Lehrmodul: Topic Modeling mit dem DARIAH Topics Explorer lehren

Jan Horstmann [1]

1. Universität Münster

Thema:	Topic Modeling	DOI:	10.48694/fortext.3716
Jahrgang:	1	Ausgabe:	8
Erscheinungsdatum:	2024-07-10	Erstveröffentlichung:	2019-07-22 auf fortext.net
Lizenz:			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.

Eckdaten des Lehrmoduls

- Thema der Sitzung: Themen und Topics bei Friedrich Schiller und Wilhelm Hauff
- Lernziele: Kenntnisse über die Methode des Topic Modeling, sicherer Umgang mit dem DARIAH Topics Explorer, kritische Bewertung der Methode, Autoren- und Epochenkenntnisse (Sturm und Drang, Weimarer Klassik, Romantik)
- Phasen: Einführende Begriffsdiskussion (Themen vs. Topics), Vorstellung und Diskussion der Methode, Demonstration der Toolfunktionen, Gruppenarbeit, Gruppenpräsentationen
- Sozialformen: Diskussion, Vortrag, Gruppenarbeit
- Medien/Materialien: Alle Lernenden müssen einen Laptop haben, auf dem der DARIAH Topics Explorer installiert ist; Lehrende benötigen einen Laptop und Beamer
- Dauer des Lehrmoduls: 2 x 90 Minuten
- Schwierigkeitsgrad des Tools: leicht bis mittel

Bausteine

- Verlaufsraster des Lehrmoduls: Aus welchen Phasen setzt sich das Lehrmodul zusammen? Dem Verlaufsplan entnehmen Sie Inhalte und Schwerpunkte.
- Anwendungsbeispiel: Anhand welcher Texte unterrichten Sie Topic Modeling? Leiten Sie die Studierenden dazu an, Themen in den Dramen Friedrich Schillers und den Prosawerken Wilhelm Hauffs zu explorieren.
- Verlauf der Unterrichtseinheit(en): Wie sieht die konkrete Ausgestaltung der Phasen aus und welche Arbeitsschritte werden vorgenommen? Erfahren Sie, wie die Unterrichtseinheit strukturiert ist und welche Beispielaufgaben Sie Ihren Studierenden stellen können.
- Lösungen zu den Beispielaufgaben: Hat die Lerngruppe die Beispielaufgaben richtig gelöst? Hier erfahren Sie, wie Sie die Antworten erhalten.

Alle Materialien zu dieser Sitzung stellen wir Ihnen auf Zenodo zum Download bereit (forTEXT 2019e).

Verlaufraster des Lehrmoduls

Phase	Impulse des/der Lehrenden	Erwartete Aktivität der Lernenden	Sozialform	Medien / Materialien
Vorab und Einstieg (*ca. 15 Minuten*)	Was unterscheidet die Begriffe Thema, Stoff, Motiv, Topos und Topic?	Vorab: Methodeneintrag Topic Modeling (Horstmann 2024a) und Lerneinheit Topic Modeling mit dem DARIAH Topics Explorer (Horstmann 2024b); Beteiligung an der Diskussion	Diskussion im Plenum	Beamer, Laptop
Problematisierung (*ca. 15 Minuten*)	Warum ist ein Topic kein Thema? Wie funktioniert Topic Modeling?	Beteiligung an der Diskussion; Rückbezug auf Methodeneintrag	Diskussion im Plenum	Beamer, Laptop
Erarbeitung (*ca. 60 Minuten*)	Vorstellung der Toolfunktionen; Betreuung der Kleingruppen	Hands-on Topic Modeling in Kleingruppen; Vorbereitung einer Präsentation als Hausaufgabe	Lehrvortrag und Gruppenarbeit	Beamer, Laptop, DARIAH Topics Explorer, zwei Korpora, zwei Stoppwortlisten
Sicherung (*ca. 50 Minuten*)	Betreuung der Gruppen	gegenseitige Präsentation der Gruppenarbeitsergebnisse	Gruppenarbeit	Laptops, DARIAH Topics Explorer, zwei Korpora, zwei Stoppwortlisten
Reflexion & Transfer (*ca. 40 Minuten*)	Diskussion von Schwierigkeiten, Impulse für Transfer geben	Ergebnisse und Schwierigkeiten aus den Gruppenpräsentationen diskutieren	Diskussion im Plenum	Beamer, Laptops

Das Verlaufsraster steht als PDF-Datei zum Download auf Zotero (forTEXT 2019e) zur Verfügung.

1. Anwendungsbeispiel

In zwei Seminarsitzungen werden Sie die Methode Topic Modeling (Horstmann 2024a) lehren, die ein probabilistisches Verfahren zur thematischen Exploration größerer Textsammlungen ist. Die Studierenden probieren die Methode an einem Dramenkorpus (vgl. Korpus) von Friedrich Schiller und einer Sammlung von Erzählungen Wilhelm Hauffs mithilfe des DARIAH Topics Explorers (Schumacher 2024) praktisch in Kleingruppen aus und präsentieren ihre Ergebnisse.

2. Verlauf der Unterrichtseinheiten

2.1 Vorarbeiten

Die Studierenden sollten vorab den Methodeneintrag Topic Modeling (Horstmann 2024a) und den Tooleintrag DARIAH Topics Explorer (Schumacher 2024) gelesen, sowie die Lerneinheit Topic Modeling mit dem DARIAH Topics Explorer (Horstmann 2024b) durchgearbeitet haben. Hilfreich wäre es zudem, sie die Tutorialvideos Topic Modeling und Literaturanalyse (forTEXT 2019b; forTEXT 2019a; forTEXT 2019c) sowie die Fallstudie „Themen von Autoren und Autorinnen der Literatur des 19. Jahrhunderts“ schauen zu lassen (forTEXT 2019d). Im Zuge der Lerneinheit wird das Tool bereits installiert, sodass etwaige technische Schwierigkeiten bereits vor der Sitzung adressiert und behoben werden können. Außerdem sollten Sie Ihren Studierenden vorab das vom TextGrid Repository (Horstmann 2024c) stammende Schillerkorpus und das Hauffkorpus sowie die von uns vorbereitete Stoppwortliste für das Schillerkorpus und die Stoppwortliste für das Hauffkorpus auf der eLearning-Plattform Ihrer Institution zur Verfügung stellen. Die Materialien erhalten Sie auf Zenodo (forTEXT 2019e). Möchten Sie mit anderen digitalen Textsammlungen arbeiten, können Sie sich diese bspw. im TextGrid Repository (Horstmann

2024c) oder im Deutschen Textarchiv (DTA) (Horstmann und Kern 2024) selbst zusammenstellen. Das Korpus muss mindestens zehn Texte enthalten, damit der Topics Explorer arbeiten kann, und die Texte sollten alle das gleiche Format haben. Der Topics Explorer kann TXT- (vgl. Reintext-Version) und TEI-XML-Dateien (vgl. TEI; XML) verarbeiten. Die Beispielaufgaben in diesem Lehrmodul beziehen sich auf das Schiller- und das Hauffkorpus. Sollten Sie Ihre Seminarsitzungen nicht in einem Computerlabor abhalten, erinnern Sie Ihre Studierenden daran, einen eigenen Laptop mitzubringen. Eine 1:1-Ausstattung ist sinnvoll, da sich der Umgang mit dem Topics Explorer durch eigenhändiges Ausprobieren am besten vermitteln lässt. Es sollten auf keinen Fall mehr als zwei Studierende an einem Gerät arbeiten.
Je nachdem, welchen Lehrstil Sie persönlich präferieren, kann es sinnvoll sein, für den Einstieg in die Stunde Folien vorzubereiten, die die Diskussion über Topics und Themen mit Anschauungsbeispielen und wichtigen Schlagwörtern befördern. Wir haben Ihnen beispielhaft ein paar Folien für den Einstieg entworfen, die Sie nutzen oder weiter ausarbeiten können. Auch diese Folien erhalten Sie auf Zenodo (forTEXT 2019e).

2.2 Einstieg und Problematisierung

Um den beiden Unterrichtseinheiten (die vermutlich mit einer Woche Unterbrechung abgehalten werden) einen inneren Zusammenhang zu verleihen, sollten Sie zu Beginn transparent machen, welche Lernziele in den beiden Sitzungen erreicht werden sollen und was die Lerngruppe in den einzelnen Sitzungen erwartet.

- Es sollen Kenntnisse über die Methode des Topic Modeling und ein sicherer Umgang mit dem DARIAH Topics Explorer vermittelt werden.
- Die Studierenden sollen einerseits die Methode kritisch bewerten, andererseits aber auch Autoren- und Epochenkenntnisse (über Sturm und Drang, Weimarer Klassik und Romantik) erwerben. Sie können hierbei selbst Schwerpunkte setzen.

In einem literaturwissenschaftlichen Seminar bietet es sich an, nach Nennung der Lernziele mit einem Rückbezug auf die fachliche Tradition zu beginnen. Auf diese Weise werden literaturwissenschaftliche Fachkompetenzen vermittelt und Vorwissen aktiviert. Beim Topic Modeling kann dies über die Diskussion von Fachbegriffen geschehen. Sie starten mit einer aktivierenden Diskussion: Fragen Sie die Studierenden nach den Themen der bereits im Seminar behandelten Primärtexte oder nach den Themen von Texten, die Sie vorab ausgewählt haben. Sinnvoll ist es hierbei, Texte aus unterschiedlichen Epochen und Gattungen auszuwählen, die jedoch das gleiche Thema behandeln (so ist z. B. das Thema „Streben nach Macht“ in so unterschiedlichen Werken wie Schillers *Maria Stuart* als auch in Martins *A Song of Ice and Fire* zentral). Fragen Sie dann nach den Begriffen Stoff, Motiv und Topos in Verbindung mit den von Ihnen ausgewählten Texten und kontrastieren in der Diskussion die Begriffe. Dabei sollte deutlich werden, dass die Begriffe zwar unterschiedliche Aspekte von textlichen Inhalten beschreiben, die Begriffsverwendung auch in der Fachwissenschaft jedoch nicht einheitlich ist.
Auf Grundlage dieser Begrifflichkeiten diskutieren Sie die Dimensionen des Topic-Begriffs und mit Rückbezug auf den Methodeneintrag Topic Modeling (Horstmann 2024a) die grundsätzlichen Annahmen, die diesem Verfahren zugrunde liegen. In dieser Phase der Unterrichtseinheit können Sie zudem Beispiele von Topics zeigen, die auf bestimmten Texten oder Textsammlungen basieren (die im Methodeneintrag angegebenen Sekundärtexte und auch unsere Einstiegsfolien liefern Ihnen hierfür bei Bedarf Material). Geben Sie den Studierenden im Rahmen der ersten Unterrichtsphase die Gelegenheit, Fragen zu stellen und Schwierigkeiten, die sie mit dem Methodeneintrag hatten, zu diskutieren. Nach etwa einer halben Stunde sollten Sie langsam in die Erarbeitungsphase überleiten, indem Sie z. B. explizit nach den Erfahrungen fragen, welche die Studierenden mit der Lerneinheit Topic Modeling mit dem DARIAH Topics Explorer (Horstmann 2024b) hatten.

2.3 Erarbeitung

Mit Rückbezug auf die angesprochenen Schwierigkeiten aus der Lerneinheit führen Sie die Studierenden mit Unterstützung des Beamers einmal an Ihrem eigenen Gerät durch die einzelnen Module des DARIAH Topics Explorers. Geben Sie den Studierenden auch hierbei die Möglichkeit, Verständnisschwierigkeiten anzusprechen, um diese im Plenum beheben zu können. Dieser Vortrag sollte maximal 15 Minuten in Anspruch nehmen, damit für die anschließende Arbeit in Kleingruppen noch genügend Zeit bleibt.
Teilen Sie nun die Studierenden in Kleingruppen von 2-3 Personen ein. Größere Gruppen sind bei der gemeinsamen Arbeit mit digitalen Tools nur sinnvoll, wenn diese Tools selbst eine Funktion für kollaboratives Arbeiten haben. Da dies beim DARIAH Topics Explorer nicht der Fall ist, sollten möglichst alle Gruppenmitglieder an ihrem eigenen Laptop arbeiten, wobei gruppenintern trotzdem im Peer-to-Peer-Modus zusammengearbeitet wird. Die eine Hälfte dieser Kleingruppen wird sich im Folgenden mit dem Schillerkorpus, die andere mit dem Hauffkorpus auseinandersetzen. Achten Sie auf eine ausgewogene Aufteilung.

Geben Sie den Gruppen folgende Arbeitsaufträge an die Hand:

Aufgabe 1: Welches Dokument ist das umfangreichste in Ihrem Korpus?

Aufgabe 2: Bei welchen Einstellungen (Topicmenge und Iterationen) erhalten Sie Topics, die konsistente Themen abbilden? Stimmt die algorithmische Angabe (Log-likelihood) mit Ihrem persönlichen Eindruck überein?

Aufgabe 3: Schauen Sie sich die bereitgestellte Stoppwortliste an. Welche interpretatorischen Vorannahmen werden durch den Ausschluss der dort aufgeführten Wörter getroffen? Warum werden dort bspw. Figurennamen aufgeführt? Betrachten Sie auch die vom DARIAH Topics Explorer bereitgestellte Stoppwortliste für deutsche Texte. Welche der dort aufgeführten Wörter sollten in Bezug auf Schiller und/oder Hauff nicht „gestoppt" werden?

Aufgabe 4: Finden Sie Topics, die besonders dokumentspezifisch sind oder Dokumente, die ein bestimmtes Topic besonders stark abbilden? Was könnten Gründe dafür sein und was sind Vor- oder Nachteile einer solchen Verteilung?

Aufgabe 5: Woran können Sie anhand Ihrer Topics erkennen, welcher Gattung Ihr Korpus zugehört?

Aufgabe 6: Modellieren Sie ein Topic zum Thema Liebe und eins zum Thema Natur. Finden Sie hier autoren- oder gar epochentypische Konstellationen wieder?

Stehen Sie den Kleingruppen während der Erarbeitungsphase für Nachfragen zur Verfügung bzw. verschaffen Sie sich einen eigenen Überblick über den Stand der Gruppenarbeiten. Der erste Teil des Lehrmoduls endet nun. Zu diesem Zeitpunkt haben die Studierenden sich zunächst im Rahmen einer Diskussion mit den literaturwissenschaftlichen Fachinhalten „Thema" und „Topic" auseinandergesetzt. Darüber hinaus haben sie die Methode des Topic Modeling kennengelernt und anhand des DARIAH Topic Explorers selber ausprobiert. Beenden Sie die Sitzung, indem Sie auf Inhalte der nächsten Seminareinheit verweisen. Bis zur nächsten Sitzung sollen die Gruppen die Fragen beantwortet und eine ca. 10-minütige Präsentation mit Anschauungsmaterial (Exportfunktionen des Topics Explorers) sowie eine ausformulierte Musterlösung zu den Aufgaben vorbereitet haben. Die Erarbeitungsphase endet damit nicht mit dem Schluss der ersten Sitzung, sondern bildet ebenso die Grundlage der Hausaufgaben.

2.4 Sicherung

Die zweite Sitzung beginnt mit den Präsentationen der Arbeitsergebnisse der Kleingruppen. Diese finden nicht im Plenum statt, um eine zeitintensive und mehrfache Vorstellung ähnlicher oder gleicher Arbeitsergebnisse zu umgehen. Jeweils zwei Kleingruppen, die sich mit Schiller beschäftigt haben, und zwei Kleingruppen, die sich mit Hauff beschäftigt haben, bilden Teams und stellen sich ihre Ergebnisse gegenseitig vor (ca. 20 Minuten). Dabei sollen die Studierenden Erkenntnisse, Herausforderungen und Problemstellungen diskutieren, um eine gemeinsame Kurzpräsentation und Musterlösung zu erarbeiten. Die Methode Topic Modeling sollte in den neu zusammengesetzten Gruppen kritisch reflektiert werden, wobei die in der Erarbeitungsphase ausgeführten Aufgaben als Rahmen für diese Diskussionen fungieren.
Sie selbst gehen in dieser Phase von Gruppe zu Gruppe, hören bei den Diskussionen zu und geben Impulse. Diese Impulse sollten sowohl traditionell literaturwissenschaftlicher Art sein (Epochen-, Gattungs- und Autorenkenntnisse sind hier besonders relevant) als auch vor dem Hintergrund der Ihnen zur Verfügung gestellten Lösungsvorschläge für die Beispielaufgaben gegeben werden. Achten Sie darauf, dass in allen Gruppen eine Ergebnissicherung stattfindet – inhaltlicher Art wie auch in Bezug auf exportierte Daten aus dem Topics Explorer. Für die Gruppendiskussionen sollten Sie insgesamt ca. 50 Minuten veranschlagen.

2.5 Transfer & Reflexion

In den letzten 40 Minuten der Sitzung sollten schließlich die in der Sicherungsphase zu Tage getretenen Probleme und Erkenntnisse ins Plenum getragen und zusammenfassend diskutiert werden. Auf freiwilliger Basis sollte sich mindestens eine Schiller- und eine Hauff-Gruppe bereit erklären, ihre Ergebnisse zu präsentieren, sodass die anderen Gruppen kritisch dazu Stellung nehmen können. Die kurzen Vorträge (ca. 10 Minuten) sollten exportierte Grafiken enthalten und über Erwartungen und Überraschungen im Prozess des Topic Modelings informieren. Welche methodischen wie literaturhistorischen und gattungsspezifischen Erkenntnisse haben die Studierenden gewonnen? Was sind Gemeinsamkeiten und Unterschiede im Schiller- und Hauff-Korpus? Was wären alternative bzw. traditionelle Wege gewesen, um diese Erkenntnisse zu erlangen? Welche Textsammlungen würden die Studierenden außerdem gerne mit dem Topics Explorer untersuchen? Wenn die Diskussion sich zäh gestalten sollte, können Sie weitere Gruppen auffordern, ihre Ergebnisse zu präsentieren – auch wenn es hierbei stellenweise zu Wiederholungen kommen kann. Die erarbeiteten Musterlösungen können Sie zudem als Seminarleistung einsammeln.
Abschließend geben Sie Impulse für einen Transfer der erlernten Methodik. Fragen Sie, wie die Methode in etwaigen zu erstellenden Hausarbeiten oder in Bezug auf andere im Seminar behandelte Werke oder Fragestellungen fruchtbar gemacht werden kann. Betonen Sie dabei die Relevanz einer Kombination von traditionellem Fachwissen mit der neu erlernten Methode. Ein Distant Reading-Verfahren wie das Topic Modeling kann z. B. den Blick auf einen Autor/eine Autorin, eine Epoche oder eine Gattung schärfen und erweitern. Kenntnisse,

die man bei der intensiven – bspw. diskursanalytisch ausgerichteten – Lektüre eines Textes erwirbt, können durch Methoden des Distant Reading in einem größeren Kontext gespiegelt, verfeinert, evtl. sogar revidiert werden. Der Computer gibt nicht auf Knopfdruck literaturwissenschaftliches Fachwissen aus; vielmehr sollte der interpretatorisch überformte Modellierungsprozess im Topic Modeling als Pendant und mithin als Ergänzung (vgl. Scalable Reading) zum traditionellen Close Reading gesehen werden.

3. Lösungen zu den Beispielaufgaben

Die Lösungen zu dieser Einheit erhalten Sie auf Zenodo (forTEXT 2019e).

Externe und weiterführende Links

- Materialien zum Lehrmodul auf Zenodo: https://zenodo.org/records/12530205 (Letzter Zugriff: 19.09.2024)

Bibliographie

forTEXT. 2019a. Tutorial: DARIAH Topics Explorer installieren. Topic Modeling und Literaturanalyse. 11. Februar. doi: 10.5281/zenodo.10371074, https://doi.org/10.5281/zenodo.10371074.

———. 2019b. Tutorial: DARIAH Topics Explorer zur Literaturanalyse nutzen. Topic Modeling und Literaturanalyse. 25. Februar. doi: 10.5281/zenodo.10372228, https://doi.org/10.5281/zenodo.10372228.

———. 2019c. Tutorial: Drei Methoden für bessere Topics beim Topic Modeling. Topic Modeling und Literaturanalyse. 11. März. doi: 10.5281/zenodo.10378213, https://doi.org/10.5281/zenodo.10378213.

———. 2019d. Themen von Autoren und Autorinnen der Literatur des 19. Jahrhunderts. 25. März. doi: 10.5281/zenodo.10276975, https://doi.org/10.5281/zenodo.10276975.

———. 2019e. Topic Modeling mit dem DARIAH Topics Explorer lehren. 22. Juli. doi: 10.5281/zenodo.10518659, https://zenodo.org/records/12530205.

Horstmann, Jan. 2024b. Lerneinheit: Topic Modeling mit dem DARIAH Topics Explorer. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 8. Topic Modeling (7. Oktober). doi: 10.48694/fortext.3729, https://fortext.net/routinen/lerneinheiten/topic-modeling-mit-dem-dariah-topics-explorer.

———. 2024a. Methodenbeitrag: Topic Modeling. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 8. Topic Modeling (7. Oktober). doi: 10.48694/fortext.3717, https://fortext.net/routinen/methoden/topic-modeling.

———. 2024c. Ressourcenbeitrag: TextGrid Repository. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 11. Bibliografie (30. 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 (30. November). doi: 10.48694/fortext.3791, https://fortext.net/ressourcen/textsammlungen/deutsches-textarchiv-dta.

Schumacher, Mareike. 2024. Toolbeitrag: DARIAH Topics Explorer. Hg. von Evelyn Gius. *forTEXT* 1, Nr. 8. Topic Modeling (7. Oktober). doi: 10.48694/fortext.3728, https://fortext.net/tools/tools/dariah-topics-explorer.

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.

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).

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.

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.

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.

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

LDA LDA steht für *Latent Dirichlet Allocation* und ist ein generatives, statistisches Wahrscheinlichkeitsmodell, welches zum Topic Modeling angewendet werden kann. Bei der LDA werden auf Grundlage eines Wahrscheinlichkeitsmodells Wortgruppen aus Textdokumenten erstellt. Dabei wird jedes Dokument als eine Mischung von verborgenen Themen betrachtet und jedes Wort einem Thema zugeordnet. Wortreihenfolgen und Satzzusammenhänge spielen dabei keine Rolle.

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.

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.

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.

Scalable Reading Die Kombination aus Distant Reading- und Close Reading-Methoden, angewandt auf einen Untersuchungsgegenstand, wird als Scalable Reading bezeichnet.

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).

Topic Modeling Das Topic Modeling ist ein statistisches, auf Wahrscheinlichkeitsrechnung basierendes, Verfahren zur thematischen Exploration größerer Textsammlungen. Das Verfahren erzeugt „Topics" zur Abbildung häufig gemeinsam vorkommender Wörter in einem Text. Für die Durchführung können verschiedene Algorithmen und Modelle wie das LDA verwendet werden.

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.

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.