reading 327
Exercises 327

Chapter 17 Topic Analysis 329
17.1 Topics as Terms 332
17.2 Topics as Word Distributions 335

17.3 Mining One Topic from Text 340
17.4 Probabilistic Latent Semantic Analysis 368
17.5 Extension of PLSA and Latent Dirichlet Allocation 377
17.6 Evaluating Topic Analysis 383
17.7 Summary of Topic Models 384
Bibliographic Notes and Further Reading 385
Exercises 386

Chapter 18 Opinion Mining and Sentiment Analysis 389
18.1 Sentiment Classification 393
18.2 Ordinal Regression 396
18.3 Latent Aspect Rating Analysis 400
18.4 Evaluation of Opinion Mining and Sentiment Analysis 409
Bibliographic Notes and Further Reading 410
Exercises 410

Chapter 19 Joint Analysis of Text and Structured Data 413
19.1 Introduction 413
19.2 Contextual Text Mining 417
19.3 Contextual Probabilistic Latent Semantic Analysis 419
19.4 Topic Analysis with Social Networks as Context 428
19.5 Topic Analysis with Time Series Context 433
19.6 Summary 439
Bibliographic Notes and Further Reading 440
Exercises 440

PART IV UNIFIED TEXT DATA MANAGEMENT ANALYSIS SYSTEM 443

Chapter 20 Toward A Unified System for Text Management and Analysis 445
20.1 Text Analysis Operators 448
20.2 System Architecture 452
20.3 META as a Unified System 453

Appendix A Bayesian Statistics 457
A.1 Binomial Estimation and the Beta Distribution 457
A.2 Pseudo Counts, Smoothing, and Setting Hyperparameters 459
A.3 Generalizing to a Multinomial Distribution 460

A.4 The Dirichlet Distribution 461
A.5 Bayesian Estimate of Multinomial Parameters 46
A.6 Conclusion 464

Appendix B Expectation- Maximization 465

B.1 A Simple Mixture Unigram Language Model 466
B.2 Maximum Likelihood Estimation 466
B.3 Incomplete vs. Complete Data 467
B.4 A Lower Bound of Likelihood 468
B.5 The General Procedure of EM 469

Appendix C KL-divergence and Dirichlet Prior Smoothing 473

C.1 Using KL-divergence for Retrieval 473
C.2 Using Dirichlet Prior Smoothing 475
C.3 Computing the Query Model $p(w \mid \widehat{\theta}_Q)$ 475

References 477

Index 489

Authors' Biographies 509