

Inhaltsverzeichnis

1	Text und Text Mining	1
1.1	Text Mining	1
1.1.1	Text ist Wissensrohstoff	1
1.1.2	Text Mining und Text-Mining-Werkzeuge	3
1.1.3	Text-Mining-Umgebungen	4
1.1.4	Was leistet Text Mining?	6
1.2	Text und Big Data	8
1.3	Aufbau und Struktur von Text	10
1.3.1	Arten von Text und ihre Merkmale	10
1.3.2	Zeichen, Types und Wörter	11
1.3.3	Nachricht, Information	12
1.3.4	Information und Wissen	14
1.3.5	Beispieltex te und Textressourcen	15
1.4	Redundanz in Texten	18
1.4.1	Arten von Redundanz	18
1.4.1.1	Nahe Wiederholung von Wörtern im Text	18
1.4.1.2	Wiederholung von Wortteilen	19
1.4.1.3	Wiederholung in der Struktur: Eigennamen	19
1.4.1.4	Kongruenz	20
1.4.1.5	Feste Wendungen	20
1.4.1.6	Explizite Wiederholung typischer Zusammenhänge in verschiedenen Sätzen oder Texten	21
1.4.2	Wirkung von Redundanz	21
1.4.2.1	Wiederholte Wörter: Wichtige Namen oder Schlagwörter	22
1.4.2.2	Wiederholte Substrings im Text: Klassifikation von Texten und Wörtern	22
1.5	Linguistische Strukturen	23
1.5.1	Texte und linguistische Ebenen	23

1.5.2	Warum erfordert die Verarbeitung natürlicher Sprache linguistisches Wissen?	24
1.5.3	Zwei Ansätze für die Repräsentation und Verarbeitung linguistischen Wissens.	27
	Literatur.	31
2	Linguistische Repräsentationen.	35
2.1	Theoretische Grundlage: Strukturalismus	36
2.1.1	Was ist die Grundidee des Strukturalismus?	36
2.1.2	Kontexte.	37
2.2	Morphologie.	40
2.2.1	Grundbegriffe.	40
2.2.2	Verarbeitungsparadigmen für die Morphologie.	43
2.3	Syntaktische Repräsentationen	46
2.3.1	Begriffsbestimmung: Was sind syntaktische Strukturen?	46
2.3.2	Konstituenten-Syntax.	47
2.3.3	Dependenzen	50
2.3.4	Probabilistisches Parsen.	51
2.4	Semantische Repräsentationen.	53
2.4.1	Grundbegriffe und Definition	53
2.4.2	Semantische Relationen.	57
2.5	Fachtexte und Terminologie.	62
2.5.1	Fachtexte	62
2.5.2	Terminologie	63
2.6	Die Rolle von Ausnahmen in der Sprache	65
2.6.1	Falsche Schreibweisen.	65
2.6.2	Seltene Ausnahmen	66
2.6.3	Auswirkungen dieser Sonderfälle	67
	Literatur.	68
3	Maschinelle Verarbeitung von Text	73
3.1	Verarbeitungsparadigmen für Text.	73
3.1.1	Regelbasierte Verarbeitung	75
3.1.2	Überwachte Statistische Verarbeitung	79
3.1.3	Neuronale Verarbeitung.	81
3.2	Die Linguistische Pipeline	85
3.2.1	Pipeline-Modell	85
3.2.2	Einlesen und Vorverarbeitung	89
3.2.3	Segmentierung	93
3.2.4	Morphologische und Syntaktische Verarbeitung	99
3.2.4.1	Grundformreduktion und Stammformreduktion.	99
3.2.4.2	Tagging mit Wortarten.	101

3.2.4.3	Chunking	102
3.2.4.4	Syntaxparsing	103
3.2.5	Semantische Verarbeitung	106
3.2.5.1	Eigennamenerkennung	106
3.2.5.2	Entity Linking	107
3.2.5.3	Koreferenzauflösung	108
3.2.5.4	Wortbedeutungsdisambiguierung	109
3.2.6	Anwendungsorientierte Verarbeitung	110
3.2.6.1	Pipeline für Terminologieextraktion	111
3.2.6.2	Pipeline für Entitätenzentriertes Retrieval und Facettierte Suche	111
3.2.6.3	Pipeline für Sentimentanalyse	112
3.2.6.4	Pipeline für Open Information Extraction	113
3.3	Skalierung auf große Datenmengen	114
3.3.1	Datenparallelität	114
3.3.2	Zusammenhang von Korpusgröße und Qualität	117
3.3.2.1	Der More-Data-Effect	118
3.3.2.2	Quantitatives Wachstum von Resultatmengen	119
3.3.2.3	Qualitative Verbesserung von Resultatmengen	122
	Literatur	123
4	Sprachdaten: Lexika und Korpora	131
4.1	Korpusauswahl	131
4.1.1	Generische Korpora	132
4.1.2	Selbst erstelltes Korpus	133
4.1.3	Dokumente oder Sätze? Nur wohlgeformte Sätze?	134
4.1.4	Repräsentativität und Ausgewogenheit der Zusammensetzung oder zufällige Auswahl?	135
4.1.5	Parallele Korpora	136
4.2	Satzkorpuserstellung auf Webdaten	136
4.2.1	Crawling	137
4.2.2	Beschränkung auf HTML-Dokumente	137
4.2.3	Text aus HTML-Dokumenten extrahieren	137
4.2.4	Qualitätssicherung	138
4.2.4.1	Sprachseparierung auf Dokumentenebene	139
4.2.4.2	Sprachüberprüfung auf Satzebene	140
4.2.4.3	Umgang mit Dubletten und Quasidubletten	140
4.2.4.4	Musterbasierte Entfernung nicht-wohlgeformter Sätze	146
4.2.5	Evaluiierung und Ranking von Sätzen	149
4.2.5.1	Ranking mittels GDEX	149
4.2.5.2	Typische Sätze	152

4.3	Speicherformate	153
4.4	Indexierung	156
4.4.1	Indexstrukturen	157
4.4.1.1	Klassisch: Einzelwort-Index, evtl. mit zusätzlichen Wortgruppen	157
4.4.1.2	Klassisch: Einzelwort-Index mit genauer Position ...	157
4.4.1.3	Spezielle Datenstrukturen für Textsuche: PAT Trees und die Anwendung in der NoSketch-Engine.	157
4.4.1.4	Ranking der Suchergebnisse	158
4.4.2	Verschiedene Typen von Suchanfragen	158
4.4.2.1	Beispielsätze für Wörter, Wortgruppen oder Kookkurrenzen	158
4.4.2.2	Beispielsätze für Grundformen	159
4.4.2.3	Kombination mehrerer Suchkriterien	159
4.4.2.4	Extraktion von Wortlisten mit bestimmten Eigenschaften	159
4.4.2.5	Extraktion von Relationen.	160
4.4.2.6	Suche unter Verwendung von Satzsignaturen	160
4.5	Manuell erstellte lexikalische Ressourcen	162
4.5.1	Klassische Wörterbücher	163
4.5.2	Verzeichnisse	164
4.5.2.1	Verzeichnisse von Personen	165
4.5.2.2	Verzeichnisse von Vornamen und Nachnamen	165
4.5.2.3	Geographische Eigennamen	165
4.5.3	Relationen zwischen Mengen von Wörtern.	166
4.5.3.1	Synsets	166
4.5.3.2	Erweiterung von Wortnetzen: Semantische Ähnlichkeit mit Word Embeddings.	167
4.6	Automatische Erweiterung lexikalischer Ressourcen	167
4.6.1	Klassische Wörterbuchangaben	168
4.6.1.1	Lemmatisierung: Vollform zu Grundform	168
4.6.1.2	Flexionsklasse zu Grundform	168
4.6.1.3	Wortart zu Grundform: POS-Tagging und NER.	169
4.6.1.4	Grammatisches Geschlecht zu Nomen	170
4.6.1.5	Kompositazerlegung	170
4.6.2	Statistische Angaben	171
4.6.2.1	Worthäufigkeiten	171
4.6.2.2	Wortpaare: Kookkurrenzen	172
4.6.2.3	Sentiment	173
4.6.2.4	Sachgebietsangaben	173
	Literatur.	174

5 Sprachstatistik	177
5.1 Statistische Messungen und ihre Zuverlässigkeit	178
5.1.1 Aufgaben aus der Sprachstatistik	178
5.1.2 Messgrößen der Sprachstatistik	179
5.1.3 Präsentation der Ergebnisse	179
5.1.4 Messungen und Messwerte	180
5.1.4.1 Beschreibung und Nachvollziehbarkeit der Messung	180
5.1.4.2 Abhängigkeit von der Wortdefinition am Beispiel der Type-Token-Ratio	181
5.1.4.3 Abhängigkeit von der Korpusgröße	182
5.1.4.4 Zählen mit oder ohne Wiederholungen: Mittlere Wortlänge	183
5.1.5 Abhängigkeit von der Zusammensetzung des Korpus	185
5.1.5.1 Abhängigkeit vom Text-Genre	185
5.1.5.2 Fehlen gesprochener Sprache	186
5.1.6 Abhängigkeit von Optionen bei der der Korpuserstellung	186
5.1.6.1 Verzerrte Ergebnisse durch mangelnde Qualität der Rohdaten	186
5.1.6.2 Fragwürdige Optionen	188
5.1.6.3 Umgang mit seltenen Ereignissen	188
5.2 Zipfsches Gesetz	189
5.2.1 Zusammenhang zwischen Rang, Häufigkeit und Wortlänge	189
5.2.2 Vorhersagen und Anwendungen	192
5.3 Kookkurrenzen	195
5.3.1 Struktur von signifikanten Kookkurrenzen	195
5.3.2 Maße für die statistische Signifikanz einer Kookkurrenz	197
5.3.3 Plausibilität der Ergebnisse	200
5.3.4 Andere Signifikanzmaße	200
5.3.5 Signifikante Kookkurrenzen – Beispiele und Anwendungen	201
5.3.6 Erste Anwendungen für signifikante Kookkurrenzen	203
5.3.6.1 Fremdsprachliche Daten im Text	203
5.3.6.2 Mundart	203
5.3.7 Signifikante Kookkurrenzen und Polysemie	204
5.3.8 Semantische Relationen zwischen signifikanten Kookkurrenzen	205
5.3.9 Visualisierung von signifikanten Kookkurrenzen	206
5.4 Distributionelle Semantik	208
5.5 Sprachmodelle	212
5.6 Dense Vector Embeddings	218
5.6.1 Statische Word Embeddings	219
5.6.2 Statische Word Embeddings: Wortähnlichkeit und Analogie	222

5.6.3	Kontextualisierte Word Embeddings	224
5.6.4	Embeddings für Sätze und Texte	226
5.6.5	Evaluation von Embeddings	227
5.7	Wortbedeutungsinduktion	228
5.8	Qualitätsmaße für Korpora.	231
5.8.1	Statistische Abweichungen von der erwarteten Verteilung	232
5.8.2	Verwendung von Vergleichskorpora	236
5.8.3	Musterbasierte Abweichungen.	240
5.9	Statistischer Korpusvergleich.	242
5.9.1	Voraussetzungen und Ziele	242
5.9.2	Wortvergleiche.	243
5.9.2.1	Log-likelihood-Ratio.	244
5.9.2.2	tf-idf (term frequency/inverse document frequency)	246
5.9.2.3	Statistische Hypothesentests und Burrows Zeta	246
5.9.3	Kontextvergleiche	247
5.9.3.1	Embeddingbasierte Verfahren	247
5.9.3.2	Kookkurrenzstatistik	248
	Literatur.	251
6	Maschinelles Lernen für Sprachverarbeitung	257
6.1	Merkmalsextraktion	258
6.2	Clustering.	261
6.3	Beispiele für Clustering	264
6.3.1	Hierarchisches Clustern von Wortarten	265
6.3.2	Wortbedeutungsinduktion mit Graph Clustering.	267
6.3.3	Eventerkennung mit inkrementellem Clustering.	268
6.3.4	Dokumentclustering mit k-Means	269
6.4	Topic-Modelle	270
6.4.1	Dokumente enthalten Themenstränge (<i>Topics</i>)	270
6.4.2	Modellierung von Topics	272
6.4.3	Evaluation und Best Practices	275
6.5	Evaluation von Clustering	277
6.6	Klassifikation	279
6.6.1	Definition und Arten von Klassifikation	280
6.6.2	Aufteilung der Beispieldaten.	280
6.6.3	Beispiele für Klassifikationsalgorithmen.	283
6.6.3.1	Naïve Bayes	283
6.6.3.2	Entscheidungsbäume.	284
6.6.3.3	Neuronale Netzwerke für die Klassifikation.	285
6.7	Annotation: Erstellung von Trainingsdaten	286
6.7.1	Durchführen von Annotationsprojekten.	286

6.7.2	Annotationsebenen und Tools zur manuellen Annotation	288
6.7.3	Datensatzerstellung mit Crowdsourcing	290
6.8	Sequenzklassifikation.	291
6.9	Evaluation von Klassifikation	295
6.9.1	Mengenorientierte Evaluationsmaße	296
6.9.2	Andere Evaluationsmaße im Text Mining	298
6.10	Neuronale Methoden: End-to-End, Transfer Learning	299
6.10.1	End-to-End-Lernen	299
6.10.2	Transferlernen	301
	Literatur.	302
7	Beispielanwendungen	311
7.1	Terminologieextraktion	312
7.1.1	Arbeitsablauf der Terminologieextraktion.	312
7.1.2	Verfahren der Terminologieextraktion.	313
7.2	Facettierte Suche mit Eigennamen.	317
7.2.1	Tagesnetzwerk: Visuelle Nachrichtenzusammenfassung.	318
7.2.2	Storyfinder: Eigene Lesehistorie	319
7.2.3	Investigativtool New/s/leak	319
7.2.4	Zusammenfassung	321
7.3	Sentimentanalyse	322
7.3.1	Begriffsbestimmung, Einsatzbereiche, Herausforderungen	322
7.3.2	Lösungsansätze und Aufgaben	323
7.4	Trendanalysen und News Monitoring	327
7.4.1	Trends und schwache Signale	327
7.4.2	News Monitoring.	327
7.4.3	Wörter des Tages	328
7.5	Neologismen	332
7.5.1	Neologismenwörterbücher.	333
7.5.2	Hochfrequente Neologismen 2010–2020	337
7.6	Kontextvolatilität	338
7.6.1	Kurze Zusammenfassung des Verfahrens	338
7.6.2	Beispielanwendung	340
	Literatur.	342
	Glossar	347
	Stichwortverzeichnis	381