Glossar

Adjazenzmatrix: (Pfister 2004, S. 44f.) ist eine 2-dimensionale Tabelle, deren Kopfzeile und 1. Spalte identisch sind und bestimmte Beziehungen zwischen je einem Objekt in der Kopfzeile und in der 1. Spalte anzeigen.
Beispiel 1: Bei einem Drama stehen die handelnden Figuren in der Kopfzeile und 1. Spalte. Ein Eintrag aij (i-te Zeile, j-te Spalte) für ein Paar zweier Figuren kann anzeigen, wie oft die beiden Figuren in den Szenen miteinander auf der Bühne stehen.
Beispiel 2: In der Stilometrie stehen die Texte in der Kopfzeile und 1. Spalte. Ein Eintrag aij für ein Paar zweier Texte kann anzeigen, wie ähnlich diese Texte zueinander sind. Die Ähnlichkeit wird mit einem Distanzmaß gemessen, siehe den Artikel zur Stilometrie.

Algorithmus: Umsetzung einer Methode in einem Programm, z.B. programmtechnisches Verfahren zur Aufspaltung eines Textes in Wörter.

Annotation: manuelle oder automatische Annotation – Hinzufügung von Zusatzinformationen (Anmerkungen, Kommentaren) zu einem Text. Die manuelle wird händisch unter Nutzung von Tagsets vorgenommen, die halbautomatisierte kann mit Machine-Learning-Verfahren (KI) durchgeführt werden

Distant Reading und **Close Reading**: Distant Reading ist die Verarbeitung (Analyse) großer Textmengen, ohne dass die Texte von einem Menschen gelesen werden. Close Reading bezeichnet das sorgfältige Lesen eines Textes.

Bei **EDA** (explorative Datenanalyse) werden Daten, Texte und Korpora auf bestimmte Auffälligkeiten (oder Trends) hin explorativ untersucht, vgl. dazu die Einleitung von A. Lucke im vorliegenden Band. Dazu kommen viele der hier genannten Tools zum Einsatz.

Entwicklungsumgebung: enthält die Tools zur Softwareentwicklung mit einer bestimmten Programmiersprache, z.B. Java, C++, Python) wie Editor, Compiler, Testsystem, Spracherweiterungen usw.

Funktion: ist ein Teil eines Programms und realisiert eine bestimmte Methode/Anforderung, z.B. die Aufspaltung eines Textes in Wörter.

Häufigkeit eines Wortes: ist ein statistisches Merkmal eines Textes und wird für Textanalysen benötigt. Es werden absolute (Anzahl der Vorkommen des Wortes im Text) und relative (absolute Häufigkeit / Anzahl Wörter im Text) Häufigkeiten verwendet.

Häufigkeitsverteilung: zeigt die Verteilung der Häufigkeit eines Wortes in Textabschnitten, die gleich lang oder durch Kapitel, Szenen o.ä. definiert sind.

Konfigurationsmatrix: ist eine Tabelle, die in der i-ten Zeile und der j-ten Spalte eine 1 hat, wenn in einem Drama die Figur i in der Szene j auftritt (Binärmatrix, Szenenstruktur).

KWIC: Keywords in Context, Nachbarn / Kontext eines Wortes im Text

Lemmatisierung: bezeichnet die Reduktion der Wortform auf ihre Grundform.

Methode: ist ein bestimmtes Verfahren zur Lösung einer Aufgabenstellung, also der Weg zu einem angestrebten Ziel. Eine Methode wird programmtechnisch mittels Algorithmen in Form von Funktionen umgesetzt. Die Methode ist damit die abstrakte Vorgehensweise, während ein Algorithmus zur Programmierung dieser Methode sehr detailliert ist und u.a. mit Pseudocode beschrieben werden kann.

Named Entity Recognition: erkennt eindeutig benannte Größen im Text, z.B. Personen, Länder, Orte, Produkte, Organisationen, Buchtitel.

Netzwerkanalyse: Ein Netzwerk besteht aus Knoten und Kanten. Damit können mittels Dicke und Farbe bestimmte Zusammenhänge dargestellt werden. Mit solchen Netzen können z.B. gemeinsame Auftritte von je zwei Figuren in Dramen oder die Ähnlichkeiten von Literaturtexten in der Stilometrie angezeigt werden (z.B. mit dem Simmelian Backbone Network, siehe Studie von Weitin 2021).

NLP: Natural Language Processing enthält eine Menge von Tools zur Sprachverarbeitung.

Package: sind fertige Programmteile, diein einer bestimmten Programmiersprache entwickelt sind und die ein oder mehreren Funktionen enthalten. Sie können in die Entwicklungsumgebung eingebunden werden, z.B. stylo in „R" für Stilometrie, oder spaCy in der Python-Entwicklungsumgebung für NLP-Funktionen.

POS-Tagging (Part-of-Speech-Tagging): ist die Zuordnung von Wörtern und Satzzeichen eines Textes zu Wortarten.

Preprocessing / Vorverarbeitung: Bevor die eigentliche Textanalyse startet, wird der Text passend aufbereitet, z.B. in seine Wörter zerlegt (Tokenisierung). Ein andere Form der Vorarbeiten ist die Aufteilung von Texten in Trainings- und Testdaten zum Einsatz von KI-Tools.

Programm / Tool: ist eine IT-Lösung zu einer vorgegebenen Aufgabenstellung, z.B. Ermittlung der Ähnlichkeit von Novellentexten mithilfe von bestimmten Abstandsmaßen auf Basis der häufigsten Wörter (Stilometrie).

Programmiersprache: ist Teil der Entwicklungsumgebung zur Programmierung eines Programms für eine vorgegebene Menge von Anforderungen, z.B. Java oder Python.

Sentimentanalyse: sucht nach Informationen im Text wie verbalisierte menschliche Gefühle, Empfindungen und/oder Meinungen.

Stoppwortliste: enthält alle diejenigen Wörter, die für eine Textanalyse ausgeblendet werden.

TEI-Format: ist ein Standard für die Darstellung, den Austausch und die Speicherung von Texten in digitaler Form und basiert auf XML

Tokenisierung: Ermittlung der Wörter (oder Sätze, Phrasen, Absätze u.a.) in einem Text mittels Segmentierung in Einheiten der Wortebene.

Topic Modeling: in einer großen Textsammlung sollen ähnliche Textteile in Form von Wortgruppen gefunden werden. Eine Wortgruppe ist ein Topic, z.B. „Theater, Schauspieler und Stück" oder „Euro, Bank und Wirtschaft".

Visualisierung: macht einen Sachverhalt, etwa das Ergebnis einer Analyse, sichtbar z.B. mittels Grafiken (Netzwerkgraphen) oder farbig abgestuften Tabellen.

Vorverarbeitung: siehe Preprocessing

Wordcloud: ist eine Form der Visualisierung und zeigt eine Wolke mit Wörtern aus einem Text/einer Textmenge, wobei die Häufigkeit des Wortes durch die Schriftgröße des Wortes in der Cloud dargestellt wird und das häufigste Wort in der Mitte steht.

Worteinbettung (Word Embedding) oder Vektorisierung: damit werden die Wörter eines Textes in numerische Vektoren umgewandelt, um u.a. Zusammenhänge zwischen den Wörtern mathematisch besser berechnen zu können.