

Chengqing Zong • Rui Xia • Jiajun Zhang

Text Data Mining

 Springer

Contents

1	Introduction	1
1.1	The Basic Concepts	1
1.2	Main Tasks of Text Data Mining	3
1.3	Existing Challenges in Text Data Mining	6
1.4	Overview and Organization of This Book	9
1.5	Further Reading	12
2	Data Annotation and Preprocessing	15
2.1	Data Acquisition.....	15
2.2	Data Preprocessing	20
2.3	Data Annotation	22
2.4	Basic Tools of NLP	25
2.4.1	Tokenization and POS Tagging	25
2.4.2	Syntactic Parser	27
2.4.3	<i>N</i> -gram Language Model	29
2.5	Further Reading	30
3	Text Representation	33
3.1	Vector Space Model	33
3.1.1	Basic Concepts.....	33
3.1.2	Vector Space Construction	34
3.1.3	Text Length Normalization.....	36
3.1.4	Feature Engineering	37
3.1.5	Other Text Representation Methods	39
3.2	Distributed Representation of Words.....	40
3.2.1	Neural Network Language Model	41
3.2.2	C&W Model	45
3.2.3	CBOW and Skip-Gram Model	47
3.2.4	Noise Contrastive Estimation and Negative Sampling....	49
3.2.5	Distributed Representation Based on the Hybrid Character-Word Method.....	51

3.3	Distributed Representation of Phrases	53
3.3.1	Distributed Representation Based on the Bag-of-Words Model	54
3.3.2	Distributed Representation Based on Autoencoder	54
3.4	Distributed Representation of Sentences	58
3.4.1	General Sentence Representation	59
3.4.2	Task-Oriented Sentence Representation	63
3.5	Distributed Representation of Documents	66
3.5.1	General Distributed Representation of Documents	67
3.5.2	Task-Oriented Distributed Representation of Documents	69
3.6	Further Reading	72
4	Text Representation with Pretraining and Fine-Tuning	75
4.1	ELMo: Embeddings from Language Models	75
4.1.1	Pretraining Bidirectional LSTM Language Models	76
4.1.2	Contextualized ELMo Embeddings for Downstream Tasks	77
4.2	GPT: Generative Pretraining	78
4.2.1	Transformer	78
4.2.2	Pretraining the Transformer Decoder	80
4.2.3	Fine-Tuning the Transformer Decoder	81
4.3	BERT: Bidirectional Encoder Representations from Transformer	82
4.3.1	BERT: Pretraining	83
4.3.2	BERT: Fine-Tuning	86
4.3.3	XLNet: Generalized Autoregressive Pretraining	86
4.3.4	UniLM	89
4.4	Further Reading	90
5	Text Classification	93
5.1	The Traditional Framework of Text Classification	93
5.2	Feature Selection	95
5.2.1	Mutual Information	96
5.2.2	Information Gain	99
5.2.3	The Chi-Squared Test Method	100
5.2.4	Other Methods	101
5.3	Traditional Machine Learning Algorithms for Text Classification	102
5.3.1	Naïve Bayes	103
5.3.2	Logistic/Softmax and Maximum Entropy	105
5.3.3	Support Vector Machine	107
5.3.4	Ensemble Methods	110

5.4	Deep Learning Methods	111
5.4.1	Multilayer Feed-Forward Neural Network	111
5.4.2	Convolutional Neural Network	113
5.4.3	Recurrent Neural Network	115
5.5	Evaluation of Text Classification	120
5.6	Further Reading	123
6	Text Clustering	125
6.1	Text Similarity Measures	125
6.1.1	The Similarity Between Documents	125
6.1.2	The Similarity Between Clusters	128
6.2	Text Clustering Algorithms	129
6.2.1	<i>K</i> -Means Clustering	129
6.2.2	Single-Pass Clustering	133
6.2.3	Hierarchical Clustering	136
6.2.4	Density-Based Clustering	138
6.3	Evaluation of Clustering	141
6.3.1	External Criteria	141
6.3.2	Internal Criteria	142
6.4	Further Reading	143
7	Topic Model	145
7.1	The History of Topic Modeling	145
7.2	Latent Semantic Analysis	146
7.2.1	Singular Value Decomposition of the Term-by-Document Matrix	147
7.2.2	Conceptual Representation and Similarity Computation	148
7.3	Probabilistic Latent Semantic Analysis	150
7.3.1	Model Hypothesis	150
7.3.2	Parameter Learning	151
7.4	Latent Dirichlet Allocation	153
7.4.1	Model Hypothesis	153
7.4.2	Joint Probability	155
7.4.3	Inference in LDA	158
7.4.4	Inference for New Documents	160
7.5	Further Reading	161
8	Sentiment Analysis and Opinion Mining	163
8.1	History of Sentiment Analysis and Opinion Mining	163
8.2	Categorization of Sentiment Analysis Tasks	164
8.2.1	Categorization According to Task Output	164
8.2.2	According to Analysis Granularity	165

8.3	Methods for Document/Sentence-Level Sentiment Analysis	168
8.3.1	Lexicon- and Rule-Based Methods	169
8.3.2	Traditional Machine Learning Methods	170
8.3.3	Deep Learning Methods	174
8.4	Word-Level Sentiment Analysis and Sentiment Lexicon	
	Construction	178
8.4.1	Knowledgebase-Based Methods	178
8.4.2	Corpus-Based Methods	179
8.4.3	Evaluation of Sentiment Lexicons	182
8.5	Aspect-Level Sentiment Analysis	183
8.5.1	Aspect Term Extraction	183
8.5.2	Aspect-Level Sentiment Classification	186
8.5.3	Generative Modeling of Topics and Sentiments	191
8.6	Special Issues in Sentiment Analysis	193
8.6.1	Sentiment Polarity Shift	193
8.6.2	Domain Adaptation	195
8.7	Further Reading	198
9	Topic Detection and Tracking	201
9.1	History of Topic Detection and Tracking	201
9.2	Terminology and Task Definition	202
9.2.1	Terminology	202
9.2.2	Task	203
9.3	Story/Topic Representation and Similarity Computation	206
9.4	Topic Detection	209
9.4.1	Online Topic Detection	209
9.4.2	Retrospective Topic Detection	211
9.5	Topic Tracking	212
9.6	Evaluation	213
9.7	Social Media Topic Detection and Tracking	215
9.7.1	Social Media Topic Detection	216
9.7.2	Social Media Topic Tracking	217
9.8	Bursty Topic Detection	217
9.8.1	Burst State Detection	218
9.8.2	Document-Pivot Methods	221
9.8.3	Feature-Pivot Methods	222
9.9	Further Reading	224
10	Information Extraction	227
10.1	Concepts and History	227
10.2	Named Entity Recognition	229
10.2.1	Rule-based Named Entity Recognition	230
10.2.2	Supervised Named Entity Recognition Method	231
10.2.3	Semisupervised Named Entity Recognition Method	239
10.2.4	Evaluation of Named Entity Recognition Methods	241

10.3	Entity Disambiguation	242
10.3.1	Clustering-Based Entity Disambiguation Method	243
10.3.2	Linking-Based Entity Disambiguation	248
10.3.3	Evaluation of Entity Disambiguation	254
10.4	Relation Extraction	256
10.4.1	Relation Classification Using Discrete Features	258
10.4.2	Relation Classification Using Distributed Features	265
10.4.3	Relation Classification Based on Distant Supervision	268
10.4.4	Evaluation of Relation Classification	269
10.5	Event Extraction	270
10.5.1	Event Description Template	270
10.5.2	Event Extraction Method	272
10.5.3	Evaluation of Event Extraction	281
10.6	Further Reading	281
11	Automatic Text Summarization	285
11.1	Main Tasks in Text Summarization	285
11.2	Extraction-Based Summarization	287
11.2.1	Sentence Importance Estimation	287
11.2.2	Constraint-Based Summarization Algorithms	298
11.3	Compression-Based Automatic Summarization	299
11.3.1	Sentence Compression Method	300
11.3.2	Automatic Summarization Based on Sentence Compression	305
11.4	Abstractive Automatic Summarization	307
11.4.1	Abstractive Summarization Based on Information Fusion	307
11.4.2	Abstractive Summarization Based on the Encoder-Decoder Framework	313
11.5	Query-Based Automatic Summarization	316
11.5.1	Relevance Calculation Based on the Language Model ...	317
11.5.2	Relevance Calculation Based on Keyword Co-occurrence	317
11.5.3	Graph-Based Relevance Calculation Method	318
11.6	Crosslingual and Multilingual Automatic Summarization	319
11.6.1	Crosslingual Automatic Summarization	319
11.6.2	Multilingual Automatic Summarization	323
11.7	Summary Quality Evaluation and Evaluation Workshops	325
11.7.1	Summary Quality Evaluation Methods	325
11.7.2	Evaluation Workshops	330
11.8	Further Reading	332
	References	335