Toolbeitrag: Voyant

Marie Flüh [1]

1. Universität Hamburg

forTEXT

Thema:	Textvisualisierung	DOI:	10.48694/fortext.3775
Jahrgang:	1	Ausgabe:	5
Erscheinungsdatum:	2024-08-07	Erstveröffentlichung:	2018-11-26 auf fortext.net
Lizenz:	CC BY SA		open access

Allgemeiner Hinweis: Rot dargestellte Begriffe werden im Glossar am Ende des Beitrags erläutert. Alle externen Links sind auch am Ende des Beitrags aufgeführt.

Der Workflow von Voyant lässt sich in vier Schritte unterteilen: 1) Vorab: Vorbereitung der zu untersuchenden Textsammlung oder des Textes; 2) Input: Hochladen der Textdokumente; 3) Interface: „Architektur" der Analyselandschaft gestalten, indem die Tools für die Panels ausgewählt werden; 4) Output: Export der Ergebnisse über den Exportbutton in jedem Panel

- **Systemanforderungen:** Voyant ist webbasiert (vgl. Webanwendung) mit verschiedenen Browsern (vgl. Browsersuchfunktion) (Firefox, Internet Explorer, Google Chrome, Safari) nutzbar, während der VoyantServer als Desktopapplikation nach dem Download – als sogenannte Standalone-Version – auch offline verwendet werden kann. Der VoyantServer (vgl. Server) ist mit Mac-, Windows- oder Linux-Betriebssystemen nutzbar, wobei Java installiert sein muss.
- **Stand der Entwicklung:** Im Jahr 2003 veröffentlicht, derzeitig läuft die Version Voyant 2. 4, die 28 Tools beinhaltet und stetig weiterentwickelt wird. Der VoyantServer 2. 4 M7 wurde im Juni 2018 veröffentlicht.
- **Herausgeber:** Geoffrey Rockwell (University of Alberta) und Stéfan Sinclair (McGill University).
- **Lizenz:** Voyant ist eine kostenlos nutzbare Open-Source-Software.
- **Weblink:** Onlineversion: https://voyant-tools.org, Standalone-Version: https://github.com/voyanttools/VoyantServer
- **Im- und Export:** Import durch die direkte Texteingabe in das Textfeld (Dateiformate: TXT (vgl. Reintext-Version), HTML, XML, Eingabe von URLs (vgl. URI)) oder durch das Hochladen der Dateien (Dateiformate: TXT, HTML, XML, PDF (geOCRt (vgl. OCR)), RTF, RDF, Word, Excel, CSV, ODT, Pages, TEI, RSS, DToc, Atom- und Archivdateien wie z. B. ZIP, TAR, TGZ). Die Ergebnisse können als HTML-Quellcode, fertig generierte URL, bibliografische Zitation oder tabellarisierte bzw. rohe Datensätze exportiert werden (Dateiformate wie z. B. CSV). Jedes Tool weist in dem entsprechenden Panel einen tooleigenen Exportbutton auf. Über den generierten Link können einzelne Tools mit den individuellen Feineinstellungen weitergeleitet werden. Darüber hinaus kann über einen die gesamte Textsammlung (vgl. Korpus) betreffenden Exportbutton ein Link abgerufen werden, der die Textsammlung und die jeweils ausgewählten Tools – allerdings in ihren Grundeinstellungen – beinhaltet.
- **Sprachen:** Siehe https://voyant-tools.org/docs/#!/guide/languages

1. Für welche Fragestellungen kann Voyant eingesetzt werden?

Voyant ist eine computerbasierte Textanalyselandschaft (vgl. Text Mining). Pro Arbeitseinheit lassen sich jeweils fünf unterschiedliche Tools (vgl. Feature) miteinander kombinieren, die eine quantitative Untersuchung (vgl. Distant Reading) von Texten oder Textsammlungen sowie verschiedene Formen der Textvisualisierung (Horstmann und Stange 2024) ermöglichen. Gängige Fragestellungen sind zum Beispiel: Welche Begriffe werden in einem Text oder einer Textsammlung am häufigsten verwendet? In welchem Kontext kommen ausgewählte Wör-

ter vor? Auf welche Art und Weise ballen sich Wörter in einer Textsammlung zusammen? Welche Verbindungen bestehen zwischen Figuren, Orten oder Organisationen?

2. Welche Funktionalitäten bietet Voyant und wie zuverlässig ist das Tool?

Funktionen: Voyant, dessen originär englisches Interface mittlerweile in neun weitere Sprachen übersetzt wurde, ermöglicht die Untersuchung, Analyse und Visualisierung zweier vorbereiteter Textsammlungen – bestehend aus acht Werken Jane Austens oder aus 27 Werken Shakespeares – oder die Arbeit an eigens zusammengestellten Textsammlungen (z. B. TextGrid Repository (Horstmann 2024), Deutsches Textarchiv (DTA) (Horstmann und Kern 2024)). Je nach Auswahl der Tools bietet Voyant eine Vielzahl von Funktionen, dazu zählen:

- Verschiedene Formen der Tokenisierung (vgl. Type/Token), also die Segmentierung des Textes in einzelne Wörter
- Erstellen erweiterbarer Stoppwortlisten (vgl. Stoppwortliste)
- unterschiedliche Varianten der Analyse, die sich gegenseitig beeinflussen (Wortsuche (vgl. Query), Wortkontexttools (vgl. Kollokation), Häufigkeit von Wörtern und Phrasen in einer Sammlung oder einem Text)
- Zahlreiche Formen der Visualisierung

Zuverlässigkeit: Voyant arbeitet relativ zuverlässig. Bei der Implementierung großer Textsammlungen kann es allerdings zu Verzögerungen, Ausfällen oder Fehlermeldungen kommen. Darüber hinaus sind einige Tools noch nicht vollends funktionstüchtig, worauf die Entwickler jedoch hinweisen. Die Visualisierungen einiger Tools – wie z. B. TextualArc oder Knots – können auf den ersten Blick überladen und unübersichtlich wirken. Da die Interpretation der Ergebnisse einen wichtigen Teil der Arbeit mit Voyant darstellt, empfiehlt es sich für Einsteiger*innen, zunächst auf die gängigen und einfach auszuwertenden Tools (z. B. Summary, Cirrus, ScatterPlot) zurückzugreifen oder die bestehende Benutzeroberfläche (vgl. GUI) beizubehalten (Cirrus, Reader, Trends, Summary, Contexts).

3. Ist Voyant für DH-Einsteiger*innen geeignet?

Checkliste	✓ / teilweise / –
Methodische Nähe zur traditionellen Literaturwissenschaft	teilweise
Grafische Benutzeroberfläche	✓
Intuitive Bedienbarkeit	✓
Leichter Einstieg	✓
Handbuch vorhanden	✓
Handbuch aktuell	✓
Tutorials vorhanden	✓
Erklärung von Fachbegriffen	teilweise
Gibt es eine gute Nutzerbetreuung?	teilweise

Voyant ermöglicht einen Brückenschlag zwischen innovativen digitalen Analyseverfahren von Texten und deren (literaturwissenschaftlicher) Interpretation. Hierbei steht die computerbasierte Textanalyse durch die Erhebung statistischer Daten im Fokus, deren Interpretation in einem zweiten Schritt und quasi offline – nämlich durch die Nutzer*innen – geschieht. Die methodische Nähe zur traditionellen Literaturwissenschaft besteht vor allem in der Interpretation der erhobenen Daten.
Kleinere Hilfestellungen und Erklärungen, deren Lesbarkeit durch eine zu kurze Anzeigezeit allerdings erschwert wird, sind als fester Bestandteil der GUI jedem Panel beigefügt und leiten bei Bedarf zum Handbuch weiter. Rückfragen oder Anmerkungen sind über Twitter und GitHub möglich, eine Mailadresse für Anfragen oder andere Formen der Unterstützung werden derzeit nicht angeboten.

4. Wie etabliert ist Voyant in den (Literatur-)Wissenschaften?

Voyant ist in den Wirkungsfeldern der digitalen Geisteswissenschaften fest etabliert. Da digitale Texte allgegenwärtig sind, wird Voyant darüber hinaus in verschiedenen Sachgebieten eingesetzt. Im Oktober 2016 wurde der VoyantServer von 81.686 Menschen aus insgesamt 156 Ländern besucht, wobei 1.173.252 Tool-Anwendungen umgesetzt wurden und der VoyantServer über 2000 Mal heruntergeladen wurde. Darüber hinaus ist Voyant auf der Homepage von sechs Universitätsbibliotheken als wichtige Ressource aufgelistet. Der Einbezug in

fachdidaktische Unterrichtskonzeptionen (Lechner, Henning und Müller 2018; Kemann 2016; Kühner 2017), journalistische Arbeiten (Sinclair und Rockwell 2012, 19) oder kreative Projekte (Sample 2013) verweist auf die mannigfachen Anwendungsbereiche des Tools. Ein Blick auf eine Auswahl von Blog-Posts, Artikeln, Kongressen, Workshops, Praxisbeispielen aus Schulen und Universitäten, in denen Voyant-Tools auf unterschiedliche Art und Weise Anwendung finden, verdeutlicht, dass Voyant kein reines Werkzeug der digitalen Geisteswissenschaften darstellt. Der Leitspruch „see through your text" lässt sich mittels der Voyant-Tools in ganz unterschiedlichen Kontexten in die Tat umsetzten.

5. Unterstützt Voyant kollaboratives Arbeiten?

Nein, der Arbeitsmodus mit Voyant ist die Einzelarbeit.

6. Sind meine Daten bei Voyant sicher?

Voyant liegt auf keinem gesicherten Server. Das Tool wurde in Kanada entwickelt, die entsprechenden Server befinden sich außerhalb der EU. Um Datenschutz und Sicherheit zu erhöhen, empfiehlt sich die Installation des VoyantServers, der offline und auf dem eigenen Rechner ausgeführt wird. Angaben zu der eigenen Person müssen weder für die Verwendung des VoyantServers noch für die Nutzung der webbasierten Version gemacht werden. Die IP-Adresse wird erhoben, um den Verlauf der Arbeitssitzungen nachvollziehen zu können. Zwecks Tooloptimierung, Bugentfernung und Ermittlung der Toolanwendungen wird Google Analytics verwendet. Sofern Daten für Forschungszwecke herangezogen werden, geschieht dies in anonymisierter wie aggregierter Form. Bei der Nutzung von Voyant werden somit Datenschutzrechte geltend gemacht, die nicht den europäischen Standards entsprechen. Auf diesen Sachverhalt wird in der Datenschutzerklärung mit dem Verweis auf die spezifischen Datenschutzrichtlinien von Google hingewiesen. Die bearbeitete Textgrundlage wird gespeichert, um eine weiterführende Bearbeitung zu ermöglichen. Der hierfür zufällig generierte Link kann zusätzlich mit einem Zugangspasswort versehen werden. Sofern ein Link mindestens einmal im Monat aufgerufen wird, kann die Speicherung der hinterlegten Daten andauern. Texte oder Textsammlungen können auf Nachfrage per E-Mail gelöscht werden.

Externe und weiterführende Links

- Voyant Github: https://web.archive.org/save/https://github.com/sgsinclair/Voyant (Letzter Zugriff: 18.06.2024)
- VoyantTools Help: https://web.archive.org/save/https://voyant-tools.org/docs/#!/guide (Letzter Zugriff: 18.06.2024)
- VoyantServer Information: https://web.archive.org/save/https://voyant-tools.org/docs/#!/guide/server (Letzter Zugriff: 18.06.2024)
- Voyant Sprachen: https://web.archive.org/save/https://voyant-tools.org/docs/#!/guide/languages (Letzter Zugriff: 18.06.2024)
- Voyant Standalone-Version – VoyantServer: https://web.archive.org/save/https://github.com/voyanttools/VoyantServer (Letzter Zugriff: 18.06.2024)
- VoyantTools Twitter: https://web.archive.org/save/https://twitter.com/VoyantTools (Letzter Zugriff: 18.06.2024)
- Voyant Webseite: https://web.archive.org/save/https://voyant-tools.org (Letzter Zugriff: 18.06.2024)

Bibliographie

Herrmann, J.B. 2012. Literatur rechnen. Ein Wiki zur digitalen Textanalyse. http://litre.uni-goettingen.de/index.php/Hauptseite (zugegriffen: 25. Oktober 2018).

Horstmann, Jan. 2024. Ressourcenbeitrag: TextGrid Repository. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 11. Bibliografie (29. November). doi: 10.48694/fortext.3794, https://fortext.net/ressourcen/textsammlungen/textgrid-repository.

Horstmann, Jan und Alexandra Kern. 2024. Ressourcenbeitrag: Deutsches Textarchiv (DTA). Hg. von Evelyn Gius. *forTEXT* 1, Nr. 11. Bibliografie (29. November). doi: 10.48694/fortext.3791, https://fortext.net/ressourcen/textsammlungen/deutsches-textarchiv-dta.

Horstmann, Jan und Jan-Erik Stange. 2024. Methodenbeitrag: Textvisualisierung. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 5. Textvisualisierung (7. August). doi: 10.48694/fortext.3772, https://fortext.net/routinen/methoden/textvisualisierung.

Kemann, Max. 2016. A Republic of Emails: What are the contents. https://www.maxkemman.nl/2016/11/a-republic-of-emails-what-are-the-contents/ (zugegriffen: 25. Oktober 2018).

Kühner, Janina. 2017. Fachdidaktisches Essay: Beispielhafte Konzeption einer Literaturunterrichtseinheit mit Voyant. *Skriptum* 6, Nr. 1: 41–57. http://www.skriptum-geschichte.de/fileadmin/user_upload/Ausgaben/2017/Heft_1/PDFs/essay_kuehner.pdf (zugegriffen: 1. November 2018).

Lechner, Renée, Urs Henning und Emil Müller. 2018. Distant Reading mit Voyant. *Web2-Unterricht*. https://web2-unterricht.ch/2018/05/distant-reading-mit-voyant/ (zugegriffen: 25. Oktober 2018).

Sample, Mark. 2013. no life no life no life no life: the 100,000,000,000,000 stanzas of House of Leaves of Grass. *@samplereality*. http://www.samplereality.com/2013/05/08/no-life-no-life-no-life-no-life-the-100000000000000-stanzas-of-house-of-leaves-of-grass/ (zugegriffen: 23. Oktober 2018).

Samsel, Laurie. 2018. Voyant Tools. *Music References Services Quarterly* 21, Nr. 3: 153–157. doi: 10.1080/10588167.2018.1496754, https://www.tandfonline.com/doi/pdf/10.1080/10588167.2018.1496754?needAccess=true (zugegriffen: 26. Oktober 2018).

Sinclair, Stefan und Geoffrey Rockwell. 2012. Teaching Computer-Assisted Text Analysis: Approaches to Learning New Methodologies. In: *Digital Humanities Pedagogy: Practices, Principles and Politics*, hg. von Brett D. Hirsch, 3: Digital Humanities Series. OpenBook Publishers. (zugegriffen: 1. November 2018).

Sinclair, Stéfan und Geoffrey Rockwell. 2016a. *Hermeneutica. Computer-Assisted Interpretations in the Humanities*. The MIT Press. (zugegriffen: 25. Oktober 2018).

———. 2016b. Text Analysis and Visualization: Making Meaning Count. In: *A New Companion to Digital Humanities*, hg. von Susan Schreibman, Ray Siemens, und John Unsworth, 274–290.

Glossar

Annotation Annotation beschreibt die manuelle oder automatische Hinzufügung von Zusatzinformationen zu einem Text. Die manuelle Annotation wird händisch durchgeführt, während die (teil-)automatisierte Annotation durch Machine-Learning-Verfahren durchgeführt wird. Ein klassisches Beispiel ist das automatisierte PoS-Tagging (Part-of-Speech-Tagging), welches oftmals als Grundlage (Preprocessing) für weitere Analysen wie Named Entity Recognition (NER) nötig ist. Annotationen können zudem deskriptiv oder analytisch sein.

Browser Mit Browser ist in der Regel ein Webbrowser gemeint, also ein Computerprogramm, mit dem das Anschauen, Navigieren auf, und Interagieren mit Webseiten möglich wird. Am häufigsten genutzt werden dafür Chrome, Firefox, Safari oder der Internet Explorer.

Browsersuchfunktion Um eine Browser-Suchfunktion durchzuführen, und beispielsweise eine Webseite auf bestimmte Suchbegriffe zu filtern, klicken Sie auf Ihrem Mac „cmd“ + „F“ und auf Ihrem Windows PC „Strg“ + „F“. In das sich öffnende Suchfenster tragen Sie Ihren Suchbegriff ein und die jeweils geöffnete Seite wird darauf hin durchsucht.

Close Reading Close Reading bezeichnet die sorgfältige Lektüre und Interpretation eines einzelnen oder weniger Texte. Close Reading ist in der digitalen Literaturwissenschaft außerdem mit der manuellen Annotation textueller Phänomene verbunden (vgl. auch Distant Reading als Gegenbegriff).

Commandline Die Commandline (engl. *command line interface* (CLI)), auch Kommandozeile, Konsole, Terminal oder Eingabeaufforderung genannt, ist die direkteste Methode zur Interaktion eines Menschen mit einem Computer. Programme ohne eine grafische Benutzeroberfläche (GUI) werden i. d. R. durch Texteingabe in die Commandline gesteuert. Um die Commandline zu öffnen, klicken Sie auf Ihrem Mac „cmd“ + „space“, geben „Terminal“ ein und doppelklicken auf das Suchergebnis. Bei Windows klicken Sie die Windowstaste + „R“, geben „cmd.exe“ ein und klicken Enter.

CSV CSV ist die englische Abkürzung für *Comma Separated Values*. Es handelt sich um ein Dateiformat zur einheitlichen Darstellung und Speicherung von einfach strukturierten Daten mit dem Kürzel `.csv`, sodass diese problemlos zwischen IT-Systemen ausgetauscht werden können. Dabei sind alle Daten zeilenweise angeordnet. Alle Zeilen wiederum sind in einzelne Datenfelder aufgeteilt, welche durch Trennzeichen wie Semikola oder Kommata getrennt werden können. In Programmen wie Excel können solche Textdateien als Tabelle angezeigt werden.

Data Mining Data Mining gehört zum Fachbereich Information Retrieval und bezieht sich auf die systematische Anwendung computergestützter Methoden, die darauf abzielt, in vorhandenen Datenbeständen Muster, Trends oder Zusammenhänge zu erkennen. Textbasierte Formen des Data Minings sind u. a. Text Mining, Web Mining und Opinion Mining.

Distant Reading Distant Reading ist ein Ansatz aus den digitalen Literaturwissenschaften, bei dem computationelle Verfahren auf häufig große Mengen an Textdaten angewandt werden, ohne dass die Texte selber gelesen werden. Meist stehen hier quantitative Analysen im Vordergrund, es lassen sich jedoch auch qualitative Metadaten quantitativ vergleichen. Als Gegenbegriff zu *Close Reading* wurde der Begriff insbesondere von Franco Moretti (2000) geprägt.

Feature Unter Features können Einzelfunktionen eines Tools verstanden werden, die beispielsweise komplexe

Funktionen wie die Visualisierung eines Textes als Wordcloud ermöglichen, oder auch kleinere Funktionseinheiten wie den Abgleich einzelner Spracheigenschaften (Properties) mit annotierten Beispieltexten darstellen.

GUI GUI steht für *Graphical User Interface* und bezeichnet eine grafische Benutzeroberfläche. Ein GUI ermöglicht es, Tools mithilfe von grafischen Schaltflächen zu bedienen, um somit beispielsweise den Umgang mit der Commandline zu umgehen.

HTML HTML steht für *Hypertext Markup Language* und ist eine textbasierte Auszeichnungssprache zur Strukturierung elektronischer Dokumente. HTML-Dokumente werden von Webbrowsern dargestellt und geben die Struktur und Online-Darstellung eines Textes vor. HTML-Dateien können außerdem zusätzliche Metainformationen enthalten, die auf einer Webseite selbst nicht ersichtlich sind.

Information Retrieval Die Teildisziplin der Informatik, das Information Retrieval, beschäftigt sich mit der computergestützten Suche und Erschließung komplexer Informationen in meist unstrukturierten Datensammlungen.

IP-Adresse Die Vernetzung von Computern wird in einem Internetprotokoll (IP) festgehalten, woraufhin jedes angebundene Gerät in diesem Computernetz eine IP-Adresse erhält. So werden die Geräte adressierbar und erreichbar gemacht. Die IP gehört zu den personenbezogenen Daten, da über sie auf Ihre Identität geschlossen werden kann.

Kollokation Als Kollokation bezeichnet man das häufige, gemeinsame Auftreten von Wörtern oder Wortpaaren in einem vordefinierten Textabschnitt.

Korpus Ein Textkorpus ist eine Sammlung von Texten. Korpora (Plural für „das Korpus“) sind typischerweise nach Textsorte, Epoche, Sprache oder Autor*in zusammengestellt.

Lemmatisieren Die Lemmatisierung von Textdaten gehört zu den wichtigen Preprocessing-Schritten in der Textverarbeitung. Dabei werden alle Wörter (Token) eines Textes auf ihre Grundform zurückgeführt. So werden beispielsweise Flexionsformen wie „schneller“ und „schnelle“ dem Lemma „schnell“ zugeordnet.

Machine Learning Machine Learning, bzw. maschinelles Lernen im Deutschen, ist ein Teilbereich der künstlichen Intelligenz. Auf Grundlage möglichst vieler (Text-)Daten erkennt und erlernt ein Computer die häufig sehr komplexen Muster und Gesetzmäßigkeiten bestimmter Phänomene. Daraufhin können die aus den Daten gewonnen Erkenntnisse verallgemeinert werden und für neue Problemlösungen oder für die Analyse von bisher unbekannten Daten verwendet werden.

Markup (Textauszeichung) Die Textauszeichnung (eng. *Markup*) fällt in den Bereich der Daten- bzw. Textverarbeitung, genauer in das Gebiet der Textformatierung, welche durch Auszeichnungssprachen wie XML implementiert wird. Dabei geht es um die Beschreibung, wie einzelne Elemente eines Textes beispielsweise auf Webseiten grafisch dargestellt werden sollen.

Markup Language Markup Language bezeichnet eine maschinenlesbare Auszeichnungssprache, wie z.B. HTML, zur Formatierung und Gliederung von Texten und anderen Daten. So werden beispielsweise auch Annotationen durch ihre Digitalisierung oder ihre digitale Erstellung zu Markup, indem sie den Inhalt eines Dokumentes strukturieren.

Metadaten Metadaten oder Metainformationen sind strukturierte Daten, die andere Daten beschreiben. Dabei kann zwischen administrativen (z. B. Zugriffsrechte, Lizenzierung), deskriptiven (z. B. Textsorte), strukturellen (z. B. Absätze oder Kapitel eines Textes) und technischen (z. B. digitale Auflösung, Material) Metadaten unterschieden werden. Auch Annotationen bzw. Markup sind Metadaten, da sie Daten/Informationen sind, die den eigentlichen Textdaten hinzugefügt werden und Informationen über die Merkmale der beschriebenen Daten liefern.

Named Entities Eine Named Entity (NE) ist eine Entität, oft ein Eigenname, die meist in Form einer Nominalphrase zu identifizieren ist. Named Entities können beispielsweise Personen wie „Nils Holgerson“, Organisationen wie „WHO“ oder Orte wie „New York“ sein. Named Entities können durch das Verfahren der Named Entity Recognition (NER) automatisiert ermittelt werden.

OCR OCR steht für *Optical Character Recognition* und bezeichnet die automatische Texterkennung von gedruckten Texten, d. h. ein Computer „liest“ ein eingescanntes Dokument, erkennt und erfasst den Text darin und generiert daraufhin eine elektronische Version.

Opinion Mininig Unter Opinion Mining, oder Sentiment Analysis, versteht man die Analyse von Stimmungen oder Haltungen gegenüber einem Thema, durch die Analyse natürlicher Sprache. Das Opinion Mining gehört zu den Verfahren des Text Minings.

PDF PDF steht für *Portable Document Format* . Es handelt sich um ein plattformunabhängiges Dateiformat, dessen Inhalt auf jedem Gerät und in jedem Programm originalgetreu wiedergegeben wird. PDF-Dateien können Bilddateien (z. B. Scans von Texten) oder computerlesbarer Text sein. Ein lesbares PDF ist entweder ein OCRter Scan oder ein am Computer erstellter Text.

POS PoS steht für *Part of Speech* , oder „Wortart“ auf Deutsch. Das PoS- Tagging beschreibt die (automatische) Erfassung und Kennzeichnung von Wortarten in einem Text und ist of ein wichtiger Preprocessing-Schritt, beispielsweise für die Analyse von Named Entities.

Preprocessing Für viele digitale Methoden müssen die zu analysierenden Texte vorab „bereinigt“ oder „vorbereitet“ werden. Für statistische Zwecke werden Texte bspw. häufig in gleich große Segmente unterteilt (*chunking*), Großbuchstaben werden in Kleinbuchstaben verwandelt oder Wörter werden lemmatisiert.

Property Property steht für „Eigenschaft“, „Komponente“ oder „Attribut“. In der automatischen Annotation dienen konkrete Worteigenschaften wie Groß- und Kleinschreibung zur Klassifizierung von Wörtern oder Phrasen. Durch die Berücksichtigung solcher Eigenschaften in den Features eines Tools kann maschinelles Lernen bestimmter Phänomene umgesetzt werden. In der manuellen Annotation können als Properties auch Eigenschaften von Annotationen benannt werden.

Query *Query* bedeutet „Abfrage“ oder „Frage“ und bezeichnet eine computergestützte Abfrage zur Analyse eines Textes. Um Datenbestände zu durchsuchen, werden Abfragesprachen eingesetzt, die *Queries* (Anfragen) an den Datenbestand senden. So bilden alle möglichen Queries zusammen die *Query Language* eines Tools.

Reintext-Version Die Reintext-Version ist die Version eines digitalen Textes oder einer Tabelle, in der keinerlei Formatierungen (Kursivierung, Metadatenauszeichnung etc.) enthalten sind. Reintext-Formate sind beispielsweise TXT, RTF und CSV.

Server Ein Server kann sowohl hard- als auch softwarebasiert sein. Ein hardwarebasierter Server ist ein Computer, der in ein Rechnernetz eingebunden ist und der so Ressourcen über ein Netzwerk zur Verfügung stellt. Ein softwarebasierter Server hingegen ist ein Programm, das einen spezifischen Service bietet, welcher von anderen Programmen (Clients) lokal oder über ein Netzwerk in Anspruch genommen wird.

Stoppwortliste Stoppwörter sind hochfrequente Wörter, meist Funktionswörter, die, aufgrund ihrer grammatisch bedingten Häufigkeit, beispielsweise die Ergebnisse von inhaltlichen oder thematischen Analysen verzerren können. Deshalb werden diese Wörter, gesammelt in einer Stoppwortliste, bei digitalen Textanalysen meist nicht berücksichtigt.

TEI Die *Text Encoding Initiative* (TEI) ist ein Konsortium, das gemeinsam einen Standard für die Darstellung von Texten in digitaler Form entwickelt. Die TEI bietet beispielsweise Standards zur Kodierung von gedruckten Werken und zur Auszeichnung von sprachlichen Informationen in maschinenlesbaren Texten (siehe auch XML und Markup).

Text Mining Das Text Mining ist eine textbasierte Form des Data Minings. Prozesse & Methoden, computergestützt und automatisch Informationen bzw. Wissen aus unstrukturierten Textdaten zu extrahieren, werden als Text Mining zusammengefasst.

Type/Token Das Begriffspaar „Type/Token“ wird grundsätzlich zur Unterscheidung von einzelnen Vorkommnissen (Token) und Typen (Types) von Wörtern oder Äußerungen in Texten genutzt. Ein Token ist also ein konkretes Exemplar eines bestimmten Typs, während ein Typ eine im Prinzip unbegrenzte Menge von Exemplaren (Token) umfasst.
Es gibt allerdings etwas divergierende Definitionen zur Type-Token-Unterscheidung. Eine präzise Definition ist daher immer erstrebenswert. Der Satz „Ein Bär ist ein Bär.“ beinhaltet beispielsweise fünf Worttoken („Ein“, „Bär“, „ist“, „ein“, „Bär“) und drei Types, nämlich: „ein“, „Bär“, „ist“. Allerdings könnten auch vier Types, „Ein“, „ein“, „Bär“ und „ist“, als solche identifiziert werden, wenn Großbuchstaben beachtet werden.

URI *Uniform Resource Identifier* (URI) ist ein Identifikator zur eindeutigen Erkennung von Online-Ressourcen wie Webseiten. Im „Raum“ des Internets können so alle Inhalte eindeutig identifiziert werden, unabhängig davon, ob es sich dabei beispielsweise um eine Seite mit Text oder Video handelt. Die am häufigsten verwendete Form eines URI ist die Webseitenadresse, die URL.

Web Mining Unter Web Mining versteht man die Anwendung von Techniken des Data Mining zur Extraktion von Informationen aus dem World Wide Web. Das Web Mining ist ein Teilbereich des Data Minings und zählt zu einem der wichtigsten Anwendungsgebiete für das Text Mining.

Webanwendung Eine webbasierte Anwendung ist ein Anwendungsprogramm, welches eine Webseite als Schnittstelle oder Front-End verwendet. Im Gegensatz zu klassischen Desktopanwendungen werden diese nicht lokal auf dem Rechner der Nutzer*innen installiert, sondern können von jedem Computer über einen Webbrowser „online“ genutzt werden. Webanwendungen erfordern daher kein spezielles Betriebssystem.

Wordcloud Eine *Wordcloud* , oder auch Schlagwortwolke, ist eine Form der Informationsvisualisierung, beispielsweise von Worthäufigkeiten in einem Text oder einer Textsammlung. Dabei werden unterschiedlich gewichtete Wörter, wie die häufigsten Wörter, i.d.R. größer oder auf andere Weise hervorgehoben dargestellt. Die horizontale/vertikale Ausrichtung und die Farbe der dargestellten Wörter hat meistens allerdings keinen semantischen Mehrwert.

XML XML steht für *Extensible Markup Language* und ist eine Form von Markup Language, die sowohl computer- als auch menschenlesbar und hochgradig anpassbar ist. Dabei werden Textdateien hierarchisch strukturiert dargestellt und Zusatzinformationen i. d. R. in einer anderen Farbe als der eigentliche (schwarz gedruckte) Text dargestellt. Eine standardisierte Form von XML ist das TEI-XML.

ZIP ZIP steht für ein Dateiformat (zip = engl. Reißverschluss), in welchem mehrere Einzeldateien verlustfrei, komprimiert zusammengefasst werden. ZIP-Dateien werden beim Öffnen entweder automatisch entpackt oder lassen sich per Rechtsklick extrahieren.