Inhaltsverzeichnis

Vorwort **I**

Inhaltsverzeichnis **III**

1 Von Natural Language Processing zu Agentic NLP **11**

1.1 Was ist Natural Language Processing? 11

1.2 Warum natürliche Sprache schwierig ist 13

1.3 Verarbeitungsebenen natürlicher Sprache 14

1.4 Historische Entwicklung des NLP 16

1.5 Typische Aufgaben des NLP 18

1.6 Anwendungsfelder in Organisationen 19

1.7 Der typische Ablauf eines NLP-Projekts 21

1.8 Werkzeuge und Ökosysteme 23

1.9 Prozessmodell für agentische NLP-Systeme 24

1.10 Was Agentic NLP auszeichnet 25

1.11 Zusammenfassung 26

2 Sprachwissenschaftliche Grundlagen für NLP **27**

2.1 Phonetik und Phonologie 28

2.2 Morphologie: der Aufbau von Wörtern 29

2.3 Lexika, Wortarten und Mehrwortausdrücke 32

2.4 Syntax 33

2.5 Semantik 36

2.6 Pragmatik 38

2.7 Diskurs 39

2.8 Mehrdeutigkeit als zentrales NLP-Problem 41

2.9 Zusammenfassung 42

3 Machine-Learning-Grundlagen für NLP **43**

3.1 Von sprachlichen Daten zur Lernaufgabe 43

3.2 Merkmale, Labels und Zielvariablen 44

3.3 Überwachtes und unüberwachtes Lernen 46
3.4 Training, Validierung und Test 47
3.5 Klassifikation und Regression 48
3.6 Overfitting und Generalisierung 49
3.7 Zentrale Bewertungsmaße 50
3.8 Machine Learning, Deep Learning und generative KI 54
3.9 Machine-Learning-Pipeline im NLP 55
3.10 Zusammenfassung 57
4 Entwicklungsumgebung und Open-Source-Stack 58
4.1 NLP-Projekte systematisch entwickeln 58
4.2 Entwicklungsumgebung 64
5 Python kompakt für NLP 102
5.1 Text und Tokens 102
5.2 Elementare Datentypen 104
5.3 Zeichenketten, Unicode und Kodierungen 106
5.4 Datenstrukturen 107
5.5 Kontrollstrukturen 110
5.6 Funktionen und Methoden 113
5.7 Dateien und Pfade 115
5.8 Dataframes und Series 118
5.9 Arrays und Sparse-Matrizen 119
5.10 Tensoren für neuronale NLP-Modelle 122
5.11 Reguläre Ausdrücke 124
5.12 Generatoren und Batches 126
5.13 Zusammenfassung 127
6 Textvorverarbeitung 129
6.1 Dokumente einlesen und Nutztext extrahieren 130
6.2 Textverarbeitung mit spaCy 136

6.3 Tokenisierung 147
6.4 Satzsegmentierung 155
6.5 Normalisierung von Schreibvarianten 162
6.6 Lemmatisierung und Stemming 166
6.7 Stoppwörter 169
6.8 Emojis, Hashtags und andere Sonderformen 173
6.9 Rauschen, Fehler und sprachliche Variation 178
6.10 Technische Datenqualität und Qualitätskontrolle 186
6.11 Reproduzierbare Verarbeitungsketten 191
6.12 Durchgängiges Fallbeispiel: Wartungsberichte 194
6.13 Zusammenfassung 197
7 Korpora, Daten und Annotation 199
7.1 Korpusbegriff und Korpustypen 199
7.2 Frage, Grundgesamtheit und Stichprobe 200
7.3 Datensammlung und Importformate 202
7.4 Metadaten und Dokumentidentifikation 204
7.5 Annotationsebenen und Repräsentationen 205
7.6 Annotationsrichtlinien und Validierung 207
7.7 Übereinstimmung zwischen Annotierenden 209
7.8 Datensplits ohne Datenleckage 211
7.9 Klassenverteilung und Datenaugmentation 212
7.10 Abdeckung, Repräsentativität und Bias 214
7.11 Datenschutz und Nutzungsrechte 216
7.12 Korpusdokumentation und Korpuskarten 218
7.13 Versionierung und Reproduzierbarkeit 220
7.14 Zusammenfassung 220
8 Klassische Textrepräsentationen 222
8.1 Vom Text zum Merkmalsvektor 223

8.2 Bag-of-Words und Dokument-Term-Matrix 224
8.3 Zählwerte und binäre Merkmale 226
8.4 Wort-N-Gramme 227
8.5 TF-IDF-Gewichtung 229
8.6 Begrenzung des Merkmalsraums 230
8.7 Kosinusähnlichkeit 231
8.8 Dünn besetzte Matrizen und Effizienz 233
8.9 Statistische Sprachmodelle als Baseline 234
8.10 Überwachtes Machine Learning als Baseline 243
8.11 Zusammenfassung 274
9 Klassische NLP-Aufgaben 277
9.1 Part-of-Speech-Tagging 278
9.2 Named Entity Recognition 280
9.3 Syntaktisches Parsing und Abhängigkeiten 283
9.4 Clustering 287
9.5 Textklassifikation 292
9.6 Sentimentanalyse 294
9.7 Information Retrieval 296
9.8 Information Extraction 298
9.9 Relationen und Ereignisse extrahieren 301
9.10 Zusammenfassung 305
10 Embeddings und semantische Suche 307
10.1 Distributionelle Semantik und Kontextfenster 308
10.2 Kookkurrenzmatrizen 309
10.3 Word2Vec 310
10.4 FastText und Teilwortinformationen 322
10.5 Unbekannte Wörter und Fachkomposita 323
10.6 Statische und kontextuelle Embeddings 325

10.7 Satz- und Dokument-Embeddings 327
10.8 Zusammenfassung 328
11 Deep Learning für NLP 330
11.1 Tensoren, Rechengraphen und Gradienten 331
11.2 Feedforward-Netze für Textklassifikation 332
11.3 Embedding-Schichten als trainierbare Eingabe 334
11.4 Attention-Mechanismen 336
11.5 Regularisierung und stabiler Trainingsbetrieb 339
11.6 Training, Evaluation und Modellvergleich 342
11.7 Durchgängige Fallstudie: Wartungsberichte klassifizieren 346
11.8 Rechenressourcen, Reproduzierbarkeit und Sicherheit 349
11.9 Zusammenfassung 350
12 Transformer 352
12.1 Von rekurrenten Netzen zu Transformern 352
12.2 Self-Attention und Multi-Head-Attention 353
12.3 Positionskodierungen 354
12.4 Masken für variable und gerichtete Sequenzen 356
12.5 Der Transformer-Encoderblock 357
12.6 Encoder, Decoder und Encoder-Decoder 359
12.7 Encoder für Textklassifikation 363
12.8 Decoder-Modelle und kausale Attention 364
12.9 Decoder GPT-Modelle 365
12.10 BERT: Bidirektionale Encoder 367
12.11 Encoder-Decoder-Modelle 369
12.12 T5 und das Text-zu-Text-Prinzip 371
12.13 Vortraining, Fine-Tuning und Transfer 372
12.14 Tensorformen in Transformer-Programmen 373
12.15 Zusammenfassung 374

13 Open-Source-LLMs 376
13.1 Tokenisierung und Kontextfenster 379
13.2 Autoregressives Vortraining 380
13.3 Skalierung, Parameter und Ressourcen 382
13.4 Modellanpassung 384
13.5 Prompts und strukturierte Ausgaben 407
13.6 Halluzinationen und Wissensgrenzen 412
13.7 Zusammenfassung 416
14 Retrieval-Augmented Generation für Unternehmenswissen 419
14.1 RAG als Systemarchitektur 419
14.2 Vom Retrieval zum RAG-Prompt 420
14.3 Chunking, Retrieval und Reranking 425
14.4 Belegtreue Antworten mit RAG 426
14.5 RAG versus Fine-Tuning und LoRA 432
14.6 Zusammenfassung 433
15 Agentic NLP: Tools, Zustände und Workflows 434
15.1 Vom LLM-Aufruf zum Agenten 434
15.2 Agent, Workflow und Automation unterscheiden 435
15.3 Tool Calling und kontrollierte Werkzeuge 436
15.4 State und Memory 441
15.5 Agentische Abläufe und Kontrollschleifen 441
15.6 Deterministischer Workflow oder Agent? 447
15.7 Grenzen, Freigaben und Beobachtbarkeit 448
15.8 Zusammenfassung 448
16 Generative Sprachsysteme und Dialoganwendungen 451
16.1 Maschinelle Übersetzung 452
16.2 Extraktive Zusammenfassung 454
16.3 Abstraktive Zusammenfassung 456

16.4 Frage-Antwort-Systeme 458
16.5 Dialogsysteme und Chatbots 461
16.6 Steuerung generativer Textausgaben 464
16.7 Rechtschreibprüfung und Umformulierung 467
16.8 Zusammenfassung 470
17 Mehrsprachige und multimodale Sprachsysteme 472
17.1 Mehrsprachigkeit als Systemanforderung 472
17.2 Schriftsysteme, Unicode und Transliteration 473
17.3 Sprachidentifikation 474
17.4 Mehrsprachige Korpora und Datenbalance 477
17.5 Sprachabhängige Normalisierung 480
17.6 Multilinguale Tokenisierung 483
17.7 Translation Memory, LLM und Kontrolle 484
17.8 Sprachübergreifende semantische Suche 488
17.9 Low-Resource-Sprachen 490
17.10 Zusammenfassung 494
18 Praxisprojekte: NLP-Lösungen 496
18.1 Vom Einzelverfahren zur Praxisarchitektur 496
18.2 Projektstruktur und externe Ressourcen 497
18.3 Serviceanfragen klassifizieren 499
18.4 Sentimentanalyse 502
18.5 Informationsextraktion 525
18.6 Unternehmensdokumente durchsuchen 533
18.7 Berichte mit Instruct-LLM verdichten 536
18.8 RAG für Unternehmenswissen 540
18.9 Chatbot erstellen 543
18.10 Sprachgesteuerter Serviceassistent 558
18.11 Zusammenfassung 566

19 Praxisprojekte: NLP-Anwendungen 569
19.1 Systemarchitektur der NLP-Anwendung 570
19.2 Modellgewichte einmalig beschaffen 574
19.3 Neuronale Übersetzung mit MarianMT 575
19.4 FAISS-Index-Datei laden 576
19.5 Beleggestützte Antwortgenerierung mit Qwen3 578
19.6 Grafische Windows-Anwendung mit Tkinter 579
19.7 Windows-Anwendung erzeugen 582
19.8 Windows-Anwendung testen 584
19.9 Verteilstruktur 588
19.10 Zusammenfassung 588
20 Ausblick 590
20.1 Von Modellleistung zu Systemleistung 590
20.2 Entwicklungslinien für Agentic NLP 590
20.3 Rolle des Menschen 591
20.4 Schlussbetrachtung 592
Literaturverzeichnis 593