Glossar

Kevin Wunsch[a] und Christopher A. Nunn[b]

[a] https://orcid.org/0000-0003-1491-747X, [b] https://orcid.org/0000-0001-7208-8636

Algorithmus
Ein Algorithmus bezeichnet ein schrittweises, determiniertes Verfahren zur Lösung einer bestimmten Aufgabe. Bei gleichbleibenden Daten wird das Ergebnis über jede Iteration des Algorithmus dasselbe sein. In der digitalen Editorik existiert etwa das Prinzip *E-V-A* (= **E**ingabe, **V**erarbeitung und **A**usgabe der XML-Datei, die am Ende als HTML-Struktur ausgegeben wird).

Alignierung
Die Alignierung ist der Bioinformatik entlehnt, wo sie Ähnlichkeiten innerhalb biologischer Sequenzen erfassen soll. In den Digital Humanities bezeichnet man damit den Abgleich von Daten aus verschiedenen Quellen. Die Alignierung hilft, komplexe Zusammenhänge sichtbar zu machen.

Allographen
Unter Allographen versteht man Varianten eines Zeichens, die denselben sprachlichen Wert repräsentieren. Oftmals kommen Allographen in verschiedenen Schriftsystemen vor. Die Unterschiede sind häufig klein und haben keine Auswirkungen auf die Bedeutung.

APC (Article Processing Charge)
APC bezeichnet die Kosten, die an Verlage zu entrichten sind, wenn eine Open Access-Veröffentlichung angestrebt wird.

API
Eine API bezeichnet eine Form von Schnittstelle, mit der man z. B. mit den Daten einer Website interagieren kann. *Information retrieval* findet meist über Schnittstellen statt. Gängige Standards für Schnittstellen im Web sind REST *(Representational State Transfer)* und SOAP *(Simple Object Access Protocol).*

Kevin Wunsch und Christopher A. Nunn: Glossar, in: Christopher A. Nunn und Frederike van Oorschot (Hrsg.): Kompendium Computational Theology 1. Forschungspraktiken in den Digital Humanities, Heidelberg: heiBOOKS, 2024, S. 503–517. DOI: https://doi.org/10.11588/heibooks.1459.c21929

Artefakt
Ein als historisch wahrgenommenes Überbleibsel, das in irgendeiner Form Gegenstand der Forschung sein kann.

Augmented Reality
Die computergestützte Überlagerung und Erweiterung der Realität mit einer oder mehreren Schichten künstlicher Realität, die die Realität erweitern und anreichern.

Backend
Das Backend ist der unsichtbare Teil eines Softwaresystems, verantwortlich etwa für Datenverarbeitung und -verwaltung.

Boxplot
Ein Boxplot ist die visuelle Zusammenfassung von Variabilität. Neben dem Median werden Quartile und Minimal- beziehungsweise Maximalwerte sowie auffallende Ausreißer dargestellt. Abb. 1 z. B. ist der Dissertation von Keli Du entnommen und präsentiert 2 000 Zeitungsartikel aus DIE ZEIT, die zwischen 2001 und 2014 veröffentlicht wurden. Mit den Boxplots wurde hier die durchschnittliche Textlänge von je 200 Dokumenten illustriert, die von der ZEIT-Redaktion unterschiedlichen Klassen zugeordnet worden waren.

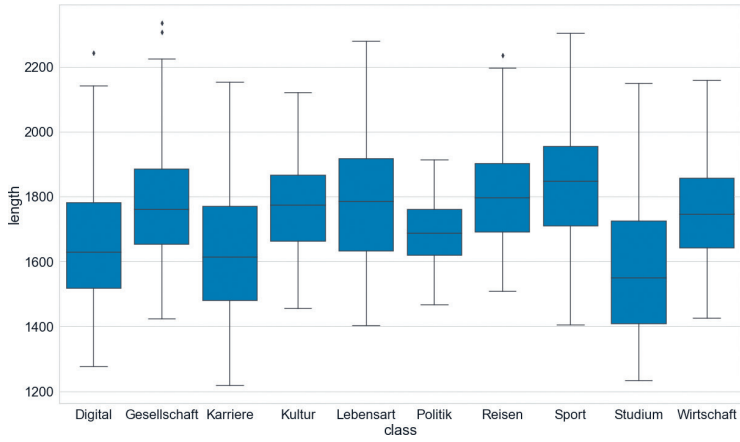

Abb. 1 Verteilung der Textlänge in den zehn Klassen der Zeitungsartikel.

Close Reading
Das *klassische* Lesen (nach Franco Moretti, der dieses dem Distant Reading unversöhnlich gegenüberstellt). Close Reading ist eine Methode der Textanalyse, bei der ein Text detailliert und sorgfältig untersucht wird, um seine formalen und sprachlichen Merkmale sowie die Bedeutungsebenen zu verstehen. Diese Methode gilt als konzeptioneller Vorläufer des *New Criticism,* einer literaturtheoretischen Richtung, die das formalisierte Close Reading als eine zentrale Analysetechnik verwendet.[1]

Codec
Kofferwort aus *Coder* und *Decoder.* Es beschreibt ein Algorithmenpaar, das digitale Daten kodiert und dekodiert. Dies dient oftmals dem Zweck der effizienteren Datenübertragung.

Container
Containersysteme ermöglichen Entwickler*innen die einfachere Skalierung ihrer Software. Ein weitverbreitetes Container-System ist die open-source Lösung *Docker*[2]. Mit der Hilfe von Container-Systemen wird das Problem „bei mir funktioniert es aber" abgeschwächt, da die Entwicklungs- und die Deployment-Umgebung identisch sind.

Content Management System (CMS)
Eine Software(-sammlung), mit deren Hilfe Inhalte im digitalen Raum auch ohne technisches Wissen verwaltet, bearbeitet und veröffentlicht werden können. *Wordpress* ist ein weit verbreitetes CMS.

Crowdsourcing
Mit Crowdsourcing bezeichnet man die Auslagerung von Ressourcen an die Community oder *Crowd.* So kann etwa Archivmaterial durch Ehrenamtliche transkribiert werden (z. B. im Bürgerprojekt *Consilium Communis* zur Transkription historischer Dokumente im Stadtarchiv Neuss[3]).

Daten
Daten sind digitale Repräsentation von Informationen aus den Forschungsrichtungen. Es können Texte, Bilder, Tonaufnahmen, Videoaufnahmen, aber auch Metadaten oder aggregierte Daten sein. Grundsätzlich kann zwischen strukturierten und unstrukturierten Daten unterschieden werden.

1 Vgl. Moretti, F. (2000) Conjectures on World Literature, New Left Review, 1, 54–68; Herrnstein-Smith, B. (2016). What was „Close Reading"? A Century of Method in Literary Studies, Minnesota Review, 87, 57–75.
2 S. https://www.docker.com, zuletzt aufgerufen am 14. 07. 2024.
3 S. https://www.stadtarchiv-neuss.de/nachrichten-detail/198.html, zuletzt aufgerufen am 14. 07. 2024.

Daten (strukturiert)
Ein gutes Beispiel für strukturierte Daten sind XML-Daten, die die hierarchische Struktur des Textes abbilden.

Daten (unstrukturiert – messy/fuzzy)
Unstrukturierte oder *messy/fuzzy* Daten sind solche, die unvollständig oder inkonsistent sind. Ihre Interpretation und Analyse ist gemeinhin als schwerer zu betrachten, als die von strukturierten Daten. Die algorithmische Auswertung solcher Daten ist ebenfalls aufwändiger als die von strukturierten Daten.

Deep Learning
Deep Learning ist ein Teilgebiet des *Machine Learning*, das auf neuronalen Netzen basiert. Diese vielschichtigen Netzwerke sind optimiert für die Verarbeitung großer Datenmengen und können automatisiert etwa Muster und Merkmale in Daten erkennen.

Diasystem
Ein Diasystem bezeichnet ein System, in dem Varianten und Varietäten einer Sprache (auch Dialekte) eines definierten Bereiches vorkommen, etwa geografische oder soziale Systeme. Dabei können die Varietäten alle Bereiche der Linguistik umfassen. In der Linguistik sind Diasysteme ein Mittel zur Analyse der Vielfalt innerhalb einer Sprache.

Digital Turn
Der Begriff beschreibt die Hinwendung von traditionellen Methoden und Praktiken hin zur großflächigen Nutzung digitaler Technologien und Prozesse. Der *digital turn* revolutioniert viele Bereiche der Wissenschaft.

Disintermediation
Man könnte die Disintermediation auch als „Cutting the middle man" bezeichnen. Es wird also das Vermeiden und Umgehen von klassischen Vermittlern beschrieben.

Distant Reading
Distant Reading beinhaltet die Analyse von Texten *aus der Ferne.* Algorithmen und computergestützte Methoden ermöglichen die Auseinandersetzung mit gigantischen Korpora *(big data).* Wo sich das *nahe* Lesen (vgl. Close Reading) auf einzelne Texte konzentriert, werden beim Distant Reading oftmals quantitative Ansätze, Text-Mining oder Muster-Analyse genutzt, um Aussagen über größere Korpora zu treffen.

Entität
In der Informatik werden hierunter Objekte oder Sachen verstanden, die durch Daten beschrieben werden können. Sie können real, aber auch Abstrakta sein. In den Di-

gitalen Geisteswissenschaften gibt es zudem mehrere Vokabulare zur Beschreibung von Objekten. Der Standard *CIDOC-CRM*[4] definiert eine Entität als Ding und definiert davon ausgehend Entitäten wie etwa Orte, Personen oder Dokumente.

Feature
Als Feature wird ein bestimmtes Charakteristikum eines Gegenstandes (bzw. einer Entität) bezeichnet, etwa eines Artefaktes oder eines Textes.

Figurenideolekt
Darunter versteht man die individuelle Sprachweise einer literarischen Figur. Diese macht die Figur unverwechselbar und betont die Persönlichkeit sowie soziokulturelle Hintergründe.

Framework
Ein Framework ist im weitesten Sinne eine festgelegte Sammlung von wiederverwendbaren (Code-)Bausteinen, die die Softwareentwicklung vereinfacht. Frameworks sind als Grundgerüst vieler Bereiche unerlässlich, da sie nicht nur den Entwicklungsprozess vereinfachen und beschleunigen, sondern den Entwickler*innen auch ermöglichen, sich auf die Spezifika zu konzentrieren.

Frontend
Unter Frontend versteht man die *sichtbaren* Teile einer Website oder Anwendung. Beliebte Frameworks für die Frontendentwicklung in der Webentwicklung sind *ReactJS*[5] oder *Angular*[6].

Geokodierung
Geokodierung beschreibt das Kodieren von Ortsdaten, sodass diese von Computern ausgelesen werden können. Es ermöglicht die Visualisierung der Ortsdaten auf Karten und in Geoinformationssystemen. Die bekanntesten Tools zur Verarbeitung solcher Daten sind *ArcGis*[7] und das Open-source Tool *QGIS.*[8]

GLAM (Galleries, Libraries, Archives, and Museums)
Gemeinhin fallen alle Institute des kulturellen Erbes unter diesen Sammelbegriff.

4 S. https://www.cidoc-crm.org, zuletzt aufgerufen am 14. 07. 2024.
5 S. https://react.dev, zuletzt aufgerufen am 14. 07. 2024.
6 S. https://angular.dev, zuletzt aufgerufen am 14. 07. 2024.
7 S. https://www.arcgis.com/index.html, zuletzt aufgerufen am 15. 07. 2024.
8 S. https://www.qgis.org/, zuletzt aufgerufen am 15. 07. 2024.

Graphdatenbank
Eine Graphdatenbank ist eine Datenbank, die Daten in Knoten (Entitäten) und Kanten (Beziehungen) speichert. Sie ermöglicht das Modellieren und Abfragen komplexer Beziehungsstrukturen.[9] Neben XML-Datenbanken sind Graphdatenbanken ein verbreiteter Datenbanktyp zum Erstellen digitaler Editionen. XML kann allerdings auch als gerichteter Graph betrachtet werden, dessen Elemente und Attribute Knoten darstellen, während die gerichteten Kanten die hierarchischen Beziehungen wiederspiegeln.

GUI (Graphical User Interface)
GUI bezeichnet eine grafische Nutzeroberfläche, den wohl gebräuchlichsten, aber nicht immer effizientesten Weg zur Interaktion mit einer Anwendung.[10]

Heatmap
In einer Heatmap werden Daten farblich hervorgehoben, z. B. die Positionen, die ein Fußballspieler im Verlauf des Spiels einnimmt.[11]

Hidden Markov Model
Hidden Markov-Modelle werden oftmals in der Sprach- oder Bioinformatik verwendet, um Muster zu erkennen. Dabei wird von einer Serie von beobachtbaren Ereignissen und Datenpunkten auf verborgene Zustände geschlossen. Das Modell besteht aus Zuständen, Übergangswahrscheinlichkeiten sowie Emissionswahrscheinlichkeiten. Sie beschreiben die Wahrscheinlichkeit, mit der ein bestimmtes beobachtbares Ereignis eintritt.

HTML (Hypertext Markup Language)
HTML ist die Auszeichnungssprache des Internets, d. h. die beschreibende Sprache, aus der Webseiten bestehen.

Inference Server
Ein Inference Server ist ein GPU *(Graphics Processing Unit)*-unterstützter Spezialserver, der möglichst schnell – idealerweise annähernd in Echtzeit – mit Hilfe maschineller Modelle Vorhersagen treffen kann.

9 Vgl. Kuczera, A. (2017). Graphentechnologien in den Digitalen Geisteswissenschaften, ABI Technik, 179–196. https://doi.org/10.1515/abitech-2017-0042, zuletzt aufgerufen am 14. 07. 2024.

10 Vgl. Drucker, J. (2011). Humanities Approaches to Graphical Display, *digital humanities quarterly,* 5(1), 1–52. http://www.digitalhumanities.org/dhq/vol/5/1/000091/000091.html, zuletzt aufgerufen am 14. 07. 2024.

11 S. hierzu z. B. Memmert, D., & Raabe, D. (Hrsg.). (2019). Revolution im Profifußball. Mit Big Data zur Spielanalyse 4.0. 2. Aufl (S. 92–93). Berlin: Springer. https://doi.org/10.1007/978-3-662-59218-2, zuletzt aufgerufen am 14. 07. 2024

Information Retrieval
Typischerweise meint *Information Retrieval* den algorithmengestützten Prozess des Datenfindens und -gewinnens, um daraus Anfragen zu bearbeiten, z. B. zu beantworten.

Java
Eine kompilierte Programmiersprache, die Grundlage vieler Datenbankanwendungen ist. *ExistDB*[12], das *Standardtool* zum Erstellen digitaler Editionen, oder *ediarum*[13], ein Framework zum Erstellen digitaler Editionen, sind in Java geschrieben.

Javascript
Javascript war lange als Sprache zur Manipulation von Websites bekannt, etwa zur Änderung einer Button-Farbe beim Klick auf denselben. Seit einigen Jahren wird es vermehrt in Backend-Anwendungen genutzt. Es hat, obwohl der Name anders vermuten lässt, keine Gemeinsamkeiten mit Java.

Jupyter Notebooks
Jupyter Notebooks bezeichnen eine interaktive Webanwendung, die es Nutzern ermöglicht, alle Aspekte der Arbeit in einem Webdokument zu sammeln. Besonders in den Bereichen *Data Science* und den *data driven* Wissenschaften sind Jupyter Notebooks der Standard für die Entwicklung und Präsentation von Ergebnissen.

Kernel
Der Kern jedes Unixbasierten Betriebssystems (Linux, MacOS). Er enthält grundständige Informationen und Operatoren für den Computer. Der Kernel ist quasi das Herz des Rechners, der die grundlegenden Funktionen des Systems steuert.

Lemmatisierung
Bei der Lemmatisierung wird ein Wort auf seine Grundform (Lemma, d. h. auf die Wörterbuchform) reduziert, z. B. *gegangen → gehen.*

Linked (Open) Data
Linked Open Data ist ein Teil des Semantic Webs, welches wiederum als ein *Internet für Maschinen* beschrieben werden kann und aus mehreren Ebenen besteht. Im Semantic Web geht es nicht darum, Daten und Informationen menschenverarbeitbar bereitzustellen, sondern darum, dass *intelligente Maschinen* die Daten nutzen können. Dies wird etwa durch die Verknüpfung von Daten *(linked data)* ermöglicht.

12 S. http://exist-db.org/exist/apps/homepage/index.html, zuletzt aufgerufen am 14. 07. 2024.
13 S. https://www.ediarum.org, zuletzt aufgerufen am 14. 07. 2024.

Machine Learning
Machine Learning, ein Teilgebiet der Künstlichen Intelligenz, ermöglicht, dass Computer aus Daten lernen und ohne explizite Programmierung Aufgaben bewältigen. Es verwendet Algorithmen, um aus Daten Vorhersagen zu extrahieren und Muster zu erkennen. Dabei unterscheidet man zwischen *supervised learning* und *unsupervised learning*. Beim bewachten Lernen werden Algorithmen mit gelabelten Daten trainiert, während beim unüberwachten Lernen Algorithmen genutzt werden, um Muster und Strukturen in nicht gelabelten Daten zu erkennen. Beim *reinforcement* werden negative und positive Ergebnisse gewertet, etwa ein gewonnenes Spiel, sodass ähnliche Strategien zur Problemlösung häufiger und ineffiziente seltener genutzt werden.

Macroanalysis
Unter *Macroanalysis* versteht man den Ansatz, große Textcorpora mit quantitativen Methoden (etwa Distant Reading) zu analysieren. Prominent wurde dieser Begriff durch die gleichnamige Monographie von Matthew Jockers aus dem Jahr 2013.

Mel Frequency Cepstral Coefficients
Diese Koeffizienten sind eines der zentralen Merkmale in der computergestützten Audioverarbeitung. Sie bieten eine effiziente Darstellung wesentlicher Merkmale von Audiosignalen.

Mid-Range-Reading
Mid-Range-Reading steht methodisch zwischen Close Reading und Distant Reading.[14] Im deutschsprachigen Raum ist eher der Begriff des *Scalable Reading* geläufig, eine Leseweise, die je nach Gegenstand und Fragestellung zwischen Close und Distant Reading changiert.[15]

Mixed Methods
Methoden verschiedener wissenschaftlicher Traditionen werden kombiniert – ein häufiges Beispiel ist die gleichzeitige Nutzung von qualitativen und quantitativen Methoden.

14 S. Booth, A. (2020). Mid Range Reading. Not a Manifesto, *Publications of the Modern Language Association of America,* 132(3), 620–627. https://doi.org/10.1632/pmla.2017.132.3.620, zuletzt aufgerufen am 14. 07. 2024.

15 S. z. B. Krautter, B. (2024). The Scales of (Computational) Literary Studies. Martin Mueller's Concept of Scalable Reading in Theory and Practice. In F. Armaselu & A. Fickers (Hrsg.), *Zoomland. Exploiring Scale in Digital History and Humanities* (S. 261–286, v. a. 262). Berlin/Boston: De Gruyter Oldenbourg [= *Studies in Digital History and Hermeneutics,* 7].

Neuronales Netzwerk
Der Begriff ist von biologischen neuronalen Netzwerken des Gehirns inspiriert. Im informationswissenschaftlichen Kontext besteht ein neuronales Netzwerk aus verschiedenen in Schichten organisierten Einheiten zur Informationsverarbeitung. Sie gewichten dabei Eingaben, um anhand von Mustern Vorhersagen zu treffen.

N-Gramm
Mit einem N-Gramm werden zusammenhängende Sequenzen aus einem Text beschrieben: „Hier steht" ist ein Bigram, während „Hier steht Text" ein Trigram ist. Mithilfe solcher NGramme lassen sich Muster erkennen oder aber Kontexte und Häufigkeiten von Wortketten analysieren.

Natural Language Processing (NLP)
NLP bezeichnet ein Teilgebiet der künstlichen Intelligenz, welches sich mit der Interaktion zwischen Computern und menschlicher Sprache befasst. Darunter fallen Methoden zur Verarbeitung und Analyse, aber auch Generierung natürlicher Sprache. Anwendungen reichen von der maschinengestützten Übersetzung und Textklassifikation bis hin zu Sprachgenerierung und Stimmungsanalyse. Es werden hierbei die Methoden der Linguistik und der Informatik kombiniert, um Computern das Verstehen menschlicher Sprache zu ermöglichen.

Noise
Als *Noise* wird ein *Rauschen* innerhalb eines Datensatzes bezeichnet, welches Muster stört und unterbricht. Es sind irrelevante oder zufällige Datenvariationen, die die Leistung von Modellen beeinträchtigen können. Mitunter ist es schwierig, wichtige Informationen vom Rauschen zu unterscheiden. Dabei können verschiedene Methoden der Datenbereinigung hilfreich sein.

OJS (Open Journal Systems)
OJS beinhalten Open Source Software zur Verwaltung und Veröffentlichung wissenschaftlicher Zeitschriften, die vorwiegend im Open Access-Bereich genutzt wird. Die *Zeitschrift für digitale Geisteswissenschaften*[16] basiert z. B. auf OJS.

Operationalisierung
Eine Operationalisierung beschreibt die Umwandlung von abstrakten Konstrukten in Variablen. So werden maschinelle Untersuchungen ermöglicht, da klar quantifizierbare Indikatoren definiert werden.[17]

16 S. https://zfdg.de, zuletzt aufgerufen am 14. 07. 2024.

17 Zu ausführlichen Informationen zu dieser Technik s. Krautter, B., Pichler, A., & Reiter, N. (2023). Operationalisierung. In AG Digital Humanities Theorie des Verbandes Digital Humanities im deutschsprachigen Raum (Hrsg.), *Begriffe der Digital Humanities. Ein diskursives Glossar* (o. S.).

Optical Character Recognition (OCR)
OCR dient der computergestützten Erkennung von Text. Dafür gibt es verschiedene Tools, teilweise auf dem eigenen Gerät, teilweise im Web. *Transkribus*[18], *OCR-D*[19], *Tesseract*[20] und *Abby FineReader*[21] sind die wohl bekanntesten Tools.

Overfitting
Der Effekt eines *zu gut* trainierten Modells im Machine Learning. Das Modell hat die Trainingsdaten – einschließlich Ausnahmen und Noise – so sehr verinnerlicht, dass die Generalisierung auf neuen Daten schlechter wird. Dies bedeutet, dass das Modell auf den Trainingsdaten sehr gut abschneidet, auf den Testdaten allerdings nicht.

Part-of-Speech-Tagging (POS-Tagging)
POS-Tagging nennt man die Auszeichnung der verschiedenen Wortarten in einem Text. Dies hilft beim Verständnis der Struktur eines Textes, was einen grundlegenden Schritt vieler NLP-Aufgaben darstellt.

PID (Persistent Identifier)
Persistente Identifier sind dauerhaft verfügbar und unveränderlich. Der Verweis funktioniert über eines von vielen Systemen, die bekanntesten sind PURL *(Persistent Uniform Resource Locator),* DOI *(Digital Object Identifier)* und URI *(Uniform Resource Identifier).*

Pipeline (im NLP-Bereich)
Als Pipeline wird eine Aneinanderreihung von Schritten zur Textverarbeitung oder Analyse bezeichnet. Eine mögliche Pipeline wäre z. B. Tokenisierung → POS-Tagging → Feature Extraction, um einen Rohtext zu strukturieren, der anschließend für maschinengestützte Analysen genutzt werden kann.

Plugin
Ein Plugin ist eine optionale Erweiterung einer Anwendung. Beispiele dafür sind etwa AdBlocker für Browser oder das *Zotero*-Plugin in den bekannten Office-Anwendungen, das das Einfügen von Textverweisen ermöglicht.

Wolfenbüttel: Herzog August Bibliothek [= *Zeitschrift für digitale Geisteswissenschaften. Working Papers,* 2]. https://doi.org/10.17175/wp_2023_010, zuletzt aufgerufen am 14. 07. 2024.

18 S. https://www.transkribus.org/de, zuletzt aufgerufen am 14. 07. 2024.

19 S. https://ocr-d.de, zuletzt aufgerufen am 14. 07. 2024.

20 S. https://github.com/tesseract-ocr/tesseract, zuletzt aufgerufen am 14. 07. 2024.

21 S. https://pdf.abbyy.com, zuletzt aufgerufen am 14. 07. 2024.

React
React ist ein weit verbreitetes Frontend-Framework auf Basis von *Javascript.* Es beschleunigt die Entwicklung von responsiven und leicht wartbaren Webanwendungen. Dies kann zugunsten der Barrierefreiheit gehen.

Relationale Datenbank
Eine relationale Datenbank stellt ein System dar, das Daten in Tabellen ablegt und verwaltet. Die Beziehungen zwischen den Tabellen werden durch Schlüssel definiert. Für Abfragen wird eine Form von SQL verwendet.

Retrodigitalisierung
Die Digitalisierung von analogen Medien, etwa Bücher, Fotografien, oder Tonaufnahmen, bezeichnet man als Retrodigitalisierung. Durch diese werden die Medien zugänglich, durchsuchbar und langfristig verfügbar.

Scatterplot
Ein Scatterplot ist eine grafische Darstellung zweier Variablen und deren Wertepaare in einem zweidimensionalen Koordinatensystem. Beziehungen zwischen den Variablen sollen hierbei hervorgehoben und mögliche Zusammenhänge sichtbar gemacht werden. Im folgenden Beispiel (Abb. 2) hat Christof Schöch 60 Themen und deren Vorkommen in 890 französischen Dramen zwischen 1610 und 1810 (von der Klassik bis zur Aufklärung – Grundlage ist die *Théâtre classique collection* von Paul Fièvre) untersucht. Die Ergebnisse visualisierte er in einem Scatterplot.

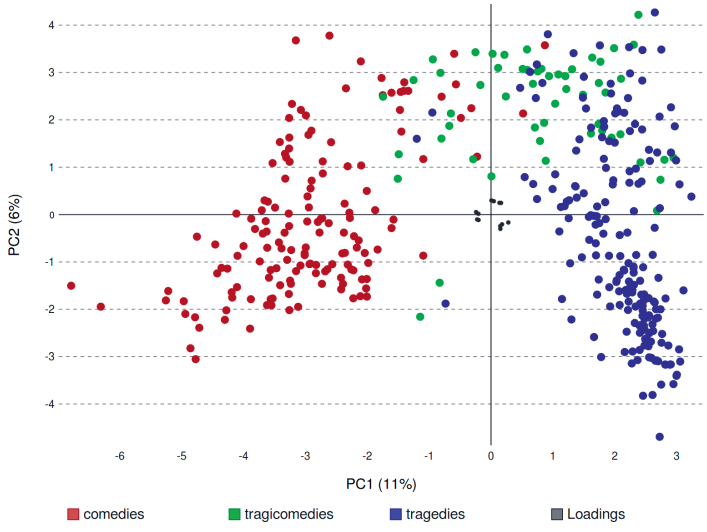

Abb. 2 Scatterplot nach einer *Principal Component Analysis* (Verfahren der multivariaten Statistik) zu französischen Dramen und der Wahrscheinlichkeit des Auftritts von 60 Themen darin

Segmentierung (von Texten)
Aufteilung eines Textes in sinnvolle kleinere Einheiten, etwa Sätze, Absätze oder einzelne Wörter. Oftmals sorgt die Segmentierung eines Textes für verbesserte Möglichkeiten zur computergestützten Analyse.

SIFT-Verfahren (Scale-Invariant-Feature Transform)
Unter einem SIFT-Verfahren wird ein Algorithmus aus der Bildverarbeitung verstanden, der zentrale Bildelemente erkennt und beschreibt. Das Verfahren wird zum Bildvergleich genutzt.

Skin
Als *Skin* wird ein durch Nutzer*innen anpassbares Design oder Erscheinungsbild bezeichnet. Dieses beeinflusst nicht die Funktionalität.

Softmax
Softmax wird insbesondere in neuronalen Netzwerken genutzt, um Wahrscheinlichkeitsverteilungen über eine Menge von Klassen zu erzeugen. Dabei werden Vektoren zu Wahrscheinlichkeitsverteilungen transformiert, sodass die Summe aller Wahrscheinlichkeiten 1 beträgt.

SQL (Structured Query Language)
SQL ist die Grundsprache von relationalen Datenbanksystemen. Weite Teile des Internets nutzen relationale Datenbanken.

Stemming
Stemming ist eine Methode aus dem NLP, die darauf abzielt, verschiedene Formen eines Wortes auf den gemeinsamen Stamm zu reduzieren. Hierbei werden Suffixe und Präfixe entfernt, sodass nur noch der Wortstamm stehen bleibt (z. B. *gegangen* → *geh*). Für Klassifizierungsfragen, etwa Sentimentanalyse, ist *Stemming* eine wichtige Methode.

Stoppwörter
Die semantische Bedeutung von Stoppwörtern wird als gering eingestuft. Sie werden daher entfernt, um Relevanz und Effizienz von Analysen und Anfragen zu verbessern. Je nach Textgattung ist es wichtig, nicht vorschnell auf automatisierte Stoppwortlisten zurückzugreifen, sondern diese lieber „fach- und sachgerecht“ zu erarbeiten. Andernfalls können zentrale Inhalte der maschinellen Analyse entgehen (man denke nur an Hamlets „Sein oder nicht sein“).[22]

22 Vgl. Ch. Schubert (2021). Digital Humanities auf dem Weg zu einer Wissenschaftsmethodik. Transparenz und Fehlerkultur, Digital Classics Online, 7, 39–53, hier: 41–42. https://doi.org/10.11588/dco.2021.7.82371, zuletzt aufgerufen am 14. 07. 2024.

TEI (Text Encoding Initiative)
TEI ist der Standard zum Kodieren von Texten nach einem festgelegten Vokabular, der fortlaufend aus der Community heraus erweitert werden und an veränderte Bedingungen angepasst werden kann. Er erfasst inhaltliche und strukturelle Einheiten in Texten in Tags und macht sie so auswertbar. Durch unterschiedliche XML-Dialekte gibt es auch Anpassungen der TEI-Guidelines für bestimmte Anwendungsbereiche, z. B. *TEI EpiDoc* für ein strukturiertes Markup von wissenschaftlichen Editionen antiker Dokumente, v. a. von Inschriften und Papyri, oder *TEI Correspondence* zur Auszeichnung von Briefen.

Tokenisierung
Bei der Tokenisierung wird der Text in eine zuvor festgelegte Einheit aufgeteilt. Häufig sind dies Wörter oder Phrasen. Dies ermöglicht eine Vielzahl computergestützter Analyseverfahren.

Toolchain
Als Toolchain bezeichnet man eine Sammlung von Werkzeugen, die in gleichbleibender Reihenfolge verwendet werden, um spezifische Aufgaben zu erledigen. Dabei übernimmt jedes Tool ein spezifisches Aufgabenset. Gemeinsam ermöglichen diese Werkzeuge den Entwicklungsprozess.

Transformer
Transformer sind ein relativ junges Konzept im Machine Learning, das 2017 entwickelt wurde. Die leistungsstarken *Large Language Models* (LLMs: BERT, Chat-GPT) wären ohne Transformer-Modelle wohl nicht möglich. Diese Modelle heben sich besonders durch Selbstaufmerksamkeit, Parallelverarbeitung, eine Encoder/Decoder-Architektur sowie Positionskodierungen und ihre hohe Skalierbarkeit von anderen Modellen ab.

Treebank
Treebanks wurden von Geoffrey Leech in den 1980ern als Begrifflichkeit und Methode zur Erfassung der Satzstruktur in einer Baumstruktur eingeführt. Der Baum teilt den Satz in seine Bestandteile auf, etwa Subjekt – Objekt – Verb.[23]

Type-Token-Ratio
Die *Type-Token-Ratio* ist das Maß für die lexikalische Vielfalt eines Textes. Sie beschreibt das Verhältnis der einzigartigen Wörter *(Types)* zur Gesamtanzahl der Wörter *(Tokens).* Je größer der Wert, desto vielfältiger ist das Vokabular, während ein niedriger Wert auf ein stark begrenztes Vokabular hinweist.

23 Vgl. Leech, G., & Eyes, E. (1997). Syntactic Annotation. Treebanks. In R. Garside, G. Leech & T. McEnery (Hrsg.), *Corpus Annotation* (S. 34–52). New York: Addison Wesley Longman.

Unicode (Universal Character Encoding)
Anders als andere Zeichensatz-Standards ist Unicode nicht auf ein Subsystem menschlicher Sprachen fokussiert, sondern auf die vollumfassende Darstellung aller Zeichen im Binärsystem. Es bildet daher auch die technische Basis für die Schriftzeichenkodierung.

Unix
Unix war ein Betriebssystem, das in den 1960er und 1970er Jahren entwickelt wurde. Aus diesem sind bekannte und weit verbreitete Betriebssysteme hervorgegangen – die bekanntesten dürften *Linux, BSD* und *macOS* sein.

XML (Extensible Markup Language)
XML ist eine Auszeichnungssprache zur Darstellung strukturierter Daten. Seit der Erstveröffentlichung hat sich XML in vielen Bereichen des digitalen Lebens durchgesetzt, so auch in der Community der Digitalen Editorik. Für gewöhnlich werden TEI-codierte Texte in XML ausgezeichnet.

Vektor
Ein Vektor bezeichnet eine mathematische Größe, die einen n-dimensionalen Raum abdecken kann. Häufig werden diese in der Algebra, Physik oder Informatik angewandt. Im NLP beschreibt ein Vektor die Dimensionen eines Wortes. Oftmals werden als Beispiel Städte angeführt: Die Städte Berlin und Paris sind jeweils Hauptstädte der Länder, in denen sie liegen (Deutschland, Frankreich), aber z. B. die Dimensionen der Bevölkerungsgröße oder gesprochenen Sprachen sind unterschiedlich.

Virtual Reality
Im Gegensatz zur *augmented reality* ist die virtuelle Realität eine künstliche Wirklichkeit. Heutzutage sind VR-Brillen, wie die *Oculus Rift* zumindest für Computer- und Konsolenspiele weit verbreitet. Die einfachste virtueller Realität ist daher wohl ein Computerspiel, das in eine nicht reale Welt entführt. Die Nutzung virtueller Realität für digitale Forschung nimmt jedoch ebenfalls immer weiter Fahrt auf.[24]

Visual Computing
Visual Computing beschreibt ein interdisziplinäres Feld der Informatik. Dieses befasst sich mit allen Aspekten der digitalen Arbeit mit Bildern, etwa der Erzeugung von Bildern und der automatisierten Auswertung von Bildinformationen *(Computer Vision)* oder aber auch der Bildverarbeitung und Visualisierung.

24 S. z. B. das *VR-Lab* am Institut für Digital Humanities der Georg-August-Universität Göttingen. https://www.uni-goettingen.de/de/vr-lab/662748.html, zuletzt aufgerufen am 14. 07. 2024. Hier werden „Virtual-, Augmented- und Mixed-Reality-Technologien [genutzt], um Räume der Vergangenheit zu rekonstruieren und diese Visualisierungen kritisch zu hinterfragen."

Bildnachweise

Abb. 1: Du, K. (2022). *Zum Verständnis des LDA Topic Modeling. Eine Evaluation aus Sicht der Digital Humanities* [Diss.]. Online: Universität Würzburg. Hier: S. 74, Abb. 5.1 (CC BY 4.0). https://opus.bibliothek.uni-wuerzburg.de/opus4-wuerzburg/frontdoor/deliver/index/docId/34826/file/Du_Keli_Dissertation.pdf (zuletzt aufgerufen am 14. 07. 2024).

Abb. 2: Schöch, Ch. (2017). Topic Modeling Genre. An Exploration of French Classical and Enlightenment Drama, *digital humanities quarterly,* 11(2), 1–53, hier: 35, fig. 10. http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html (zuletzt aufgerufen am 14. 07. 2024).