Chris Biemann · Gerhard Heyer · Uwe Quasthoff

Wissensrohstoff Text

Eine Einführung in das Text Mining

2., wesentlich überarbeitete Auflage

Springer Vieweg

Glossar

Vorwort zum Glossar: Das Glossar erklärt wichtige Begriffe des Text Mining, die nicht als allgemein bekannt vorausgesetzt werden. Dies können sowohl wiederkehrende Begriffe aus dem Text wie auch weiterführende Erklärungen für im Text nicht näher erläuterte Begriffe sein. In der alphabetischen Sortierung sind bei Wortgruppen ggf. enthaltene Adjektive nachgestellt wie bei *Lernverfahren, überwachte*. Nach dem Stichwort folgt zusätzlich die englische Übersetzung. Falls für das Verständnis nötig, wird im Definitionstext mit einem Verweispfeil ▸ auf weitere Begriffe im Glossar verwiesen.

Accuracy (accuracy) auch Treffergenauigkeit Accuracy ist ein ▸ **Evaluationsmaß** zur Bewertung von Klassifikationssystemen. Sie ist definiert als die Anzahl aller korrekten Vorhersagen dividiert durch die Gesamtanzahl an Vorhersagen. Dementsprechend ist die maximal erreichbare Accuracy 1 und wird erreicht, wenn jede Vorhersage korrekt ist, wohingegen eine Accuracy von 0 erreicht wird, wenn keine der Vorhersagen korrekt ist. Sie eignet sich nur für die Bewertung von einigermaßen balancierten Datensätzen, in denen also alle Klassen etwa gleich häufig vorkommen.

Affektwörterbuch (affect dictionary) Lexikon von typischen ▸ **Sentimentausdrücken.**

Ähnlichkeitsmaß (similarity measure) Ein Ähnlichkeitsmaß ist im ▸ **Information Retrieval** ein Maß für die Ähnlichkeit von Wörtern oder Texten (als einer Kollektion von Wörtern). Einem Paar von Textobjekten wird eine reelle Zahl S ($0 <= S <= 1$) zugeordnet, die umso größer ist, je ähnlicher die Textobjekte sind. In der Praxis häufig verwendete Maße sind das ▸ **Kosinusmaß,** der Jaccard-Koeffizient und der Dice-Koeffizient.

Allomorph (allomorph) Verschiedene Varianten eines ▸ **Morphems.**

Allgemeinsprache (general language) Unter der Allgemeinsprache versteht man die Sprache in ihrer allgemeinen und alltäglichen Verwendung. Antonym: ▸ **Fachsprache.**

Alphabet (alphabet) Ein Alphabet ist ein endlicher, geordneter Zeichenvorrat für die Bildung von Wörtern.

C. Biemann et al., *Wissensrohstoff Text,* https://doi.org/10.1007/978-3-658-35969-0

Analysekorpus (corpus for analysis) Das Analysekorpus ist der zu untersuchende Text bzw. die zu untersuchende Textmenge.

Annotation/Annotatoren (annotations/annotators) Annotationen sind Klassifizierungen nach einem vorgegebenen Kategorienschema, welche sich, im Kontext der Computerlinguistik, auf Sprache beziehen, etwa indem sie Dokumente eine Kategorie zuweisen oder Worte mit ihrer Wortart versehen. Annotatoren sind dabei Menschen oder automatische Komponenten, welche Annotationen erstellen.

Ansatz, korpusbasierter (corpus-based approach) Vorgehensweise in der Lexikographie, bei der ein Textkorpus die Grundlage bei der Auswahl von Stichwörtern und Verwendungsbeispielen bildet.

Antonyme (antonyms) Zwei Wörter, die nur in Bezug auf einen vorher definierten Begriff einen Gegensatz bilden. Syn.: Relative Gegensätze.

Anwender (user) Siehe ▸ **Benutzer.**

Anwendungssoftware (application software) Software, die Aufgaben des Anwenders mithilfe eines Computersystems löst. Setzt in der Regel auf der Systemsoftware der verwendeten Hardware auf bzw. benutzt sie zur Erfüllung der eigenen Aufgabe.

Äquivalenzklasse (equivalence class) Menge von Elementen, die im Hinblick auf eine (reflexive, symmetrische und transitive) Äquivalenzrelation äquivalent sind.

ASCII (ASCII; American Standard Code of Information Interchange) Genormter 7-Bit-Zeichensatz (128 Positionen) zur Darstellung von Ziffern, Buchstaben, Sonderzeichen und Steuerzeichen. Siehe auch ▸ **Latin** und ▸ **Unicode.**

Aspekt-basierte Sentimentanalyse (aspect-based sentiment analysis) Ansatz der ▸ **Sentimentanalyse,** bei der unterschieden wird zwischen dem sog. document level, den Aspektkategorien, den sentiment targets und den auf den Text sowie die verschiedenen Aspektkategorien bezogenen ▸ **Polaritäten.**

Asset (Asset) Ein Asset (Wert) ist ein sinnhaltiges Datenobjekt als Wirtschaftsgut. Syn.: Wert.

Attention-Mechanismus (attention mechanism) In ▸ **neuronalen Netzarchitekturen** Modellierung des Informationsflusses zwischen den Wörtern eines Satzes oder eines kurzen Textes in Abhängigkeit von den Eingangsrepräsentationen, um Ausgangsrepräsentationen zu berechnen: Das Netzwerk lernt, wie stark die anderen Positionen in der Sequenz zu beachten sind, um die aktuelle Position zu repräsentieren. Elementarer Bestandteil von ▸ **Transformer-Architekturen.**

Attribute, extrinsische (extrinsic attributes) Eigenschaften von Wörtern, welche man den typischen Kontexten der Wörter entnehmen kann, beispielsweise das grammatische Geschlecht von Nomen durch benachbarte Artikel oder typische Eigenschaften durch benachbarte Adjektive.

Attribute, intrinsische (intrinsic attributes) Eigenschaften von Wörtern, welche man der inneren Struktur der Wörter entnehmen kann, beispielsweise durch Kompositazerlegung oder morphologische Analyse.

Automat, endlicher (finite state automaton) Modell für einen Automaten, der Daten Zeichen für Zeichen einliest, das eingelesene Zeichen sofort verarbeitet und eine Ausgabe erzeugt. Ein endlicher Automat besitzt eine endliche Menge von Zuständen und keinen zusätzlichen Speicher.

Automatentheorie (automata theory) Teilgebiet der theoretischen Informatik, das Automaten als formale Systeme zur Beschreibung von Sprachen und Grammatiken zum Gegenstand hat.

Backpropagation (back propagation) Ein Algorithmus zur effizienten Berechnung der Gradienten innerhalb eines neuronalen Netzwerks gegeben der Eingabe und bezüglich der Abweichung von der gewünschten Ausgabe (beschrieben durch eine ▸ **Zielfunktion**), also die nötigen Veränderungen um die gewünschte Ausgabe zu erreichen. Basierend auf den berechneten Gradienten werden die Gewichte des Netzwerks angepasst.

Bag-of-Words Modell (bag-of-words model) Modell zur Repräsentation eines Textes anhand der in ihm vorkommenden Wörter und deren Anzahl, unabhängig von der exakten Position ihres Vorkommens im Text. Dabei gehen notwendigerweise Informationen verloren, für viele Aufgaben ist diese Repräsentation jedoch bereits ausreichend.

Baseline (baseline) auch unterer Leistungswert: Die Baseline ist eine untere Schranke der Leistung eines Algorithmus. Wenn ein komplexer Algorithmus entwickelt wird, wird oft ein einfacher Algorithmus (Baseline-Verfahren) dagegen gestellt. Der komplexe Algorithmus muss dann bessere Ergebnisse erzielen als der einfache Algorithmus.

Basismorphem (base morpheme) Morpheme, die Sachverhalte der außersprachlichen Welt bezeichnen. Zu ihnen gehören Nomina wie *Kind* und *Mantel,* aber auch Verbstämme wie *seh.* Syn.: Stamm.

Bayessche Statistik (Bayesian statistics) Die Wahrscheinlichkeit eines Ereignisses wird in der Bayesschen Statistik als Erwartungswert interpretiert, der sich aus einer Bewertung bisheriger Beobachtungen ableitet. Dabei werden das Vorwissen und sog. A-Priori-Annahmen explizit ausdrückt.

Begriff (notion) Ein Begriff ist die Bedeutung eines Ausdruckes. Der Begriffsumfang wird als Extension, der Begriffsinhalt als Intension bezeichnet.

Benutzer und Benutzerinnen (user) Personen, die ein Computersystem unmittelbar einsetzen und selbst bedienen. Syn.: Anwender und Anwenderinnen.

Big Data (big data) Technologien und Algorithmen für die Speicherung und Auswertung digitaler Massendaten. Zu den Verarbeitungsdimensionen zählen nach dem sog. V-Modell das Volumen der Datenmenge, die Vielzahl der abzubildenden Inhalte und die Verarbeitungsgeschwindigkeit der Recherche- und Analyseschritte.

Bi-Gramm (bi-gram) Ein Bi-Gramm ist ein spezieller Typ von ▸ **n-Grammen,** das aus zwei aufeinander folgenden Wörtern oder Buchstaben besteht. Siehe ▸ **n-Gramm** und ▸ **Tri-Gramm.**

BiLSTM (BiLSTM) Bidirektionelle Variante der ▸ **Long Short-Term Memories.** Dabei wird die Ausgabe zu jedem Schritt in der Sequenz durch eine Kombination aus zwei Sequenzmodellen generiert, wobei das eine Modell seine Schritte vom Start der Sequenz hin zum Ende macht und das andere aus der Gegenrichtung, also vom Ende aus, seine Eingaben erhält. So kann z. B. jedes Wort eines Satzes in Abhängigkeit sowohl seiner Vorgänger als auch seiner Nachfolger modelliert werden.

BIO-Schema (BIO scheme) Begin-Inside-Outside ist ein Kodierung für ▸ **Sequenzklassifikation,** dabei wird jeweils der Beginn einer ▸ **Spannenanotation** (also etwa der erste Token) mit dem „Begin" (Beginn) Label versehen, jedes weitere Element der zu annotierenden Spanne erhält ein „Inside" (Innen) Label. Alle Elemente, die von keiner Annotation abgedeckt sind, erhalten das „Outside" (Außen) Label.

Bootstrapping (bootstrapping) Bootstrapping ist die allgemeine Bezeichnung für einen Prozess, bei dem unter Zuhilfenahme einer kleinen Anfangskonfiguration eine größere, umfassendere Konfiguration erzeugt wird. Im engeren Sinne ist es eine Bezeichnung für ein maschinelles Lernverfahren, bei dem neue Information generiert wird aus einer Startmenge an Information und einer Regelmenge, wie durch die schon bekannte Information neue Information gefunden werden kann.

Buchstaben (letters) Die in einem Alphabet verwendeten Zeichen.

Chomsky-Grammatik (Chomsky grammar) System von Ersetzungsregeln für die Erzeugung bzw. Analyse von formalen und natürlichen Sprachen. Abhängig davon, welche Einschränkungen bei der Ersetzung von Zeichenketten zu beachten sind, unterscheidet man Grammatiken für reguläre, kontextsensitive, kontextfreie und unbeschränkte Sprachen.

Chunking (chunking) Beim Chunking werden grammatikalisch zusammenhängende Phrasen als solche markiert. Im Gegensatz zu Dependenzrelationen bilden Chunks dabei keine hierarchischen, sondern stattdessen eine flache Struktur aus sich nicht überschneidenden ▸ **Spannenannotationen.**

Chunks (chunks) Chunks im Sinne des ▸ **Chunking** sind zusammenhängende grammatikalische Phrasen, etwa Nominalphrasen wie „der schnelle Hund" aber auch Verbal- und Präpositionalphrasen.

Cluster-Analyse (cluster analysis, clustering) Bei der Cluster-Analyse im Text Mining wird eine Menge sprachlicher Elemente (Texte, Sätze, Wörter) durch Gruppierung zusammengehöriger Elemente in homogene Teilmengen – die Cluster (Gruppen, Kategorien, Klassen) – unterteilt. Im Gegensatz zur Vorgehensweise bei der Klassifikation werden die Cluster aus der Struktur der zu analysierenden Datenmenge selbst abgeleitet.

Clustering (Clustering, cluster analysis) Siehe ▸ **Cluster-Analyse.**

Codepages (code pages) Tabellen zur Enkodierung von Zeichensätzen wobei verschiedene Codepages verwendet werden um Alphabete verschiedener Sprachen abzubilden, so bildet die ▸ **Latin** Codepage etwa westeuropäische Sprachen ab. Dementsprechend ist es für die korrekte Dekodierung eines (binär) kodierten Textes notwendig, die richtige Codepage zu verwenden.

Cohens Kappa (Cohen's kappa) Dies ist ein Maß für ▸ **Interrater-Reliabilität** zwischen zwei Annotierenden auf demselben Datensatz.

Compiler (compiler) Systemprogramm zur Überführung eines Computerprogramms in ausführbaren Maschinencode.

Content (content) Content (Inhaltsobjekt) enthält Informationen als sinnhaltige Datenobjekte. Syn.: Inhaltsobjekt.

Content-Management-System (content management system, CMS) Softwaresysteme zur Erstellung, Verwaltung, Strukturierung und Publikation verschiedener Medien. Im organisationsinternen Kontext ist darunter oft eine Dokumentenverwaltungssoftware zu verstehen, wobei CMSe auch verwendet werden, um Webseiteninhalte zu verwalten.

Continuous Bag-of-Words (CBOW) Modell (CBOW model) Neuronales Netzwerk zum Erlernen von ▸ Word **Embeddings** in Form von ▸ **dichtbesetzten Vektoren.** Die Trainingsaufgabe des Netzwerkes ist das Vorhersagen eines Wortes anhand seines Kontextes (Wörter links und rechts des aktuellen Wortes) mittels eines ▸ **Sliding Window** Verfahrens. Die Reihenfolge der Wörter im Kontext ist dabei irrelevant.

Convolutional Neural Networks Neuronale Netzwerke, in denen Feed-Forward Operationen mit den gleichen Gewichten auf mehreren, lokal begrenzten, Teilen der Eingabe ausgeführt werden. Dabei müssen die Eingabedaten strukturiert, etwa eindimensional (z. B. Zeitreihe) oder zweidimensional (wie z. B. in einem Bild) vorliegen. Sie werden insbesondere in der Bildverarbeitung, aber auch in der Sprachverarbeitung verwendet.

Crawler, auch Webcrawler (crawler/webcrawler) Computerprogramm zum automatischen Herunterladen großer Mengen von Webseiten auf der Basis vorgegebener Kriterien. Ausgehend von einer Menge von Startseiten werden die in den Webseiten gefundenen Links jeweils weiterverfolgt.

Crawlingstrategie (crawling strategy) Reihenfolge, nach der die bekannten, aber noch nicht heruntergeladenen Webseiten durch einen Crawler aufgesucht werden. Diese Warteliste kann viele Millionen Seite umfassen und sollte in einer Reihenfolge abgearbeitet werden, dass die verschiedenen Bereiche des Web möglichst gleichmäßig besucht werden.

Data Mining (data mining) Unter dem Begriff des Data Mining werden Verfahren aus der Statistik und künstlichen Intelligenz zusammengefasst, mit denen sich in strukturierten Datenbeständen Muster und statistische Zusammenhänge berechnen lassen.

Data Warehouse (data warehouse) Kopie operativer Daten, speziell für Anfragen und Analysen strukturiert. Die Datenorganisation erfolgt in Hyperwürfeln.

Daten (data) Daten im Sinne der Informatik und Datenverarbeitung (EDV) sind (maschinen-) lesbare und bearbeitbare Repräsentationen von Information. Die Information wird dazu in Zeichenketten kodiert, deren Aufbau strengen Regeln folgt. Daten werden zu Information, wenn sie unter Bezug auf einen Interpretationsschlüssel interpretiert werden.

DAWG (directed acyclic word graph) Ein DAWG ist eine Datenstruktur zum Speichern von Wörtern. Das Funktionsprinzip ist ähnlich dem eines Trie, aber zusätzlich werden gleiche Wortenden zusammengeführt und damit komprimiert gespeichert.

Deklination (declination) Flexion (Beugung) von Nomina und Adjektiven. Verändert wird dabei die Anzahl (Numerus) und der Fall (Kasus), bei Adjektiven kann zusätzlich auch das grammatische Geschlecht (Genus) verändert werden.

Dendrogramm (dendrogram) Ein Dendrogramm dient der schematischen Darstellung einer ▸ **Cluster-Analyse.** In ihm werden die ermittelten Teilmengenbeziehungen dargestellt. Je höher im Baum der Verschmelzungspunkt zweier Mengen liegt, umso größer ist die Distanz zwischen ihnen im Vektorraum und umso kleiner ist die Ähnlichkeit zueinander.

Dependenzen (syntactic dependencies) Die zwischen den Wörtern eines Satzes bestehenden Abhängigkeiten im Sinne der Dependenzgrammatik. Dabei wird angenommen, dass jedes Wort in einem Satz ein, und nur ein übergeordnetes Wort hat, von dem es abhängt. Die Relation zwischen übergeordneten und untergeordneten Wort heißt Head-Modifier-Relation.

Dependenz-Struktur (dependency structure) Darstellung der Dependenzen zwischen den Wörtern eines Satzes in Form eines Baumes, wobei jeder Knoten im Baum einem terminalen Wort entspricht. Als Wurzel einer Dependenzbaum-Struktur wird immer das Verb (und nicht eine abstrakte Kategorie Satz) angesetzt. Ähnlich der Konstituentenstruktur-Syntax können dabei auch einzelne Knoten benannt werden, etwa mit den Grundkategorien N, V, A, P, Art und Mod.

Derivation (derivation) Morphologischer Prozess, durch den neue Wörter durch Anfügung (Affigierung) von Derivativen an Wortstämme entstehen.

Derivationsaffix (derivational affix) Siehe ▸ **Derivative.**

Derivative (derivatives) Gebundene Morpheme, die bei der Derivation verwendet werden. Die meisten Derivative haben eine charakteristische Bedeutung, die sie dem Stamm hinzufügen. Dabei ändern Derivative meist die Wortart des Stamms. Syn.: Derivationsaffixe.

Development-Menge (development set) auch: Validationsmenge (validation set): Bestandteil des Datensatzes aus Instanzen und Labels, bezüglich dessen ein Modell während seiner Entwicklung evaluiert wird. Dabei kann es vorkommen, dass das Modell implizit auf die Development-Menge und nicht generell auf Daten außerhalb der ▸ **Trainingsmenge** optimiert wird, weshalb es nötig ist einen finalen Test auf der ▸ **Testmenge** durchzuführen. Siehe auch Trainingsmenge und Testmenge.

Differenzanalyse (differential analysis) Verfahren zur Ermittlung von statistisch signifikanten Unterschieden in der Verwendung von Vokabularen, siehe auch Korpusvergleich.

Dirichlet-Verteilung (Dirichlet distribution) Multinomiale Wahrscheinlichkeitsverteilung. Die Steuerung erfolgt über einen ▸ **Hyperparameter.**

Disambiguierung (disambiguation) Disambiguierung beschreibt die Auflösung der Mehrdeutigkeit von Ausdrücken. Siehe auch: Wortbedeutungsdisambiguierung.

Dispersionsmaße (degree of dispersion) Häufig in der Stylometrie verwendete Verfahren, bei denen die Positionierung von Wörtern im Analyse- und Referenzkorpus miteinander verglichen werden.

Distanzmaß (distance measure) Angabe des relativen Abstands zwischen zwei Objekten. Typische Distanzmaße für reellwertige Vektoren sind z. B. die Euklidische Distanz oder die ▸ **Kosinusähnlichkeit.** Siehe ▸ **Tanimoto-Ähnlichkeit** für binäre Vektoren und ▸ **Levenshtein-Distanz (Editierdistanz)** für den Abstand zwischen Zeichenketten bzw. Wörtern.

Distributionelle Hypothese (distributional hypothesis) Annahme, dass Wörter, die in gleichen Kontexten auftreten, ähnliche Bedeutungen haben. Basis für die Operationalisierung der in der Tradition des Strukturalismus stehenden distributionelle Semantik, welche die Bedeutung von Wörtern nur anhand ihrer Kontexte charakterisiert.

Distributionale semantische Modelle (distributional semantic models) Modelle die Wörtern oder anderen ▸ **Instanzen** durch Darstellungen im Vektorraum repräsentieren, dabei werden Instanzen, basierend auf der ▸ **distributionellen Hypothese** durch ihren Kontext im ▸ **Trainingskorpus** beschrieben.

Dokumentenähnlichkeit (document similarity) Als Dokumentenähnlichkeit bezeichnet man die Gruppierung von Wörtern oder Texten nach inhaltlichen Kriterien in Bezug auf ein vorgegebenes Ähnlichkeitsmaß.

Dokumentkorpus (corpus of documents) Ein Korpus aus natürlichsprachlichen Texten mit oder ohne Annotationen und Metadaten, in dem gesamte Textdokumente abgelegt sind.

Drift (drift) Die Veränderung der Eigenschaften eines Datensatzes über die Zeit hinweg. So können sich in einem ▸ **Korpus** statistische Eigenschaften wie die Häufigkeit mancher Worte ändern, ebenso können sich etwa die Kontexte von Worten ändern, da sich ihre Semantik sich über die Zeit geändert hat.

Dublin Core (Dublin Core) De-facto Standard für Metadaten der Dublin Core Metadata Initiative (DCMI). Wesentliche Elemente des Standards sind Felder für einen Identifier, Technische Daten, Beschreibung des Inhalts, Festlegung beteiligter Personen und Rechte, Aspekte der Dokumentvernetzung und des Dokumentlebenszyklus. Alle Felder sind optional, können mehrfach auftauchen und in beliebiger Reihenfolge stehen.

Early-Stopping-Methode (early stopping) Eine Methode zur Vermeidung von Overfitting, dabei wird das Training eines Modells dann vorzeitig beendet, wenn, basierend auf Evaluation auf der ▸ **Development-Menge,** keine weitere Verbesserung des Modell mehr stattfindet. Oft wird dabei basierend auf einem „Patience“ (Geduld)- ▸ **Hyperparameter** nach entsprechend vielen Schritten ohne Verbesserung das Training beendet und der bis dahin beste Zustand des Modells gewählt.

Eigenname (named entity) Eigennamen sind referenzierende Bezeichner von Objekten verschiedener Kategorie (z. B. Personen, Firmen und Institutionen, Produkte und

Orte). Morphologisch sind Eigennamen nicht produktiv. Syntaktisch treten sie meist ohne Artikel auf und sind im Satz ersetzbar durch Proformen wie *er/sie/es* oder *dieser.*

Eigennamenerkennung (named entity recognition, NER) Verarbeitungsschritt, in dem Eigennamen wie Personennamen, Ortsnamen, Organisationsnamen und sonstige Namen im Text markiert werden. Wird meist mit Hilfe von ▸ **Sequenzklassifikation** durchgeführt, wobei Mehrwortnamen mithilfe des ▸ **BIO-Schemas** kodiert werden.

Einleseprozess (data ingestion) Der Prozess des Einlesens von Daten in ein System. Dabei gilt es, als Teil dieses Prozesses, Daten in ein einheitliches Format zu bringen, um weitere Verarbeitungsschritte zu vereinfachen. Im Text Mining kann dies das Überführen in ein reines Textformat bedeuten und kann auch das Einlesen von Metadaten beinhalten.

Embedding (embedding) Repräsentation von lexikalischen Einheiten wie Wörtern, Sätzen oder Texten, welche diese Einheiten in einen hochdimensionalen Vektorraum einbetten. Ein Embedding besteht aus einem reellwertigen Vektor, welcher als Punkt in diesem Raum interpretiert werden kann. Embeddings werden typischerweise durch Trainieren auf ▸ **Hintergrundkorpora** generiert und sind besonders als Eingabe für neuronale Netze geeignet.

End-to-End-System (end-to-end system) Systeme, in denen neuronale Modelle genutzt werden ohne dass Einzelschritte explizit, etwa mithilfe einer linguistischen Pipeline, modelliert werden. Dabei werden etwa Vorverarbeitungsschritte durch selbstständig gelernte, modellinterne, Repräsentationen ersetzt.

Entity Linking (entity linking) Der Prozess des Verbindens von ▸ **Eigennamen** mit externen Informationen bezüglich ihrer Identität. So wird z. B. die Erwähnung einer prominenten Person mit dem Eintrag zu dieser Person in einer Wissensbasis verbunden.

Escape-Sequenz (escape sequence) Eine Reihe von Symbolen die ausdrückt, dass ein oder mehrere darauf folgende Zeichen nicht als Daten im ursprünglichen ▸ **Alphabet** betrachtet werden sollen. So wird z. B. mit dem ▸ **regulären Ausdruck** „.“ nicht als die Suche nach einem Punkt sondern die nach einem beliebigen Zeichen aufgefasst, mittels eines Backslash als Escape-Sequenz kann mit „\.“ nach einem Punkt gesucht werden.

Evaluationsmaß (evaluation measure) Maße, um die Qualität von Ergebnissen, etwa Klassifikationsergebnisse eines maschinellen Lernsystems, zu bewerten; dabei findet ein Vergleich mit ▸ **Gold-Daten** statt. Je nach Beschaffenheit der Daten bzw. der Aufgabe eignen sich unterschiedliche Evaluationsmaße.

Evidenzbasierter Ansatz Auffassung von Wahrscheinlichkeit, bei dem anstelle von Häufigkeiten Wahrscheinlichkeitsverteilungen verwendet werden, in denen Vorwissen und a-priori-Annahmen explizit im Modell ausgedrückt werden. Syn.: Bayesscher Ansatz.

Expertensystem (expert system) Ein Expertensystem ist ein Computersystem, das auf einem speziellen Wissensgebiet die Kompetenz von menschlichen Experten nachbildet und als Beratungs- und Problemlösungssystem eingesetzt werden kann.

Explosion, kombinatorische (combinatorial explosion) Kombinatorische Explosion beschreibt die dramatische Vergrößerung des Möglichkeitenraumes einer Berechnung oder von Zuständen, etwa durch die Einführung einer weiteren Eingabedimension. Dies zieht oft nach sich, dass erschöpfende Suchansätze nicht mehr mit realistischem Speicher- und Zeitaufwand erfolgreich sind und andere Methoden, etwa ▸ **Heuristiken,** zum Einsatz kommen müssen.

Extension (extension) In der Referenzsemantik Bezeichnung für das von einem sprachlichen Ausdruck denotierte Objekt bzw. die denotierte Menge von Objekten.

Facettensuche oder Facettierte Suche (facetted search) beschreibt die Möglichkeit, in einer Information-Retrieval-Anwendung Suchergebnisse durch das Definieren von Filtern auf Metainformationen einzuschränken.

Fachausdruck (expression in technical language) Ein Fachausdruck (auch Fachbegriff) ist ein Wort bzw. eine Phrase, das nach einem vorgegebenen Kriterium für ein Fachgebiet charakteristisch ist. Syn.: Fachbegriff.

Fachbegriff (technical term) Siehe ▸ **Fachausdruck.**

Fachsprache (technical language) Eine Fachsprache ist eine Untermenge der von einem Sprachsystem erzeugbaren Strukturen, wie sie insbesondere in Fachtexten verwendet wird. Fachsprachen unterscheiden sich von der Allgemeinsprache durch messbare linguistische Abweichungen im Hinblick auf das Lexikon, die Syntax und die Semantik.

Fachterminologie (terminology) Die Fachterminologie ist das Begriffs- und Benennungssystem eines Fachgebietes, das alle Fachausdrücke umfasst, die allgemein üblich sind.

Fachtext (domain-specific text) Textsorte, die dazu dient, andere Fachleute desselben Faches oder Anwendungsbereiches zu informieren oder die Kommunikation mit Vertretern und Vertreterinnen anderer Disziplinen oder Laien über fachliche Sachverhalte zu ermöglichen. Fachtexte unterscheiden sich von allgemeinsprachlichen Texten durch das Vokabular sowie eine für den Anwendungsbereich charakteristische Syntax und Morphologie.

Feature (feature) Siehe ▸ **Merkmal.**

Feed-Forward-Netz (feed forward network) Ein neuronales Netzwerk, in welchem es im Gegensatz zum ▸ **rekurrenten Netzwerk** keine Verbindungen zu vorherigen Ebenen (oder, je nach Betrachtungsweise, zu vorherigen Zeitschritten) gibt. Das Netzwerk ist also aus aufeinanderfolgenden Schichten aufgebaut.

Fehlerfortpflanzung (error propagation) Fehler, die in einem Verarbeitungsschritt der linguistischen Pipeline auftreten, produzieren in dem nachfolgenden Verarbeitungsschritt möglicherweise weitere Fehler, welche ihrerseits wiederum Fehler im nächsten Verarbeitungsschritt verursachen können. Die Fehlerfortpflanzung ist ein typisches Problem starrer Pipelines, bei denen in den einzelnen Schritten auf Basis lokaler Informationen Entscheidungen getroffen werden, die sich erst später bzw. global als falsch herausstellen.

Fine-Tuning (fine tuning) bei neuronalen Netzwerken bezeichnet die Anpassung vortrainierter Embeddings an die vorliegende Klassifikationsaufgabe durch Weitertrainieren der Embeddings innerhalb der Netzwerkarchitektur.

Fleiss' Kappa (Fleiss' kappa) ist ein Maß für ▸ **Interrater-Reliabilität** zwischen zwei oder mehr Annotierenden.

Flexion (flexation) Bezeichnung für die Ableitung grammatischer Vollformen aus einem Stamm. Im Deutschen geschieht dies meist durch Anhängen von Flexionssuffixen sowie die Veränderung des Stamms durch Um- und Ablaute.

Flexiv (bound morpheme) Morphem, die lediglich eine grammatische und keine lexikalische Bedeutung hat. Flexive werden an Wortstämme angefügt (affigiert).

Frequentistischer Ansatz: (count-based approach) Auffassung von Wahrscheinlichkeit als Approximation der relativen Häufigkeit. Text-Mining-Verfahren, die auf diesem Ansatz aufbauen, untersuchen die Häufigkeit und statistische Verteilung von Sprachdaten.

Fügung, feste (set phrase) (auch: feste Wendung, Phraseologismus) Meist umgangssprachlich verwendete feste Wortverbindung, deren Bedeutung sich in der Regel nicht aus der Bedeutung der Einzelwörter erschließen lässt (z. B. *Spitze des Eisbergs,eingefleischter Junggeselle*).

F-Wert (F-score) Der F-Wert berechnet sich aus dem harmonischen Mittel zwischen ▸ **Precision** und ▸ **Recall** und wird oft als Maß zur Evaluation von überwachten Lernverfahren verwendet.

Gated Recurrent Units (GRU) Eine rekurrente neuronale Netzwerkarchitektur, Vereinfachung der Long Short-Term Memories, die mit weniger Rechenaufwand oft Ergebnisse von gleicher Qualität liefert.

Gazetteers (gazetteers) Listen bekannter Namen von Personen, Organisationen, Orten, Regionen etc. zur Verbesserung automatischer Eigennamenerkennungssysteme.

Gegensätze (opposites) Zwei Begriffe A und B sind Gegensätze in Bezug auf einen Oberbegriff, wenn beide einen gemeinsamen Oberbegriff C haben und die Schnittmenge zwischen der Extension von A und der Extension von B leer ist.

Gewichtungsfaktor (weight) Ein Gewichtungsfaktor regelt den Einfluss, den ein Term als Bestandteil einer Formel auf das Gesamtergebnis hat. Die Bedeutung einzelner Terme kann damit hervorgehoben, die Bedeutung von anderen hingegen abgewertet werden.

Gibbs Sampling (Gibbs sampling) Auswahl von Stichproben für das Topic Modelling, bei der für jedes Wort abhängig von allen anderen Zuordnungen seine Topic-Zuordnung berechnet wird. Die hochdimensionale Verteilung wird durch wiederholtes Ziehen von niedrigdimensionalen Variablen simuliert. Von einzelnen Wörtern ausgehend, werden so die Zuordnung von Wörtern zu Topics sowie die Zuordnung von Dokumenten zu Topics iterativ approximiert.

Gold-Daten (gold data) Daten inklusive manueller, korrekter Annotationen. Algorithmen des maschinellen Lernens werden auf derartigen Daten trainiert und

evaluiert. Meist werden Gold-Daten explizit durch menschliche ▸ **Annotatoren** erstellt.

GPU (Graphics Processing Unit) Grafikprozessoren, entwickelt für die effiziente Berechnung von Computergrafiken, sind aufgrund ihres effizienten Umgangs mit Matrizen aus Fließkommazahlen auch für das Berechnen neuronaler Netze geeignet.

Grammatik, formale (formal grammar) Eine Grammatik (formale) G definiert die Sätze einer Sprache. Sie wird gegeben durch ein Tupel G = (N,T,P,S). Hierbei ist N = Menge der nichtterminalen Symbole (syntaktische Kategorien), T = Menge der terminalen Symbole (Wörter der Sprache), S = Startsymbol und P = Menge der Produktionsregeln, die festlegt, wie sich vom Startsymbol über die syntaktischen Zwischenkategorien die Sätze der Sprache als Folge von Wörtern ableiten lassen.

Ground Truth (ground truth) Für die Evaluation von Algorithmen verifizierte Analyseergebnisse, die als Maßstab für die Bewertung der Qualität von Algorithmen dienen. Siehe auch Gold-Daten.

Grundform (base form) Ausgezeichnete Wortform als Bezeichner für ein Wort, beispielsweise Nominativ Singular für Nomina und Infinitiv für Verben. Syn.: Lemma.

Gültigkeit (validity) Eine XML-Datei heißt gültig, wenn sie wohlgeformt ist und den für diese Datei definierten XML-Regeln entspricht.

Hashverfahren (hashing) Funktion, die einer Zeichenkette einen Hashwert aus einem typischerweise großen Wertebereich zuordnet. Hashverfahren wie MD5 können so genutzt werden, um beispielsweise Sätzen mittels Hashwert eine Zahl als Identifikator zuzuordnen. Verschiedene Hashwerte stellen dann sicher, dass es sich um verschiedene Sätze handelt.

Häufigkeit, absolute (absolute frequency) Die absolute Häufigkeit eines Wortes w (Type) ist gleich der Anzahl der Vorkommen von w (Tokens) im Text.

Häufigkeit, relative (relative frequency) Die relative Häufigkeit eines Wortes w (Type) in einem Text ist gleich dem Quotienten aus der (absoluten) Häufigkeit dieses Wortes und der Gesamtzahl der Wörter (Tokens) im Text. Sie sollte für unterschiedliche Texte (der gleichen Sprache und ggf. aus dem gleichen Sachgebiet) ähnliche Werte annehmen.

Häufigkeitsklasse (frequency class) Eine Häufigkeitsklasse ist eine Einteilung der Wörter in Gruppen nach ihrer Frequenz im Korpus. Wörter der gleichen Häufigkeitsklasse treten im Korpus etwa gleich häufig auf. Oft werden die Klassen nach der Häufigkeit der darin enthaltenen Wörter geordnet und nummeriert, sodass die Nummer der Klasse eine Aussage über die Häufigkeit der zugeordneten Wörter zulässt.

Head-Modifier-Relation (head-modifier relation) Siehe ▸ **Dependenz.**

Heuristik (heuristics) Heuristiken sind vereinfachte Ansätze zur schnelleren bzw. einfacheren Lösung von Problemen. Sie können etwa die Exploration eines Problemraums unterstützen, sodass optimale Ergebnisse mit weniger Rechenaufwand berechnet werden. Heuristiken können ebenfalls angewandt werden um das optimale

Ergebnis möglichst gut zu approximieren oder ein korrektes Ergebnis in der Mehrheit der Fälle zu erreichen.

Hintergrundkorpus (background corpus) Hintergrundkorpora sind jene Korpora, die nicht spezifisch für eine Fragestellung oder Aufgabe gesammelt sind, und als solche auch keine Annotationen enthalten. Sie bestehen aus gesammelten Texten aus einer oder mehreren Domänen und sind z. B. für das Training von ▸ **Embeddings** geeignet.

HTML (HTML; hypertext markup language) Dokumentenauszeichnungssprache, die es mithilfe von HTML-Befehlen erlaubt, inhaltliche Kategorien von HTML-Dokumenten, z. B. Überschriften und Absätze, zu kennzeichnen. So ausgezeichnete Dokumente werden von Web-Browsern interpretiert und dargestellt. Die Dateiendung einer HTML-Datei lautet.html bzw.htm.

HTML-Befehl (HTML tag) Dient zur Auszeichnung von ▸ **HTML-Dokumenten** und besteht aus einer Anfangs- und einer Endemarkierung (tag), z. B. Dies ist ein Absatz . Einige Befehle benötigen keine Endemarkierung. In manchen Befehlen können zusätzlich Attribute mit Werten angegeben werden. Siehe ▸ **HTML.**

HTML-Dokument (HTML document) In ▸ **HTML** formatiertes Dokument, besteht aus Text und HTML-Befehlen, vornehmlich für die Darstellung von Webseiten. Dateiendung.html bzw. htm.

Hybrides System (hybrid system) Systeme, welche Ansätze aus verschiedenen Kategorien verbinden, werden als hybride Systeme bezeichnet. Im Kontext des maschinellen Lernens versteht man darunter Systeme, die sich in Teilen auf regelbasierte Ansätze stützen.

Hyperlink (hyperlink) Verweise auf andere Dokumente; in ▸ **Web-Browsern** meist farblich oder unterstrichen hervorgehoben; ein Mausklick auf einen Hyperlink bewirkt, dass zu dem Dokument, auf das verwiesen wird, verzweigt wird. Syn.: Link, Verweis, Referenz.

Hyperonym (hyperonym) Siehe ▸ **Oberbegriff.**

Hyperparameter (hyperparameter) Im Kontext des maschinellen Lernens sind Hyperparameter jene Parameter die nicht durch den Algorithmus approximiert oder berechnet werden, sondern a priori durch den Entwickler bzw. die Entwicklerin des Systems festgelegt werden. Bei einem neuronalen Netzwerk sind das etwa die ▸ **Netzwerktopologie,** die Lernrate und die Wahl des Trainingsalgorithmus.

Hypertext (hypertext) Text, der Sprungmarken bzw. Verweise (▸ **Hyperlinks**) auf andere Texte enthält. Hyperwürfel (hyper cube) Speicherung von Daten in mehrdimensionalen Strukturen. Jede Zelle innerhalb eines Würfels wird durch die Elemente aller Dimensionen bestimmt und kann direkt angesprochen werden (Data Warehouse).

Hyponym (hyponym) Siehe ▸ **Unterbegriff.**

Information (information) Bei Information handelt es sich um Daten, die in einem Kontext interpretiert werden und somit eine Bedeutung für den Besitzer oder Empfänger dieser Daten haben. Häufig liegen Informationen in wenig strukturierter Form als Textdokumente, Zeichnungen, Bilder etc. vor.

Information Retrieval (information retrieval) Die Auffindung (engl. retrieval) von Informationen im Sinne einer Inhaltsabfrage auf großen Dokumentenkollektionen.

Informationsextraktion (information extraction) Beschreibt im Kontext der Sprachverarbeitung den Prozess, strukturierte Informationen aus natürlicher Sprache zu gewinnen, etwa das Ergebnis eines Fußballspiels aus einem Spielbericht.

Instanz (instance, data point) Im maschinellen Lernen sind Instanzen einzelne Eingabe-Elemente auf denen operiert wird, das können z. B. komplette Bilder, eine Menge binärer Eigenschaften oder, im Text Mining, einzelne Worte, Dokumente oder Sätze sein.

Interrater-Reliabilität (Inter-Annotator-Agreement, IAA) Ein Maß für die Übereinstimmung von manuell Annotierenden bezüglich einer Annotationsaufgabe. Sie wird als obere Schranke für die Genauigkeit von auf diesen Daten trainierten Lernverfahren angesehen.

Inverse Dokumentfrequenz, IDF (inverse document frequency) Maß im Information Retrieval zur Bestimmung des Anteils von Dokumenten, in denen ein Token vorkommt, im Verhältnis zur Gesamtzahl aller Dokumente eines Korpus. Die IDF wächst, wenn ein Token nur in wenigen Dokumenten auftritt.

Inverse Liste (inverted index) Liste, in der zu jedem im Korpus vorkommenden Wort alle Positionen des Auftretens im Korpus verzeichnet sind. Sie wird während der Korpuserstellung erzeugt und ermöglicht die schnelle Suche nach Wörtern im Korpus.

Join (join) Operation auf relationalen Datenbanken, mit deren Hilfe Datensätze aus mehreren Tabellen aufgrund einer gemeinsamen Eigenschaft (der Join-Bedingung) zusammengeführt werden können.

JSON (JSON) (JavaScript Object Notation) Einfach lesbares und kompaktes Datenformat, benutzt zum Datenaustausch oder auch der Datenspeicherung.

Kategorie, syntaktische (syntactic category) Eine syntaktische Kategorie ist das Potential eines Wortes, mit anderen Wörtern derselben Sprache syntaktisch wohlgeformte Kombinationen (Phrasen) zu bilden. Zu den syntaktischen Kategorien gehören traditionell: Nomen, Verb, Adjektiv, Artikel, Präynonyposition und Konjunktion. Empirisch bilden Wörter derselben Kategorie paradigmatische Distributionsklassen. Syn.: Wortart.

Klassifikation (classification) Die Klassifikation in der automatischen Sprachverarbeitung ist das Versehen von sprachlichen Objekten (Wörter, Sätze, Absätze, Texte) mit Markierungen aus einer definierten Menge von Symbolen.

Klassifikator (classifier) Ein Klassifikator ist die Funktion, welche eine ▸ **Klassifikation** durchführt.

Kohyponym (co-hyponyme) Hat ein Begriff mehrere Unterbegriffe, stehen diese in der Relation der Kohyponymie und sind Kohyponyme.

Komplementärbegriff (complementary term) Zwei Begriffe A und B sind Komplementärbegriffe, wenn sie Gegensätze sind und das Komplement von A äquivalent zur Extension von B ist.

Komponente (component) Komponenten sind Lose gekoppelte Programme, die einzelne Verarbeitungsschritte innerhalb einer komplexen Pipeline erledigen. Typische Komponenten einer linguistischen Pipeline sind beispielsweise Textsegmentierung in Sätze, ▸ **Tokenisierung** und ▸ **Part-of-speech-Tagging.**

Komposition (composition) Morphologischer Prozess, bei dem durch das Aneinanderfügen von zwei Stämmen bzw. freien Morphemen neue Wörter gebildet werden.

Kompositionsprinzip (principle of composition) auch Frege-Prinzip: Annahme, dass die Bedeutung eines Satzes eine Funktion der Bedeutung seiner Teilausdrücke ist.

Kompositum (compound) Ein durch Komposition aus wenigstens zwei Stämmen gebildetes Wort.

Konfidenz (confidence) Ein statistisches Maß, welches angibt, wie wahrscheinlich es ist das eine gegebene Beobachtung nicht durch reinen Zufall sondern durch eine Systematik in den Eingabedaten entstanden ist.

Konjugation (conjugation) Flexion (Beugung) von Verben. Verändert wird dabei die Person, Numerus, Tempus, Modus und das sog. Genus verbi (Aktiv oder Passiv).

Konkatenation (concatenation) Begriff aus der theoretischen Informatik, mit dem das Hintereinanderschreiben zweier Zeichen bzw. Zeichenketten bezeichnet wird. Die Konkatenation zweier Wörter „x“ und „y“ ist das Wort, das sich durch Hintereinanderschreiben der beiden Wörter ergibt, also „xy“.

Konstituente (constituent) In der Linguistik Bezeichnung für die unmittelbaren Bausteine von Sätzen auf der syntaktischen Ebene, also Wörter und Kombinationen von Wörtern. Empirisch lassen sich Konstituenten durch eine Reihe von Tests bestimmen, welche hinreichende Bedingungen für das Vorliegen einer Konstituente definieren.

Konstituentenparsing (constituency parsing) Verarbeitungsschritt zum automatischen Zuweisen der ▸ **Konstituenten-Struktur** eines Satzes.

Konstituenten-Struktur (constituent structure) Syntaktische Phrasenstruktur, die mit einer kontextfreien Chomsky-Grammatik dargestellt werden kann, der Parse-Baum eines Satzes.

Kontext, globaler (global context) Der globale Kontext eines Wortes w enthält alle Wörter, die mit w statistisch signifikant häufig gemeinsam auftreten (bezogen auf ein Signifikanzmaß und einen Schwellenwert).

Kontext, lokaler (local context) Der lokale Kontext eines Wortes w ist die Menge der Wörter, mit denen w zusammen in einem Satz auftritt.

Kontextualisierte Embeddings (contextualized embeddings) Embeddings, deren Wert sich durch den Kontext der betrachteten Instanz definiert. Im Gegensatz zu statischen Embeddings gibt es nicht für jedes Element im Vokabular eine feste Representation, stattdessen wird das kontextualisierte Embedding in der Regel aus den statischen Embeddings der ▸ **Instanzen** im lokalen Kontext berechnet.

Kontextvolatilität (context volatility) Maß zur Quantifizierung von Kontextänderungen eines Wortes über die Zeit. Die Kontextvolatilität gibt insbesondere bei niederfrequenten Wörtern frühzeitig Aufschluss über semantische Verschiebungen im Verwendungskontext eines Ausgangswortes.

Konverse (converses) Inhaltliche Relation zwischen begrifflichen Gegensätzen in Bezug auf einen gemeinsamen Oberbegriff.

Kookkurrenz, signifikante (significant co-occurrence) Als signifikante Kookkurrenz bezeichnet man das statistisch auffällige gemeinsame Auftreten zweier Wörter in einem Satz oder Textfenster (bezogen auf ein Signifikanzmaß und einen Schwellenwert).

Kookkurrenzanalyse (co-occurrence analysis) Verfahren des Text Mining, bei dem für ein Korpus signifikante Kookkurrenzen berechnet und visualisiert werden.

Koreferenz (coreference) Textspannen in einem Dokument stehen in Koreferenz, wenn sie auf dieselbe Entität verweisen.

Koreferenzauflösung (coreference resolution) Der Erkennungsprozess jener Textspannen in einem Dokument, welche sich auf dieselbe Entität beziehen. Dabei werden alle Referenzen (etwa Pronomen aber auch explizite, namentliche Nennungen), die sich auf die gleiche Entität beziehen, in einer Koreferenzkette zusammengefasst.

Korpus (text corpus) Ein Korpus ist eine Sammlung von Texten.

Korpusvergleich (corpus comparison) Verfahren zur Ermittlung von statistisch signifikanten Unterschieden in der Verwendung von Vokabularen im Vergleich zwischen einem Analysekorpus und dem Referenzkorpus.

Kosinusähnlichkeit (cosine similarity) Siehe ▸ **Kosinus-Maß.**

Kosinus-Maß (cosine measure) Das Kosinus-Maß ist ein Maß für die Ähnlichkeit von Vektoren im Information Retrieval. Berechnet wird dabei der Kosinus des Winkels zwischen den beiden Vektoren zur Modellierung der Ähnlichkeit der repräsentieren Instanzen. Sind diese normiert, ist der Wert gleich dem Skalarprodukt der beiden Vektoren. Da die Komponenten der Vektoren sämtlich nichtnegativ sind, ist der Winkel zwischen den Vektoren höchstens 90° und das Kosinus-Maß liegt zwischen 0 (maximal unähnlich) und 1 (identisch).

Kreuzvalidierung (cross validation) Schema zur Aufteilung der Trainingsdaten zur Evaluation überwachter Lernverfahren. Hier wird die Trainingsmenge in n (z. B. 5 oder 10) gleich große Mengen aufgeteilt, wovon jeweils reihum eine Menge als Development-Menge und die anderen als Trainingsmenge verwendet werden. Insbesondere bei dünner Datenlage empfohlen.

Latin (Latin) Genormter 8-Bit-Zeichensatz (256 Positionen), der den ASCII-Code um 128 Positionen erweitert. Latin-1 deckt westeuropäische Sprachen ab. Siehe auch ▸ **Unicode.**

Lemma (lemma) Siehe ▸ **Grundform.**

Lernverfahren, maschinelles (machine learning algorithm) Ein Verfahren des überwachten Lernens, welches auf Basis einer ▸ **Trainingsmenge** eine Funktion approximiert, die jeden (Eingabe) Featurewert zu einem (Ausgabe) Label überführt.

Lernverfahren, überwachtes (supervised learning) Siehe ▸ **Verfahren, überwachtes.**

Lernverfahren, unüberwachtes (unsupervised learning) Siehe ▸ **Verfahren, unüberwachtes.**

Levenshtein-Distanz, auch: Editierdistanz (Levenshtein distance; edit distance) Abstand zwischen zwei Zeichenketten gemessen in der minimalen Anzahl der benötigten zeichenbasierten Einfüge-, Ersetzungs- oder Löschoperationen um die eine Zeichenkette in die andere abzuändern.

Lexikographie (lexicography) Das Fachgebiet beschäftigt sich mit der Erstellung von Wörterbüchern. Dazu zählen Konzeption, Stichwortauswahl und das Verfassen der Wörterbucheinträge.

Lexikon (lexicon) In der Linguistik und Automatischen Sprachverarbeitung die Liste der dem System bekannten Wörter einer Sprache.

Linguistische Ebene (level of linguistics) Bezeichnung für die Elemente einer hierarchischen Untergliederung von der Verarbeitung von Sprache. Üblicherweise werden die linguistischen Ebenen unterteilt in Phonologie, Phonetik, Morphologie, Syntax, Semantik und Pragmatik.

Linguistisches Wissen (linguistic knowledge) Beschreibung der für eine Sprache typischen Gesetzmäßigkeiten mit Bezug auf die linguistischen Ebenen. Für die automatische Verarbeitung natürlicher Sprache muss dieses Wissen in geeigneter Form zur Verfügung stehen.

Log-Likelihood-Maß (log likelihood measure) Statistisches Plausibilitätsmaß, welches auf der Basis beobachteter Daten die plausibelsten Parameter einer Wahrscheinlichkeit ermittelt. Im Text Mining wird das Log-Likelihood-Maß verwendet, um die statistische Signifikanz des gemeinsamen Auftretens von Wortpaaren zu ermitteln, siehe ▸ **Kookkurrenzen.**

Log-Likelihood-Ratio (log likelihood ratio) Quotient aus zwei Likelihood-Funktionen, im Korpusvergleich Quotient aus der Likelihood-Funktion der wahrscheinlichen Anzahl eines Types im Analysekorpus und der als Nullhypothese aus dem Referenzkorpus abgeleiteten erwarteten Anzahl.

Long Short-Term Memory (LSTM) Eine rekurrente neuronale Netzwerkarchitektur, in der ein expliziter Zustand vorgehalten wird. Der Zustand wird, mittels gelernter Gewichte, basierend auf der Eingabe sowie dem Zustand selbst in jedem Zeitschritt verändert. Diese explizite Modellierung erlaubt insbesondere das Modellieren temporal weit reichender Abhängigkeiten, Informationen über diese werden in einfacheren rekurrenten Architekturen oft nicht erfolgreich gelernt.

Makro-Sicht (macro view) Sicht auf sehr umfangreiche Textkorpora mit dem Ziel, diese inhaltlich zu strukturieren, z. B. durch das Clustering ähnlicher Texte oder der Identifizierung von Themen und deren Verteilung im Textkorpus.

Markdown (Markdown) Beliebte Markup-Sprache, welche das Hinzufügen von Formatierungsinformationen zu Textdateien erlaubt, ohne dabei die Lesbarkeit der Textdatei zu opfern. So werden etwa Auflistungen einfach dadurch notiert, dass aufeinanderfolgende Zeilen jeweils mit einem „*“ oder „-“ starten.

Markov-Modell (Markov model) Ein diskretes Markov-Modell erster Ordnung beschreibt eine Folge von Zuständen (hier: Tokens oder Zeichen) so, dass ein bestimmter Zustand nur von dem Zustand unmittelbar davor abhängt, nicht aber

von weiter zurückliegenden Zuständen. Markov-Modelle höherer Ordnung berücksichtigen entsprechend mehr Zustände aus der Vergangenheit. Markov-Modelle dienen zur Beschreibung einer Sequenz von Werten einer Zufallsvariable, die nicht unabhängig voneinander sind. Insbesondere hängen die Werte von den anderen in der Sequenz aufgetretenen Werten ab. Markov-Modelle lassen sich auch durch endliche Automaten mit Übergangswahrscheinlichkeiten beschreiben.

Markup (markup) Markup ist eine Auszeichnung, eine Markierung, die an bestimmten Stellen in einem Dokument eingefügt wird, um die Form der Darstellung oder die Struktur der Dokumente zu beschreiben. Die einzelnen voneinander getrennten Markup-Elemente werden als Tags bezeichnet.

Maschinelles Lernen (machine learning) Verfahren der Künstlichen Intelligenz für das Erkennen von Mustern und Gesetzmäßigkeiten in Daten und für die Generierung von Problemlösungen. Üblicherweise wird zwischen überwachtem Lernen (supervised learning) und unüberwachtem Lernen (unsupervised learning), bestärkendem (reinforcement learning) und aktivem Lernen (active learning) unterschieden.

Mehrdeutigkeit (ambiguity) Mehrdeutigkeit liegt vor, wenn es für ein Wort bzw. eine Wortfolge mehr als eine semantische Interpretation gibt. Bei der Mehrdeutigkeit ist zwischen einer lexikalischen und strukturellen Mehrdeutigkeit zu unterscheiden. Ein Wort ist lexikalisch mehrdeutig, wenn es mehr als eine Bedeutung oder Funktion hat. Ein Wort bzw. eine Wortfolge ist strukturell mehrdeutig, wenn es mehr als eine sinnvolle Zerlegung gibt.

Mehrworteinheit (multi-word unit) Eine aus mehreren Wörtern bestehende Zeichenkette, die in der Text Mining Anwendung als Einheit behandelt werden soll. Für die Extraktion von Mehrworteinheiten lassen sich meist vorher definierte ▸ **POS-Muster** nutzen.

Merkmal (feature) Repräsentation sprachlicher Einheiten als numerische oder nominale Werte zur Verwendung in statistischen oder neuronalen Modellen. Für jedes Feature wird eine Featurefunktion erstellt, welche Featurewerte für alle zu charakterisierenden sprachlichen Einheiten berechnen kann. Featurefunktionen können hierbei auf einer Vielzahl von Eigenschaften des sprachlichen Materials definiert werden. Denkbare Merkmale für Wörter sind beispielsweise Groß- und Kleinschreibung als binärer Wert, relative Häufigkeit des Wortes als reelle Zahl oder das ▸ Word **Embedding** als reellwertiger Vektor. Während solche Merkmale im maschinellen Lernen klassischerweise auf Basis der Aufgabenstellung händisch entwickelt wurden, bietet das End-to-End-Lernen die Möglichkeit, geeignete Repräsentationen automatisch im Modell erlernen zu lassen.

Metadaten (metadata) Metadaten für Dokumente sind Werte von qualifizierten Textattributen, welche Texte in technischer, organisatorischer und inhaltlicher Hinsicht beschreiben. Ein wichtiger Standard ist **Dublin Core** der Dublin Core Metadata Initiative (DCMI).

Methode, neuronale (neural method) Siehe ▸ **Verfahren, neuronale.**

Methode, regelbasierte (rule-based method) Siehe ▸ **Verfahren, regelbasierte.**

Methode, statistische (statistical method) Siehe ▸ **Verfahren, statistische.**

Mikro-Sicht (micro view) Sicht auf sehr umfangreiche Textkorpora mit dem Ziel, aus einem Text auffällige Informationen zu extrahieren und zusammen mit den Informationen aus anderen Texten zu verbinden.

Mismatch (mismatch) Ein Mismatch ist eine Nicht-Übereinstimmung von Zeichenketten. Die Stelle des ersten Mismatch zwischen zwei Wörtern ist die Position desjenigen Zeichens, vom Wortanfang aus betrachtet, in dem sich die beiden Wörter das erste Mal voneinander unterscheiden.

Morphem (morpheme) Ein Morphem ist in einer natürlichen Sprache die kleinste bedeutungstragende Einheit von Zeichenketten. Man unterscheidet zwischen freien Morphemen wie z. B. *Wort* oder *rot,* die ohne Affixe im Text auftreten können, und gebundenen Morphemen, die nur mit ergänzenden Affixen auftreten können. Zu den gebundenen Morphemen zählen alle grammatischen Morpheme und Derivative.

Morphologie (morphology) Teilbereich der Linguistik, der beschreibt, welche Morpheme in einer Sprache vorkommen und wie diese zu Wortformen kombiniert werden.

Multitask-Lernen (multitask learning) ist das Lernen verschiedener Zielfunktionen unter Zuhilfenahme von verschiedenen Trainingsmengen mit demselben Klassifizierungsalgorithmus, typischerweise einem neuronalen Netzwerk. Falls die Zielfunktionen korreliert sind, kann dies zu Verbesserungen der Genauigkeit führen.

Mundart (dialect) (auch: Dialekt) Regionale Sprachvarietät mit abweichendem Wortschatz, aber auch teilweise veränderter Grammatik und Aussprache.

Muster (pattern) Sich wiederholende Struktur in Texten, die in Form von Regeln beschrieben werden kann.

Mutual Information (mutual information) auch: Transinformation oder gegenseitige Information: Informationstheoretisches Maß für die Stärke des statistischen Zusammenhangs zweier Zufallsgrößen.

Nachbarschaftskookkurrenz (neighbour co-occurrence) Die Nachbarschaftskookkurrenz ist das statistisch auffällige gemeinsame Auftreten zweier Wörter als unmittelbare Nachbarn. Besteht zwischen zwei Wörtern eine Nachbarschaftskookkurrenz, wird das vorangehende (erste) Wort als linker Nachbar, das nachfolgende (zweite) Wort als rechter Nachbar bezeichnet.

Nachricht (message) Eine Nachricht ist eine nach vorher festgelegten Regeln zusammengestellte, endliche Folge von Zeichen und Zuständen, die eine Information vermittelt.

Named Entities (named entities) Siehe ▸ **Eigenname.**

Named Entity Recognition (named entity recognition) Siehe ▸ **Eigennamenerkennung.**

Natural Language Processing (NLP) Regelbasierte, statistische und neuronale Verfahren für die automatische Verarbeitung gesprochener und geschriebener Sprache auf

der Grundlage informatischer und linguistischer Modelle. Syn.: Automatische Sprachverarbeitung, Sprachtechnologie.

Netz, lexikalisch-semantisches (lexical semantic network) Siehe ▸ **Wortnetz.**

Netzwerktopologie (network topology) Beschreibt den Aufbau eines neuronalen Netzwerks, also Anzahl und Verbindungen der einzelnen Neuronen. In einem ▸ **Feed-Forward Netz** ist die Topologie typischerweise durch die Größen der Ein- und Ausgabevektoren sowie durch die Anzahl der Schichten des Netzwerks und ihre jeweiligen Größen gegeben.

Neuron, künstliches (artificial neuron) Die elementaren Bausteine eines künstlichen neuronalen Netzwerks sind durch biologische Neuronen inspiriert. Ein künstliches Neuron transformiert, analog zum ▸ **Perzeptron** einen reellwertigen Eingabe Vektor zu einem skalaren Ausgabe Wert, wobei eine Vielzahl von Aktivierungsfunktionen zum Einsatz kommen kann. Ein künstliches neuronales Netzwerk besteht in der Regel aus mehreren Schichten mit jeweils mehreren Neuronen.

News Monitoring (news monitoring) Sammlung und Analyse großer Nachrichtenkollektionen mit Text Mining. Als Datengrundlage dienen Texte oder in Text gewandelte gesprochene Sprache aus dem Internet oder den sozialen Medien, deren Publikationsdatum eindeutig erkennbar ist, und deren Schwerpunkt auf Nachrichten aus allen Lebensbereichen, insbesondere Politik, Wirtschaft, Unternehmen, Wissenschaft und Technologie liegt.

N-Gramm (n-gram) Ein n-Gramm besteht aus n aufeinanderfolgenden Wörtern bzw. Buchstaben.

N-Gramm-Modell (n-gram model) Einfaches Sprachmodell auf Basis von ▸ **n-Grammen.** Die Wahrscheinlichkeit eines unbekannten nachfolgenden Wortes beruht auf der relativen Häufigkeit des n-Gramms bestehend aus diesem Wort und den n-1 vorangegangen Wörtern.

Nicht-transparent (opaque) Eigenschaft von Komposita, dass sich die Bedeutung des zusammengesetzten Wortes nicht aus der Bedeutung der Bestandteile erschließen lässt. Beispiele sind *Mauerblümchen, Elfenbein, Friedhof, Fuchsschwanz.*

Nominalphrase (noun phrase) Grammatikalische Kategorie, welche die Funktion eines Nomens im Satz ausfüllt. Kann aus einem Pronomen oder einem Nomen mit modifizierenden Adjektiven und ggf. Artikel bestehen.

Oberbegriff (subordinate term) (auch Hyperonym) Ein nicht-leerer Begriff B ist ein nicht-leerer Oberbegriff eines Begriffs A, wenn die Extension von A eine echte Teilmenge der Extension von B ist.

Office-Formate (office file formats) Sammelbegriff für Formate gängiger Microsoft Office-Applikationen wie Word, Excel oder Powerpoint. Derartige Formate sind in der Regel XML-basiert und unterstützen einen großen Funktionalitätsumfang von der relativ einfachen Formatierung des enthalten Textes bis zur komplexen Einbindung von Multimediaelementen und interaktiven Komponenten.

One-Hot-Kodierung (one-hot encoding) Schema zur Vektorrepräsentation kategorische Variablen, dabei steht jedes Element des Vektors für einen der möglichen

Ausprägungen. Der Wert „1“ an n-ter Stelle besagt dass es sich um diese n-te Ausprägung des Werts handelt wobei alle anderen Werte stets „0“ sind. In der Sprachverarbeitung wird dieses Schema etwa für die Kodierung von Worten in einem gegeben Vokabular verwendet, ein solcher Vektor ist dann z. B. bei einem Vokabular der Länge 1000 auch 1000 Elemente lang.

Ontologie (ontology) Eine Ontologie ist eine explizite, meist axiomatische und standardisierte Formalisierung eines gemeinsamen Verständnisses von Schlüsselbegriffen und deren Relationen bezüglich eines Faches oder einer Anwendung. Oft beinhaltet eine Ontologie eine hierarchisch organisierte Taxonomie.

Optical Character Recognition (OCR) Erkennung von Zeilen, Wörtern und Buchstaben in einem Bild (optischer Zeichenerkennung). Im weiteren Sinne umfasst OCR mehr als die bloße Klassifikation von Pixelmustern als Buchstaben, vielmehr ist das Ziel eine möglichst umfassende und fehlerfreie Konvertierung von Bilddateien in Dokumente in einem maschinenlesbaren Format für die weitere Be- und Verarbeitung.

Out of vocabulary (OOV) Problem des ungesehenen Vokabulars, d. h. Wortformen, die nicht im Vokabular oder in den Trainingsdaten enthalten sind, aber in der Anwendung auf Testdaten auftreten. Ein unangepasstes Sprachmodell weist solchen Wortformen eine Wahrscheinlichkeit von Null zu, ein wortlistenbasiertes System kann diese Wörter nicht verarbeiten.

Overfitting (overfitting) Überanpassung eines statistischen bzw. neuronalen Modells an die Trainingsdaten. Während das Modell die Trainingsdaten exakt oder zumindest zu genau reproduziert, leidet die Fähigkeit, auf ungesehenen Daten zu generalisieren. Overfitting tritt insbesondere dann auf, wenn die Anzahl der Trainingsdaten zu gering für die Anzahl der zu lernenden Parameter des Modells ist.

Paradigmatisch (paradigmatic) In der Tradition des linguistischen Strukturalismus Bezeichnung für das Auftreten von Zeichen bzw. Zeichenketten in ähnlichen Kontexten.

Parallelisierung (parallelisation) Siehe ▸ **Skalierung, horizontale.**

Parsen, semantisches (semantic parsing) Bezeichnet einen Verarbeitungsschritt von natürlich-sprachlichen Sätzen, in dem die Bedeutungsstruktur der Prädikate (meist Verben) und deren Argumenten expliziert wird.

Parts Of Speech (POS) Wortarten wie Nomen, Verb, Adjektiv, Pronomen, Präposition, Adverb, Konjunktion, Partizip und Artikel.

Part-of-Speech-Tagging (part-of-speech tagging) oder Wortartentagging Als Part-of-Speech-Tagging (POS-Tagging) bezeichnet man das Zuordnen syntaktischer Kategorien zu Tokens. Jedes (erkannte) Token in einem Text wird um die Angabe über die Wortart ergänzt, in der das Token im gegebenen lokalen Kontext verwendet wird.

PDF (Portable Document Format) Dokumentenformat, welches ursprünglich für die einheitliche Darstellung des gleichen Dokuments auf unterschiedlichen Plattformen und Endgeräten entwickelt wurde. PDFs können sowohl Grafiken in Raster- oder Vektorform, als auch Texte, Links und interaktive Elemente wie Formulare enthalten.

Perplexität (perplexity) auch: Kreuzentropie: Ein informationstheoretisches Maß für die Qualität eines Modells bezüglich einer (empirischen) Wahrscheinlichkeitsverteilung.

Perzeptron (Perceptron) Einfaches Element eines neuronalen Netzwerks mit einer beliebigen, festen Anzahl reeller Zahlen als Eingabe. Die Ausgabe wird aus dem Eingabevektor x mittels Gewichtsvektor w und Bias-Vektor b wie folgt berechnet: x * w + b. Das Perzeptron kann mittels einer binären Aktivierungsfunktion zur binären Klassifikation benutzt werden.

Phrase (phrase) Als Phrase bezeichnet man die in einer natürlichen Sprache nach syntaktischen Regeln in sich abgeschlossenen syntaktischen Einheiten unterhalb der Satzebene.

Phrasen-Struktur-Grammatik (phrase structure grammar) Von Chomsky vorgeschlagene kontextfreie Ersetzungsgrammatik für eine natürliche Sprache. Alle Wörter werden als terminale Symbole behandelt, die syntaktischen Kategorien und Phrasen als nichtterminale Symbole und die syntaktischen Regeln werden in Form von Ersetzungsregeln angegeben.

Pointer-Generator-Networks (pointer generator networks) Neuronale Netzwerkarchitektur zur Generierung einer Ausgabe-Textsequenz basierend auf einer Eingabe-Textsequenz. Dabei kommt ein **Attention-Mechanismus** zum Einsatz und mittels der namensgebenden Pointer können Worte aus der Eingabesequenz direkt übernommen werden.

Pointwise Mutual Information (PMI) PMI ist ein Maß aus der Informationstheorie für die Stärke der Assoziation zweier Ereignisse, z. B. dem gemeinsamen Auftreten von Wörtern. Es quantifiziert den Unterschied zwischen der tatsächlichen und der erwarteten gemeinsamen Auftretenshäufigkeit unter Annahme statistischer Unabhängigkeit.

Polarität (polarity) Bewertungsdimension in der Sentimentanalyse wie *positiv – negativ.*

Porter-Stemmer (Porter stemming algorithm) Verfahren für die Stammformreduktion, bei dem nach bestimmten Regeln die Suffixe eines Wortes gelöscht werden.

POS-Muster (POS pattern) Abfolge von Wörtern einer bestimmten Wortart, wie z. B. „Nomen Nomen", „Adjektiv Nomen" oder „Adjektiv Adjektiv Nomen".

Precision (Precision) Precision (Genauigkeitsgrad) ist im ▸ **Information Retrieval** ein Effektivitätsmaß zur Bewertung von Suchmaschinen. Die Precision für eine Anfrage ergibt sich als Quotient der Anzahl der gefundenen relevanten Dokumente durch die Anzahl der gefundenen Dokumente. Precision wird stets gemeinsam mit Recall benutzt. Syn.: Genauigkeitsgrad.

Probabilistische kontextfreie Grammatik (probabilistic context free grammar) Kontextfreie Grammatik, in der jede Regel mit einer Wahrscheinlichkeit versehen ist, wobei die Summe der Wahrscheinlichkeiten aller Regeln mit demselben Symbol auf der linken Seite 1 betragen muss. Die Wahrscheinlichkeit einer Zer-

legung ist das Produkt der Wahrscheinlichkeiten der Regeln, die während des Parsens angewandt werden.

Quantor (quantifier) Bei ▸ **regulären Ausdrücken** Operator, welche die Anzahl Wiederholungen von Ausdrücken angibt, gern in Kombination mit ▸ **Wildcard.**

Rang (rank) Der Rang eines Wortes w in der häufigkeitssortierten Wortliste ist die Listenposition von w.

Rang, größter/höchster (highest rank) Die letzte vergebene Rangnummer in der häufigkeitssortierten Wortliste; die Listenposition des letzten Wortes.

Ranking (ranking) Rankings sind Sortierungen einzelnen Elemente bezüglich einer bestimmten Eigenschaft, so werden z. B. im Information Retrieval Dokumente bezüglich ihrer Relevanz zu einer Suchanfrage sortiert.

Recall (recall) Der Recall (Erschöpfungsgrad) ist im Information Retrieval ein Effektivitätsmaß zur Bewertung von Suchmaschinen. Der Recall einer Anfrage ergibt sich als Quotient der Anzahl der gefundenen relevanten Dokumente durch die Anzahl der relevanten Dokumente. Recall wird stets gemeinsam mit Precision (Genauigkeitsgrad) benutzt. Syn.: Erschöpfungsgrad.

Rechtschreibreform (spelling reform) Veränderung des amtlichen Regelwerks zur Rechtschreibung zu einem gewissen Zeitpunkt. Häufig werden dadurch vorhandene Trends im tatsächlichen Sprachwandel offiziell anerkannt. Dies betrifft oft Getrennt- und Zusammenschreibung, Groß-/Kleinschreibung, veränderte Verwendung von Buchstaben (wie ß) oder Buchstabenverbindungen (wie ph) sowie die Zeichensetzung.

Redundanz (redundancy) Redundanz in der Informationsübertragung beschreibt mehrfach vorhandene Informationseinheiten. Diese können ohne Informationsverlust weggelassen werden, erhöhen aber die Übertragungssicherheit und können zur Fehlerkorrektur benutzt werden. Unter Redundanz in Texten versteht man das wiederholte Auftreten der gleichen Information, oft in unmittelbarer Nachbarschaft. Redundanz kann syntaktischer Art (wie Kongruenz) oder inhaltlicher Art (z. B. Wiederholung von Wörtern oder Wortteilen, Pleonasmus) sein. Das wiederholte Auftreten solcher Informationen ermöglicht deren automatische Erkennung.

Referenzkorpus (reference corpus) Das Referenzkorpus besteht in der Regel aus einer Anzahl von Texten, die den allgemeinen Sprachgebrauch wiedergeben. Zum allgemeinen Sprachgebrauch können ebenfalls einige, allgemein bekannte Fachausdrücke aus den verschiedenen Fachsprachen gehören. Diese Fachausdrücke treten jedoch im Referenzkorpus relativ selten auf – verglichen mit einem Korpus der entsprechenden Fachsprache. Syn.: Vergleichskorpus.

Regelbasierter Ansatz (role-based approach) Siehe ▸ **Verfahren, regelbasierte.**

Regulärer Ausdruck (regular expression) Reguläre Ausdrücke beschreiben eine eigene, vollständig definierte Kunstsprache, welche in vielen Varianten in praktisch jeder Programmiersprache, vielen Texteditoren, vor allem aber beim Text Mining oder allgemein in der automatischen Sprachverarbeitung zum Einsatz kommt. In der Praxis

ermöglicht diese Sprache, Ausdrücke zu formulieren bzw. Muster festzulegen, nach denen gesucht werden soll.

Rekurrentes Netzwerk (recurrent network) Neuronale Netzwerkarchitektur zur Anwendung auf sequentiellen Daten, dabei wird die Eingabe als Zeitreihe aufgefasst und die Berechnung eines jeden Zeitschritts erhält Informationen (etwa die Ausgabe oder den Wert nach Anwendung einer inneren Schicht) aus vorangegangenen Schritten als zusätzliche Eingabe.

Rekursion (recursion) Siehe ▸ **Selbstaufrufender Verweis.**

Relation, logische (logical relation) Logische Relationen sind semantische Relationen, die logische Folgerungen unterstützen. Hierzu zählen insbesondere die Oberbegriffs- und Unterbegriffsbeziehung, Synonyme, Gegensätze, Antonyme, Komplementärbegriffe und Konverse. Der Begriff der logischen Relationen kann mengentheoretisch präzisiert werden.

Relation, paradigmatische (paradigmatic relation) Paradigmatische Relation ist in der Tradition des linguistischen Strukturalismus die Bezeichnung für das Auftreten zweier Wörter in ähnlichen Kontexten.

Relation, semantische (semantic relation) Semantische Relationen sind Sinnrelationen zwischen Wörtern, wie sie insbesondere in Bedeutungswörterbüchern oder Thesauren verwendet werden, also z. B. Synonymie (Bedeutungsgleichheit), Unterbegriffs- oder Oberbegriffsbeziehungen. Die Elemente von Thesauren sind allgemein beschrieben in der ISO Norm ISO 25964.

Relation, syntagmatische (syntagmatic relation) Syntagmatische Relation ist in der Tradition des linguistischen Strukturalismus die Bezeichnung für das gemeinsame Auftreten zweier Wörter in einem Satz oder Textfenster.

ROC-Kurve (receiver operating characteristic curve) auch Grenzwertoptimierungskurve ROC-Kurven können zur Evaluation der Klassifikatorqualität eingesetzt werden, indem das Verhältnis von Precision und Recall für verschiedene Konfidenzwerte dargestellt wird; die Fläche unter der Kurve charakterisiert dann die Qualität des Klassifikators unabhängig von einer konkreten Konfidenzschwelle.

Satz (sentence) In sich abgeschlossene, sinntragende und wohlgeformte sprachliche Einheit. Aus syntaktischer Sicht ist ein Satz eine Konkatenation von Phrasen.

Satz, nicht-wohlgeformter (ill-formed sentence) In zu analysierenden Texten findet man auch Sätze, die von zweifelhafter Qualität sind: Sie können unvollständig sein, syntaktische Fehler enthalten, nicht den orthographischen Regeln entsprechen und inhaltlich unverständlich oder sinnlos sein.

Satzkookkurrenz (sentence-level co-occurrence) Als Satzkookkurrenz bezeichnet man das statistisch auffällige gemeinsame Auftreten zweier Wörter in einem Satz.

Satzkookkurrenzmatrix (matrix of sentence-level co-occurrences) Die Satzkookkurrenzmatix ist eine quadratische, symmetrische Matrix, deren Zellen die Stärke der ▸ **Satzkookkurrenz** zwischen den mit den jeweiligen Zeilen und Spalten assoziierten Wörtern enthält.

Satzkorpus (corpus of sentences) Ein Korpus aus einzelnen Sätzen mit oder ohne Annotationen und Metadaten. Im Gegensatz zum ▸ **Dokumentkorpus** sind hier ganze Texte nicht rekonstruierbar durch Umordnung und dem Weglassen von Sätzen, auch im Rahmen der Qualitätssicherung.

Schwache Signale (weak signals) Konzept aus der Wirtschaftswissenschaft, mit dem frühzeitige Anzeichen für erwartbare Diskontinuitäten wirtschaftlicher Prozesse bezeichnet werden.

Selbstaufrufender Verweis (self-referring link) Siehe ▸ **Rekursion.**

Semantik (semantics) Gegenstand der Semantik ist die Bedeutung Zeichen und Zeichenketten auf allen linguistischen Ebenen. Allerdings ist der Begriff Bedeutung unscharf und bedarf einer genaueren Bestimmung durch eine semantische Theorie. Die wesentlichen Paradigmen einer semantischen Theorie sind die prozedurale, referentielle und strukturalistische Semantik.

Semantik, distributionelle (distributional semantics) Forschungsgebiet zur Entwicklung und Erforschung von datengetriebenenen Repräsentationen für die Bedeutung von sprachlichen Elementen wie Wörtern, Sätzen und Texten auf Basis der ▸ **distibutionellen Hypothese.**

Sentimentanalyse (sentiment analysis, auch Stimmungsanalyse) Anwendung des Text Mining mit dem Ziel, die Einstellung eines Autors oder einer Autorin zu einem Thema zu bestimmen. Unter Einstellung wird dabei eine Bewertung, ein emotionaler Zustand oder ein Appell verstanden.

Sentimentausdruck (sentiment expression) Ausdruck, mit dem eine subjektive Befindlichkeit oder eine Bewertung von Objekten, Ereignissen oder Meinungen (sentiment targets) ausgedrückt werden. Typische Bewertungsdimensionen sind dabei ▸ **Polaritäten** wie *positiv – negativ* für die Polaritätsanalyse oder *subjektiv-objektiv* für die Subjektivitätsanalyse.

Sentiment-Orientierung (sentiment orientation) Bezeichnung für die Bewertungsdimension von Wörtern, die anstelle einer bipolaren Bewertung wie *gut* oder *schlecht,positiv* oder *negativ* auch Zwischenwerte ausdrücken können (etwa auf einer Polaritätsskala mit einem Wert zwischen +1 für positiv und −1 für negativ mit 0 als neutral).

Seq2Seq (sequence to sequence) Systemarchitektur, in der basierend auf einer Eingabesequenz eine Ausgabesequenz generiert wird, etwa für die maschinelle Übersetzung. Dabei wird ein rekurrentes Netzwerk verwendet, um zuerst die Eingabesequenz einzulesen, und dann basierend auf dem internen Zustand und dem jeweils zuvor ausgegeben Element eine Ausgabesequenz zu generieren.

Sequenzklassifikation (sequence classification/sequence tagging) auch Sequenzmarkierung: Vorgang des automatischen Zuweisens von Labels für Instanzen innerhalb einer Sequenz, z. B. beim ▸ **Wortartentagging** oder bei der ▸ **Eigennamenerkennung.**

Signifikanzmaß (measure of significance) Das Signifikanzmaß ist ein Maß für den Nachweis statistisch verlässlicher (signifikanter) Unterschiede in empirischen Untersuchungen.

Skalierung (scaling) Leistungsanpassung der Software und Hardware um auch große Problemgrößen bzw. Datenmengen in adäquater Zeit bearbeiten zu können. Es wird dabei zwischen ▸ **vertikaler und** ▸ **horizontaler Skalierung** unterschieden.

Skalierung, horizontale (horizontal scaling) Horizontale Skalierung ermöglicht die Leistungssteigerung durch Hinzufügen zusätzlicher, gleichartiger Rechner bzw. Knoten. Dazu muss die verwendete Software für die parallelisierte Verarbeitung optimiert sein. Durch den Anteil sequentieller Verarbeitungsschritte sind der Leistungssteigerung jedoch Grenzen gesetzt.

Skalierung, vertikale (vertical scaling) Vertikale Skalierung beruht auf der Leistungssteigerung eines einzelnen Systems durch den Einsatz leistungsfähigerer Hardware. Es ist keine Anpassung der verwendeten Software erforderlich. Der Leistungssteigerung sind jedoch Grenzen gesetzt, wenn bereits die leistungsfähigste Hardware verwendet wird.

Skip-Gram-Modell (Skip-gram model) Variante der neuronale Netzwerkarchitektur word2vec zum Erlernen von ▸ Word **Embeddings** in Form von ▸ **dichtbesetzten Vektoren.** Die Trainingsaufgabe des Netzwerkes ist das Vorhersagen des Kontextes zu einem Wort, d. h. die Wörter links und rechts des aktuellen Wortes, mittels eines ▸ **Sliding-Window**-Verfahrens. Die Reihenfolge der Wörter im Kontext ist dabei relevant.

Sliding-Window (sliding window) Ein Verfahren für die elementweise Verarbeitung einer Sequenz, wobei jeweils eine feste Anzahl an Elementen links und/oder rechts des aktuellen Elements als zusätzlicher Kontext für die Verarbeitung genutzt wird.

Spam (spam) Unerwünschte Informationen, in der Regel in Form von E-Mails oder Kurznachrichten, aber z. B. auch in Form ganzen Websites. Oft beinhaltet Spam unerwünschte Werbung oder Betrugsversuche.

Spannenannotation (span annotation) Annotation, die sich auf ein oder mehrere zusammenhängende Wörter bezieht. So können z. B. mehrere aufeinanderfolgende Worte als ein ▸ **Chunk** vom Typ Nominalphrase markiert werden.

Sprachdynamik (language dynamics) Die Veränderung einer natürlichen Sprache über die Zeit aufgrund gesellschaftlicher und sprachlicher Anpassungsprozesse.

Sprache, formale (formal language) In der theoretischen Informatik Bezeichnung für eine Menge von Zeichenketten, die eine bestimmte Teilmenge der Menge aller möglichen Wörter umfasst, die nach einer vorgegebenen Konstruktionsvorschrift über einem Alphabet erzeugt werden können.

Sprache, kontextfreie (context-free language) Bezeichnung für eine ▸ **Chomsky-Grammatik,** in der in jeder Regel der Grammatik auf der linken Seite genau ein nicht-terminales Symbol steht und auf der rechten Seite eine beliebige nicht-leere Folge von terminalen und nicht-terminalen Symbolen aus dem gesamten Vokabular.

Sprache, natürliche (natural language) Von Menschen zu allen Zwecken der Kommunikation verwendete gesprochene und geschriebene Sprache. Natürliche Sprachen unterscheiden sich von formalen Sprachen u. a. dadurch, dass sie sprachstatistischen Gesetzmäßigkeiten wie dem ▸ **Zipfschen Gesetz** folgen. Außerdem ist für eine natürliche Sprache nicht immer eindeutig definiert, welche Kombinationen von Wörter zulässig sind, und die Kombinationen von Wörter können strukturell mehrdeutig sein.

Spracherkennung (speech recognition) Automatische Erkennung gesprochener Sprache. Die Wandlung gesprochener Sprache in Text wird auch als speech-to-text bezeichnet.

Sprachklassifikation (language identification) Bestimmung der natürlichen Sprache, in der ein Text oder Textabschnitt verfasst ist.

Sprachmodell (language model) Sprachmodell ist in der statistischen Sprachverarbeitung die Bezeichnung für die Gesamtheit der konkreten Wahrscheinlichkeiten sprachlicher Ereignisse.

Sprachstatistik (statistics of language) Sprachstatistik ist die Bezeichnung für empirisch validierte statistische Gesetzmäßigkeiten natürlicher Sprache. Hierzu zählen u. a. die „Zipfschen Gesetze“ und die Small-Worlds-Eigenschaft semantischer Netze.

Softmax-Funktion (softmax function) Mathematische Funktion, um alle Komponenten eines reellwertigen Vektors in den Wertebereich von 0 bis 1 zu transformieren, wobei die Summe aller Komponenten 1 ergibt, sodass das Ergebnis als Wahrscheinlichkeitsverteilung angesehen werden kann.

Stamm (stem) Siehe ▸ **Basismorphem.**

Stammformreduktion (stemming) Zuordnung verschiedener Vollformen auf denselben Stamm.

Steuerzeichen (control character) Zeichen eines Zeichensatzes für die Steuerung eines Ausgabegeräts zur Darstellung von Zeichen des Zeichensatzes, also keine darstellbaren Zeichen wie z. B. Buchstaben, Ziffern oder Satzzeichen.

Statistischer Ansatz (statistical method) Siehe ▸ **Verfahren, statistische.**

Statistischer Hypothesentest (statistical hypothesis testing) Siehe ▸ **t-Test.**

Stimulus-Response-Experiment (stimulus response experiment) Experiment zur Erforschung sprachlicher Assoziationen. Nach dem Hören eines Wortes (als Stimulus) soll die Versuchsperson schnell das erste Wort (als Response) nennen, welches ihm/ihr einfällt.

Stoppwort (stop word) Ein Stoppwort ist ein Wort, das von der Analyse ausgeschlossen werden soll. Dazu gehören im Allgemeinen Wörter aus den geschlossenen Wortklassen wie Artikel, Konjunktion und Präposition.

Strukturalismus, linguistischer (linguistic structuralism) Eine auf den Linguisten Ferdinand de Saussure zurückgehende Richtung der Linguistik, nach der Zeichenketten einer linguistischen Ebene immer nur in Bezug auf jeweils andere Zeichenketten dieser Ebene einen Sinn bzw. eine Funktion haben.

Stylometrie (stylometrie) In der Literaturwissenschaft und den Digital Humanities Bezeichnung für die Analyse stilistischer Merkmale von Texten.

Subwort-Tokenisierung (sub-word tokenisation) Tokenisierungsstrategie, bei der nicht nur ganze Wörter als Tokens verstanden werden, sondern auch einzelne Wortbestandteile. Dieser Ansatz kann mit einem relativ kleinen Vokabular sehr viele Worte abbilden, wobei seltene Wort in Subworteinheiten zerlegt werden. Dies umgeht vor allem Probleme mit ▸ **OOV** Wörtern.

Synonym (total synonym) Zwei Wörter sind synonym, wenn sie in fast allen Kontexten ausgetauscht werden können, ohne die Bedeutung des Textes zu verändern. Im engeren Sinne sind zwei Wörter synonym, wenn sie dieselbe Extension haben.

Synonym, schwaches (partial synonym) Da echte Synonymie sehr selten anzutreffen ist, spricht man im Falle Austauschbarkeit mit geringer Bedeutungsänderung von schwacher Synonymie.

Syntagmatisch (syntagmatic) In der Tradition des linguistischen Strukturalismus Bezeichnung für das gemeinsame Auftreten von Zeichen bzw. Zeichenketten.

Syntax (syntax) Gegenstand der Syntax sind syntaktische Repräsentationen für Wörter einer Sprache und deren Kombination zu Sätzen.

Tagset (tag set) Liste von zulässigen kategorischen Werten für die ▸ **Annotation** von ▸ **Tokens,** Textspannen (▸ **Spannenannotation**) oder ▸ **Dependenzrelationen.** Für jede Aufgabe bzw. Art der Annotation wird typischerweise ein eigenes Tagset verwendet, z. B. für ▸ **Part-of-Speech-Tagging** oder ▸ **Named Entity Recognition.**

Tanimoto-Ähnlichkeit (Tanimoto index) Die Tanimoto-Ähnlichkeit ist ein Maß für die Ähnlichkeit von Vektoren mit Einträgen 0 oder 1. Berechnet wird der Quotient aus der Anzahl der übereinstimmenden 1-Werte und der Anzahl der Positionen, die in mindestens einem der Vektoren den Wert 1 haben.

Taxonomie (taxonomy) Hierarchische Strukturierung von Fachbegriffen.

TEI (Text Encoding Initiative) Dokumentenformat auf der Basis von XML, entwickelt von der gleichnamigen Organisation. Eignet sich zur Auszeichnung komplexer Strukturen (z. B. in Wörterbüchern) oder editorischer Informationen.

Template-basierte Informationsextraktion (template-based information extraction) Eine Art Informationsextraktion, bei der die für das Ausfüllen vorgegebener Schemata (Templates) notwendigen Informationen automatisch aus dem Text extrahiert werden, im Gegensatz zur Open Information Extraction**.**

Tensor Processing Unit (TPU) Spezialhardware zur Beschleunigung der Berechnung von Operationen auf Tensoren, Matrizen und Vektoren. Diese Hardware wird zum Training und zur Inferenz großer neuronaler Netzwerke verwendet.

Term (term) Ein Term ist im Information Retrieval ein Wort in einem Text bzw. einer Kollektion von Texten.

Term, diskriminierender (discriminating term) Ein diskriminierender Term ist ein Wort, die nach einem vorgegebenen Kriterium für einen Text bzw. eine Kollektion von Texten charakteristisch ist und deswegen gut zur Unterscheidung von Texten eingesetzt werden kann. Bei der Differenzanalyse sind diskriminierende Terme

diejenigen Wörter, die im Fachtext wesentlich häufiger auftreten als in einem allgemeinsprachlichen Referenzkorpus.

Term-Dokument-Matrix (term document matrix) Eine Matrix für die Wortfrequenzen pro Dokument. Dabei repräsentiert jede Zeile ein Dokument und jede Spalte ein Wort des Vokabulars, sodass jedes Element der Matrix die Anzahl der Vorkommnisse eines spezifischen Worts in einem spezifischen Dokument beschreibt.

Term-Frequenz/Inverse-Dokument-Frequenz (term frequency – inverse document frequency; tf·idf) Maß im Information Retrieval zur Ermittlung statistisch signifikanter Unterschiede in der Verwendung von Token beim Vergleich eines Analyse- mit einem Referenzkorpus. Eine hohe Termfrequenz im Analysekorpus deutet darauf hin, dass das Token im Analysekorpus wichtig ist, die inverse Dokumentfrequenz bemisst die statistische Signifikanz in der unterschiedlichen Verwendung dieses Token in Bezug auf das Referenzkorpus.

Terminologie (terminology) Das Begriffs- und Benennungssystem eines Fachgebietes, das alle Fachausdrücke umfasst, die allgemein üblich sind. Die Terminologie eines Fachgebietes ist eine Sammlung von Fachausdrücken, die meist in Form eines Thesaurus beschrieben wird.

Terminologielehre (terminology theory) Beschreibung der Gesetzmäßigkeiten und die Erarbeitung der Terminologie von Fachsprachen.

Testmenge (test set) Bestandteil des Datensatzes aus Instanzen und Labels, bezüglich dessen ein Modell final evaluiert wird. Während der Entwicklung darf die Testmenge nicht zum Einsatz kommen. Siehe auch Trainingsmenge und Development-Menge.

Text (text) Ein Text ist strukturiert und besteht aus Sätzen, die nach Art und Zweck des Textes zu größeren Einheiten zusammengefasst werden (Absätze, Abschnitte, Kapitel). Ein Text besteht aus Wörtern, die ihrerseits aus den Buchstaben eines Alphabets bestehen. Text repräsentiert Wissen und stellt insofern eine wesentliche Grundlage der Wissensverarbeitung dar.

Textabdeckung (text coverage) Die Textabdeckung eines Textes durch eine vorgegebene Menge von Wörtern beschreibt den Anteil des Textes, welche diese Menge am Gesamttext hat. So haben in typischen deutschsprachigen Texten die häufigsten 10.000 Wörter eine Textabdeckung von rund 80 %.

Textkorpus (text corpus) Siehe ▸ **Korpus.**

Text Mining (Text Mining) Computergestützte Verfahren für die semantische Analyse von Texten, welche die automatische bzw. semi-automatische Strukturierung von Texten, insbesondere sehr großen Mengen von Texten, unterstützen.

Thesaurus (thesaurus) Sammlung von Fachausdrücken eines Fachgebietes unter Nennung des Sachgebietes, der Synonyme, Übersetzungen und Übersetzungsvarianten, Oberbegriffe, Kohyponyme, Antonyme und Aufzählung von Beispielen.

Token (token) Im Text Mining Vorkommen einer nach den ▸ **Tokenregeln** definierten Zeichenkette in einem Text.

Tokenisierung (tokenisation) Prozess der Anwendung von ▸ **Tokenregeln** auf einen Text.

Tokenregeln (token rules) Regeln für die Zerlegung eines Textes in Wörter, Satzzeichen und Sonderzeichen.

Topic-Modell (topic model) Evidenzbasiertes, nicht-deterministisches Verfahren für die Generierung artifizieller Topics als einer Wahrscheinlichkeitsverteilung über das vorhandene Vokabular eines Textes. Jedes Dokument wird als eine Wahrscheinlichkeitsverteilung über eben diese Topics dargestellt, welche wiederum angibt, wie wahrscheinlich ein Topic für das aktuelle Dokument ist.

Top-Level-Domain (TLD) Meist länderspezifischer Teil einer Internetadresse, bestehend aus meist zwei oder drei Buchstaben nach dem letzten Punkt. Die Einschränkung auf die Top-Level-Domains de, at und ch ist eine sinnvolle Crawlingstrategie, um hauptsächlich deutschsprachige Webseiten zu besuchen.

Trainingsmenge (training set) Eine Menge von ▸ **Instanzen,** inklusive der jeweils erwünschten Ausgaben, die als Eingabe für einen maschinellen Lernalgorithmus verwendet werden und auf deren Basis ein Modell gelernt wird. Siehe auch Development-Menge und Testmenge.

Transduktor (transductor) Endlicher Automat, welcher eine Eingabe auf einem Eingabeband in eine Ausgabe auf einem Ausgabeband überführt. Also solche eignen sich Transduktoren zur Formalisierung von Regelsystemen der morphologischen Transformation, etwa um ein Wort auf seinen Stamm zurückzuführen.

Transformation (transformation) In der Transformationsgrammatik die Bezeichnung für die Ableitung von beobachteten Wortfolgen (sog. Oberflächenstrukturen) durch Veränderungen und Verschiebungen von Phrasen aus sog. Tiefenstrukturen, welche mithilfe einer kontextfreien PSG erzeugt worden sind.

Transformer-Architektur/Transformer-Networks (transformer architecture/ netwiorks) Neuronale Netzwerkarchitektur, welche auf Basis eines Encoders und eines Decoders sowohl zur Textgenerierung als auch für Klassifikationsaufgaben verwendet werden kann. Im Gegensatz zu anderen Architekturen wie ▸ **Seq2Seq** haben Transformer-Networks grundsätzlich eine feste Eingabegröße. So können keine Sequenzen beliebiger Länge behandelt werden, allerdings birgt die Architektur Vorteile bei effizienter Berechnung und der Modellierung weitreichender Abhängigkeiten.

Translation Memory (translation memory) System für die Erzeugung von Übersetzungshypothesen unter Einsatz maschinellen Lernens auf der Grundlage von bereits erstellten Übersetzungen eines Textes.

Trainingskorpus (training corpus) Ein Trainingskorpus ist eine Textmenge, die analysiert und ausgewertet wird, um die für einen Algorithmus notwendigen Daten und Parameter zu ermitteln.

TREC (Text REtrieval Conference) Seit 1992 Serie von Konferenzen im ▸ **Information Retrieva**l und ▸ **Natural Language Processing** zur Evaluation von Verfahren für vorgegebene Retrieval- oder Verarbeitungsaufgaben wie beispielsweise das Erkennen von neuen Nachrichten im ▸ **News Monitoring.**

Trend (trend) Bezeichnung für eine sich langfristig abzeichnende Entwicklung. In der Statistik wird ein Trend als Zeitreihe dargestellt. Er beschreibt die quantitative Veränderung von Ereignissen, erklärt jedoch nicht die Hintergründe von Veränderungen.

Trie (Trie) Ein Trie (digitaler Mehrwegebaum) ist ein m-Wege-Baum, wobei m die Anzahl der Zeichen des Alphabets ist. Jeder Knoten im Baum repräsentiert einen Buchstaben. Die Buchstaben auf dem Weg von der Wurzel zu einem Knoten stehen für den Anfang eines Wortes. Die Nachfolgerknoten stehen für die entsprechenden Fortsetzungsmöglichkeiten. Syn.: digitaler Mehrwegebaum.

Tri-Gramm (tri-gram) Ein Tri-Gramm ist ein spezieller Typ von ▸ **n-Grammen,** das aus drei aufeinander folgenden Wörtern besteht. Siehe n-Gramm und Bi-Gramm.

t-Test (t-test) Statistischer Hypothesentest zur Überprüfung, ob zwei Stichproben sich statistisch signifikant unterscheiden.

Type (type) Äquivalenzklasse derjenigen Zeichenketten, die entsprechend den Tokenregeln für eine Anwendung als gleich betrachtet werden; Eintrag im Vokabular.

Übersetzung, automatische (machine translation) Übersetzung, die durch ein Computerprogramm ohne menschliche Hilfe erstellt wurde; das Forschungsgebiet der automatischen Übersetzung.

Unicode (unicode) Genormter 16-Bit-Zeichensatz (65.469 Positionen), der die Schriftzeichen aller Verkehrssprachen der Welt aufnehmen soll.

Unigramm (uni-gram) Ein Unigramm ist ein spezieller Typ von n-Grammen, das aus einem Wort oder einem Buchstaben besteht.

Unterbegriff (subordinate term) Ein nicht-leerer Begriff A ist ein nicht-leerer Unterbegriff eines Begriffs B, wenn die Extension von A eine echte Teilmenge der Extension von B ist. Syn. Hyponym.

Urheberrecht (copyright law) Eigentümer bzw. Eigentümerin aller Verwertungsrechte eines Textes ist sein Urheber bzw. seine Urheberin. Das Urheberrecht umfasst die gesetzlichen Regelungen zur Verwertung und zum Schutz der Urheberrechte.

URL (uniform resource locator; uniform resource locator) Im Web verwendete standardisierte Darstellung von Internetadressen; Aufbau: protokoll://domain-Name/Dokumentpfad.

UTF-8 Beliebtes Kodierungsschema für den ▸ **Unicode** Standard. Hier werden durch die variable Länge ▸ **ASCII** Zeichen in nur 8 Bit kodiert, wohingegen Symbole aus anderen Sprachen (etwa deutsche Umlaute, aber z. B. auch kyrillische Zeichen und Emoji) in 16 oder mehr Bit kodiert werden.

Valenzstruktur (valence structure) Fähigkeit eines Wortes oder einer Wortgruppe, andere Wörter oder Wortgruppen als Ergänzungen zu fordern oder zu erlauben. Speziell können einzelne Verben Akkusativobjekte und/oder Dativobjekte fordern *(schenken),* erlauben *(lernen)* oder auch nicht erlauben *(regnen).*

Vektor, dichtbesetzter (dense vector) Reellwertiger Vektor, in dem die meisten Zahlenwerte ungleich null sind. Solche Vektoren ermöglichen effiziente mathematische Operationen wie Addition oder Multiplikation und bilden damit die Basis für meisten Verarbeitungsschritte in neuronalen Netzwerken.

Vektorraummodell (vector space model) Das Vektorraummodell ist ein Modell zur Repräsentation von Informationen in einem hochdimensionalen metrischen Vektorraum. Es wird im Text Mining für die Repräsentation von Wörtern und von Dokumenten eingesetzt; auch ▸ **Embeddings** werden im Vektorraum repräsentiert.

Verteilung von Wörtern im Text (word distribution) Unter Verteilung der Wörter im Text versteht man die Häufigkeit der einzelnen Wörter und Wortkombinationen.

Verfahren, evidenzbasiertes (evidence-based method) Statistisches Verfahren fürs Text Mining auf Grundlage der Bayesschen Statistik, bei dem Wahrscheinlichkeitsverteilungen verwendet werden, in denen Vorwissen und A-Priori-Annahmen explizit im Modell ausgedrückt werden, beispielsweise Topic-Modelle. Syn.: Bayessche Verfahren.

Verfahren, frequentistisches (count-based method) Statistisches Verfahren fürs Text Mining, bei dem die Häufigkeit und statistische Verteilung von Sprachdaten untersucht wird. Hierzu gehören bspw. Korpusvergleiche zur Ermittlung von statistisch signifikanten Unterschieden in der Verwendung von Vokabularen sowie Kookkurrenzanalysen.

Verfahren, neuronales (neural method) Text-Mining-Verfahren unter Verwendung von reellwertigen Vektoren für die Repräsentation von Sprachdaten und neuronalen Netzen, bei denen Feature-Repräsentationen für die Lösung einer Aufgabe nicht vorgegeben, sondern vom neuronalen Netz selbst gelernt werden, um eine optimale Lösung zu finden.

Verfahren, regelbasiertes (rule-based method) Text-Mining-Verfahren zur Erkennung von Mustern von Wörtern in Texten. Die Muster werden in Form von Regeln definiert, welche es ermöglichen, Wörter und Wortfolgen in Texten zu finden. Hierfür werden eigene, vollständig definierte Kunstsprachen wie z. B. reguläre Ausdrücke definiert.

Verfahren, statistisches (statistical method) Ausnutzung sprachstatistischer Gesetzmäßigkeiten für das Text Mining. Abhängig davon, wie der Begriff der Wahrscheinlichkeit gefasst wird, kann zwischen frequentistischen und evidenzbasierten Ansätzen unterschieden werden.

Verfahren, überwachtes (supervised method) maschinelles Lernverfahren, welches auf einem Trainingsdatensatz trainiert wird und eine vorgegebene Klassifizierung lernt.

Verfahren, unüberwachtes (unsupervised method) maschinelles Lernverfahren, welches auf nicht annotierten Datensätzen operiert, Ergebnis ist typischerweise eine ▸ **Cluster-Analyse.**

Verkettung (concatenation) Siehe ▸ **Konkatenation.**

Viterbi-Algorithmus (Viterbi algorithm) ist ein effizienter Algorithmus für das Bestimmen der wahrscheinlichsten Sequenz von Labels und wird z. B. bei ▸ **Hidden-Markov-Modellen** und ▸ **Conditional Random Fields** für die ▸ **Sequenzklassifikation** eingesetzt.

Vokabular (vocabulary) Das Vokabular eines Textes besteht aus allen Types, die in einem Text auftreten. Tritt ein Type mehrfach im Text auf, fließt er trotzdem nur einmal in das Vokabular ein. Der Umfang des Vokabulars entspricht der Anzahl der in einem Text vorkommenden Types.

Vokabular, Umfang (size of vocabulary) Der Umfang des Vokabulars ist gleich der Anzahl der in einem Text vorkommenden Types.

Vollform (full form) Das Vorkommen eines flektierten Wortes im Text, vgl. ▸ **Wortform.**

Vorverarbeitung (preprocessing) Arbeitsschritt im Text Mining, bei dem Textdaten inkrementell angereichert werden, um sie für Text-Mining-Aufgaben vorzubereiten. Dies fängt mit dem Einlesen der Daten und deren Überführung in eine geeignete interne Repräsentation an und umfasst Arbeitsschritte wie das Erkennen der Sprache, die Textsegmentierung und verschiedene Bereinigungsschritte.

Wahrscheinlichkeit (probability) Die Wahrscheinlichkeit eines Ereignisses ist nach frequentistischer Auffassung die relative Häufigkeit, mit der es in einer großen Anzahl gleicher, wiederholter Experimente auftritt. Die relative Häufigkeit eines Wortes in einem Text ist gleich dem Quotienten aus der (absoluten) Häufigkeit dieses Wortes und der Gesamtzahl der Wörter im Text. Betrachtet man immer längere, gleichartige Texte, so nähert sich diese relative Häufigkeit einem konstanten Wert, der Wahrscheinlichkeit dieses Wortes.

Wahrscheinlichkeit, bedingte (conditional probability) Die bedingte Wahrscheinlichkeit P(A|B) gibt die Wahrscheinlichkeit des Ereignisses A an, unter der Voraussetzung, dass das Ereignis B bereits eingetreten ist.

Web (web; World Wide Web) Informationssystem im Internet, das auf der Hypertext-Technik basiert; ermöglicht außerdem den Zugriff auf die anderen Internet-Dienste. Der Zugang zum Web erfolgt über **Web-Browser.** Syn.: W3, WWW.

Web-Browser (web browser) Software, über die Benutzer die Dienstleistungen des Internets, insbesondere des Webs, in Anspruch nehmen können. Durch Angabe der URL wird das Computersystem, das die jeweilige Dienstleistung anbietet, eindeutig adressiert. Syn: Browser.

Webcrawler Siehe ▸ **Crawler.**

Webseite (web page) Ein Internet-Dokument mit Text- und multimodalen Inhalten und ▸ **Hyperlinks** zur Navigation.

Whitespace (white space) Sammelbegriff für verschiedene unsichtbare ▸ **Zeichen** (z. B. Leerzeichen, Tabulator, Zeilenumbruch), die nur als Leerraum dargestellt werden und typischerweise zur Trennung von ▸ **Wörtern(Tokens)** oder Formatierung von Texten verwendet werden.

Wildcard (wildcard) In regulären Ausdrücken und in Anfragesprachen sind Wildcards Zeichen, welche stellvertretend für andere Zeichen stehen.

Wissen (knowledge) Wissen ist die meist auf Erfahrung beruhende und objektiv nachprüfbare Kenntnis von Fakten und Zusammenhängen eines Weltausschnitts, die

Personen zur Lösung von Problemen einsetzen. Wissen ermöglicht die Vernetzung von Informationen.

Wissensmanagement (knowledge management) Wissensmanagement ist eine Organisationsform, die das Wissen der in einem Unternehmen beschäftigten Personen, das für einen Erfolg des Unternehmens relevant ist, erfasst, strukturiert und zum Nutzen des Unternehmens einsetzt.

Wohlgeformtheit (well-formedness) Eine ▸ **XML**-Datei heißt wohlgeformt, wenn sie entsprechend den Regeln von ▸ **XML** aufgebaut ist.

Wort (word) Ein Wort im engeren Sinne ist eine Äquivalenzklasse von Wortformen (z. B. alle flektierten Formen des Wortes *sprechen*). Im weiteren Sinne kann Wort für ▸ **Token,** ▸ **Type,** ▸ **Term,** ▸ **Wortform** und ▸ **Grundform** verwendet werden.

Wortart (part of speech) Siehe ▸ **Parts of Speech.**

Wortartentagging (part of speech tagging) Siehe ▸ **Part-of-Speech-Tagging.**

Wortbedeutungsdisambiguierung (word sense disambiguation) Bezeichnet das Auflösen lexikalischer Mehrdeutigkeiten von Wörtern im Kontext; eine der Bedeutungen wird explizit für das Wortvorkommen zugewiesen.

Wortform (word form) Eine Wortform ist die flektierte Erscheinungsform eines Wortes, wie sie in einem syntaktischen Kontext vorkommt.

Wortklasse, geschlossene (closed word class) Eine geschlossene Wortklasse ist eine Klasse von Wörtern, die nur in sehr begrenztem Maße Veränderungen unterliegt. Bei Wörtern aus geschlossenen Wortklassen steht ihre grammatische Bedeutung im Vordergrund. Sie werden auch als Funktionswörter bezeichnet. Geschlossene Wortklassen: Artikel, Hilfsverb, Interjektion, Konjunktion, Numeral, Präposition, Pronomen.

Wortklasse, offene (open word class) Eine offenen Wortklasse ist eine Klasse von Wörtern, die Veränderungen durch die morphologischen Prozesse Flexion, Derivation und Komposition unterliegt. Offene Wortklassen: Adjektiv, Nomen, Verb, Adverb.

Wort-Kontext-Matrix (word-context matrix) Anordnung von ▸ **Wörtern** (▸ **Types**) und ihren Kontexten in Form einer Matrix, aus der ersichtlich ist, in welchen Kontexten ein Wort vorkommt bzw. welche Kontexte welche Wörter enthalten.

Wortnetz (word net) Ein Wortnetz ist ein Graph mit Wörtern als Knoten. Die Kanten beschreiben Beziehungen zwischen den entsprechenden Wörtern und können manuell oder automatisch erzeugt worden sein. Im Falle von signifikanten Kookkurrenzen beschreiben die Kanten häufig syntagmatische und paradigmatische Zusammenhänge.

Wort-n-Gramm (token n-gram) Bezeichnung für n aufeinanderfolgende Wörter. Siehe ▸ **n-Gramm.**

XML (Extensible Markup Language) Sprache, mit der sich Auszeichnungssprachen definieren lassen. XML ist eine verkürzte Version der Standard Generalized Markup Language (SGML). Die Version 1.0 von XML wurde im November 1998 vom WWW Consortium (W3C) verabschiedet.

Zeichen (character) Ein Zeichen ist ein Element eines endlichen geordneten Zeichenvorrats.

Zeichenkette (string of character) Eine Zeichenkette entsteht durch die Aneinanderfügung (Konkatenation) von Zeichen.

Zeta (Zeta) Verfahren in der Stylometrie, welches auf dem ▸ **t-Test** aufbaut und diesen verfeinert. Das Grundprinzip dieses Maßes ist es, vor der Berechnung die Texte in kleinere Segmente aufzuteilen, wobei die Segmentlänge ein wichtiger Parameter ist. Dann wird für jedes Merkmal der Anteil der Segmente erhoben, in denen das Merkmal mindestens einmal vorkommt (die „document proportion"). Von diesem Anteil in der untersuchten Gruppe wird der entsprechende Anteil in der Vergleichsgruppe subtrahiert, woraus sich der Zeta-Wert ergibt.

Zielfunktion (loss function) Die Zielfunktion gibt in neuronalen Netzen die Abweichung vom gewünschten Ergebnis an, das kann etwa durch die Euklidische Distanz passieren. Die Zielfunktion wird im Rahmen von Gradientenverfahren als Basis verwendet um, mit Hilfe von ▸ **Backpropagation,** die Gewichte des Netzes anzupassen.

Zipfsches Gesetz (Zipf's law) Bezeichnung für den von George Kingsley Zipf formulierten Zusammenhang: Wenn Wörter aus einem Korpus ihrer Häufigkeit nach absteigend dass für ein geordnet sind, ergibt das Produkt aus dem Rang eines Wortes (innerhalb der häufigkeitssortierten Liste) mit seiner Häufigkeit für alle Wörter in etwa denselben Wert. Die häufigsten Wörter, also die Wörter, die in der häufigkeitssortierten Liste den niedrigsten Rang haben, sind meist sehr kurze Funktionswörter/▸ **Stoppwörter.**

Zufallsstichprobe, geschichtete (stratified sampling) Im Gegensatz zur reinen Zufallsstichprobe stellt die geschichtete Zufallsstichprobe sicher, dass die Verteilung der Kategorien der in der Probe enthaltenen Instanzen die Verteilung in der Grundgesamtheit approximiert. Wird oft zur Erstellung von ▸ **Trainings-** und ▸ **Validationsmengen** bei der Klassifikation eingesetzt.