Lerneinheit: Topic Modeling mit dem DARIAH Topics Explorer

Jan Horstmann [1]

1. Universität Münster

Thema:	Topic Modeling	DOI:	10.48694/fortext.3729
Jahrgang:	1	Ausgabe:	8
Erscheinungsdatum:	2024-07-10	Erstveröffentlichung:	2019-01-21 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 der Lerneinheit

- Anwendungsbezug: Hans Christian Andersens Märchen
- Methode: Topic Modeling in Prosatexten eines Autors
- Angewendetes Tool: DARIAH Topics Explorer
- Lernziele: Zusammenstellung und thematische Exploration einer kleinen bis mittelgroßen Textsammlung (vgl. Korpus), Installation des Tools, Auswertung/Interpretation der Ergebnisse
- Dauer der Lerneinheit: ca. 90 Minuten
- Schwierigkeitsgrad des Tools: leicht bis mittel

Bausteine

- Anwendungsbeispiel: Welche Textsammlung wird exploriert? Analysieren Sie digital die Themen in einer Auswahl von Hans Christian Andersens Märchen.
- Vorarbeiten: Was müssen Sie tun, bevor es losgehen kann? Lernen Sie, wie man den Topics Explorer installiert, wie die Textsammlung zusammengestellt und in das Tool hineingeladen wird.
- Funktionen: Welche Funktionen bietet Ihnen der Topics Explorer zur thematischen Exploration von Textsammlungen? Lernen Sie die einzelnen Module des Tools kennen und lösen Sie Beispielaufgaben.
- Lösungen zu den Beispielaufgaben: Haben Sie die Beispielaufgaben richtig gelöst? Hier finden Sie Antworten.

1. Anwendungsbeispiel

Wir werden in dieser Lerneinheit eine Sammlung von 46 Märchen des dänischen Schriftstellers Hans Christian Andersen thematisch explorieren. Topic Modeling ist ein auf Wahrscheinlichkeitsrechnung basierendes Verfahren zur Exploration größerer Textsammlungen (vgl. Distant Reading). Das Verfahren erzeugt statistische Modelle (sog. *Topics*) zur Abbildung häufiger gemeinsamer Vorkommnisse von Wörtern (vgl. Kollokation; mehr dazu siehe Topic Modeling (Horstmann 2024a)). In dieser Lerneinheit verwenden wir dafür den DARIAH Topics Explorer (Schumacher 2024a), ein 2018 erschienenes Tool, das sich noch im Prototyp-Status befindet. Als Prototyp kann es noch keine sehr großen Textsammlungen verarbeiten, es bietet jedoch den Vorteil einer grafischen Benutzeroberfläche (vgl. GUI) mit interaktiven Schaltflächen.

2. Vorarbeiten

Hinweis: Mit der Veröffentlichung von macOS 10.15 Catalina hat Apple neue Sicherheitsfunktionen eingeführt, die zu Problemen beim Starten des TopicsExplorer führen. Als vorübergehende Abhilfe folgen Sie den Anweisungen auf dieser Webseite, um die Anwendung aus dem Quellcode zu installieren.

Zunächst stellen Sie sich Ihre digitale Textsammlung selbst zusammen. Für den Topics Explorer benötigen Sie entweder Dateien im TXT-Format (vgl. Reintext-Version) oder im XML-Format. Das TextGrid Repository (Horstmann 2024b) bietet eine umfangreiche Sammlung von verlässlich edierten Textdigitalisaten, darunter auch viele Texte Hans Christian Andersens. Gehen Sie auf die Seite des Repositorys und rufen Sie rechts oben unter „Explore“ die Autorenliste auf (siehe Abb. 1):

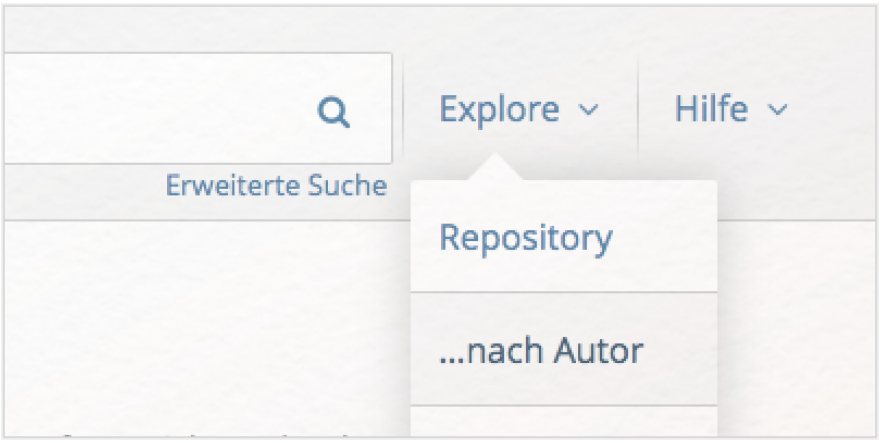

Abb. 1: TextGrid-Browser mit Explore-Funktion

Klicken Sie dann auf „Andersen, Hans Christian“. Auf der linken Seite der sich nun öffnenden Seite klicken Sie dann bei „Dateityp“ auf „text/xml“:

Abb. 2: TextGrid-Browser und Dateitypen

Nun haben Sie bereits eine Liste von 47 Texten: 46 Märchen und eine Biografie Andersens. Rechts oben unter „Alles herunterladen“ können Sie sich diese Sammlung jetzt vollständig als komprimierte ZIP-Datei herunterladen und diese dann in Ihrem Downloadordner entpacken (bzw. bei Windows per Rechtsklick „extrahieren“):

Abb. 3: ZIP-Datei herunterladen

TextGrid benennt die heruntergeladenen Dateiordner immer nach dem ersten Dokument und fügt ein „etc“ hinzu, Ihr Ordner wird also vermutlich „Der_standhafte_Zinnsoldat_etc“ heißen, insofern Sie die Anordnung

der Dateien in TextGrid zuvor nicht verändert haben.
Wenn Sie den Ordner öffnen, werden Sie feststellen, dass er sowohl XML als auch XML.meta-Dateien (vgl. Metadaten) enthält. Für unsere Lerneinheit interessieren uns ausschließlich die XML-Dateien. Sie sollten in Ihrer Ordnerübersicht daher die Ansicht geordnet nach „Art“ einstellen (bzw. bei Windows nach „Typ“):

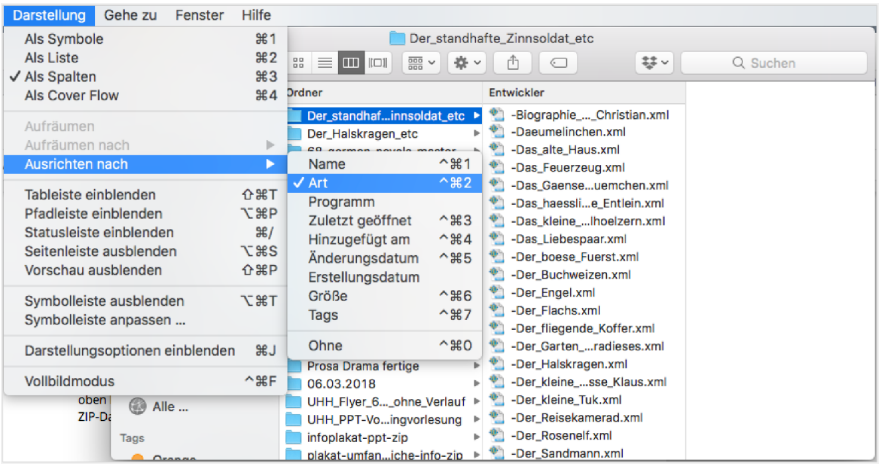

Abb. 4: XML-Dateien auswählen

So werden Ihnen alle XML-Dateien untereinander angezeigt, was die später erfolgende Textauswahl aus dem Topics Explorer heraus deutlich erleichtert.
Die Texte müssen für den Topics Explorer keinem weiteren Preprocessing unterzogen werden. Je nach wissenschaftlicher Fragestellung kann es sich jedoch anbieten, die Texte zu lemmatisieren (vgl. Lemmatisieren) (sodass alle Formen eines Worttyps als gleiches Wort behandelt werden), mithilfe einer Named Entity Recognition (Schumacher 2024b) alle Eigennamen auszuschließen, oder durch eine Part-of-Speech-Annotation (vgl. POS; Annotation) nur eine bestimmte Wortart für die Topics zu nutzen (z.B. Substantive). Da wir in dieser Lerneinheit keine so dezidierte Fragestellung haben, sondern als Erstkontakt mit Methode und Tool Andersens Märchen lediglich thematisch explorieren wollen, überspringen wir diese Schritte jedoch – Ergebnisse eines Topic Modelings müssen jedoch immer vor dem Hintergrund solcher Entscheidungen interpretiert werden.
Folgen Sie diesem Link, um sich den Topics Explorer für Ihr jeweiliges Betriebssystem (Windows, Mac oder Linux) herunterzuladen. Diese Lerneinheit arbeitet mit Version 2.0 des Topic-Explorer-Prototypen. Nach dem Entpacken der heruntergeladenen ZIP-Datei, findet sich in Ihrem Downloadordner ein Ordner mit dem Namen „dariah-topics-explorer-2.0-[Betriebssystem]“, in dem Sie das Tool finden. Mit einem Doppelklick starten sie das Tool direkt. Beim ersten Ausführen müssen Sie in Ihren Sicherheitseinstellungen vermutlich einmalig explizit das Öffnen des Programms erlauben. Sollten Sie sich unsicher sein, wie das funktioniert, schauen Sie sich unsere Videoerklärungen auf Zenodo an (für MacOS (forTEXT 2019a), für Windows (forTEXT 2019b)).

Die Startseite des Topics Explorers sieht nach dem Öffnen folgendermaßen aus:

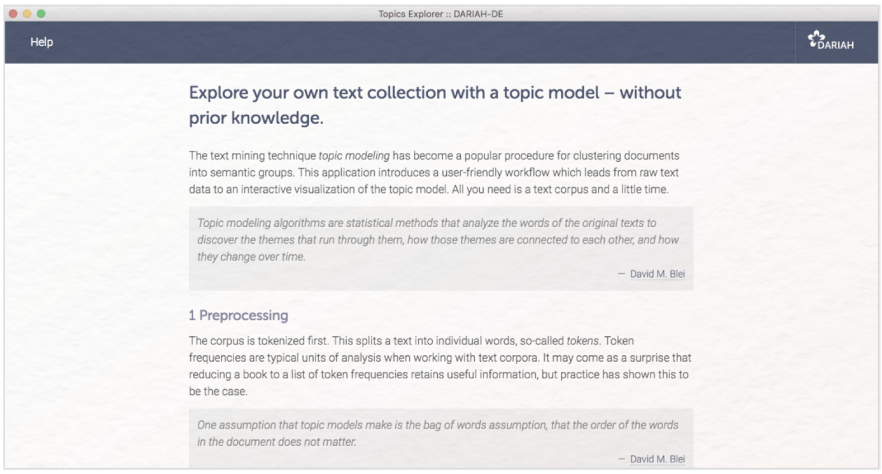

Abb. 5: TopicsExplorer Startseite

Lesen Sie den Text auf der Startseite und klicken Sie, wenn Sie soweit sind, auf die Schaltfläche „Dateien auswählen“, navigieren Sie zu Ihrem Ordner mit den Andersen-Texten und wählen Sie alle XML-Dateien aus. Da wir nur die Märchen von Andersen explorieren wollen, schließen wir die Datei „-Biographie_Andersen[...]“ in diesem Schritt aus (siehe Abb. 6).

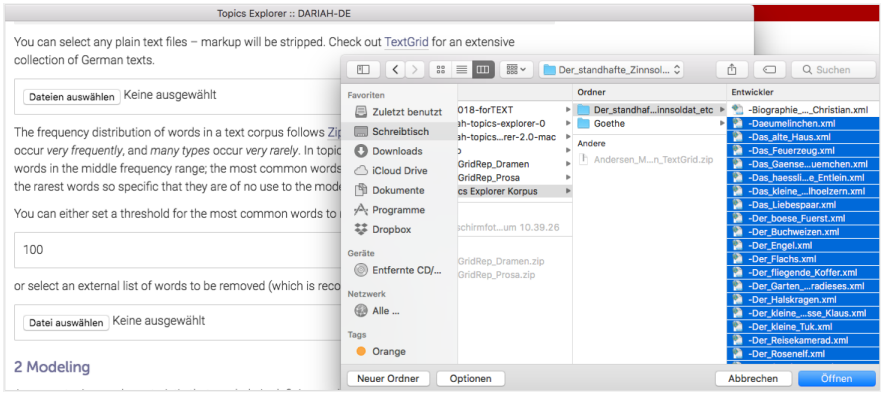

Abb. 6: TopicsExplorer: Dateien auswählen

Als nächstes legen wir fest, welche Wörter vom Topic Modeling ausgeschlossen werden sollen. Da in den meisten Fällen die häufigsten Wörter thematisch weniger relevant sind (diese sog. *most frequent words* (MFW) sind in der Regel hauptsächlich Funktionswörter wie *der*, *die*, *das*, *denn*, *da*, *weil*, *ob*, etc.), kann man entweder die Menge der häufigsten Wörter festlegen (voreingestellt sind die 100 MFW), oder - und das tun wir in dieser Lerneinheit - eine personalisierte Stoppwortliste erstellen, die für die jeweilige Textsammlung individuell angepasst wird, um möglichst konzise Topics zu erhalten. Dafür klicken Sie auf den zweiten „Datei auswählen“-Button und navigieren im Ordner des Topics Explorers über „sample-corpus“ zu „stopwords“. In diesem Ordner finden sich für verschiedene Sprachen Listen von typischen Wörtern, die beim Topic Modeling i.d.R. keine Aussagekraft haben. Da Sie diese Liste im weiteren Verlauf verändern werden, machen Sie sich am besten eine Kopie der Datei „de.txt“ und wählen diese Kopie aus. Die Datei können Sie z.B. auch in „stopwords Andersen“ o.ä. umbenennen, damit Sie später wissen, für welche Textsammlung diese Stoppwortliste angeglichen wurde.

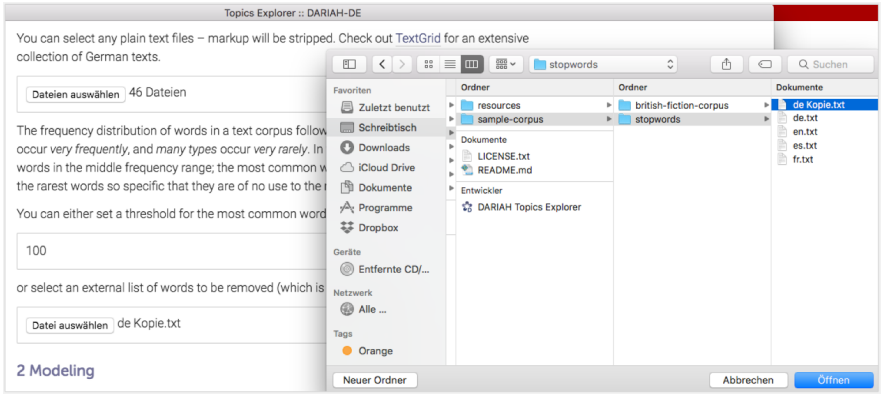

Abb. 7: TopicsExplorer – eine Kopie erstellen

Sie haben nun eine Sammlung von Texten im gleichen Format zusammengestellt und diese in den Topics Explorer geladen sowie eine Stoppwortliste hinzugefügt. Damit sind Sie startklar zum Topic Modeln.

3. Funktionen

Standardmäßig sind im Topics Explorer 10 Topics und 100 Iterationen eingestellt. Diese Werte können Sie im Folgenden jeweils verändern, um die Topics zu modellieren, bis sie Ihnen möglichst aussagekräftig erscheinen. Eine weitere Einflussmöglichkeit bietet die Stoppwortliste, die man nach jedem Durchgang durch diejenigen Wörter aus den Topics ergänzen sollte, die keine „thematische“ Aussagekraft haben. Dazu öffnen Sie einfach Ihre Andersen-Stoppwortliste und fügen am Ende der Liste die weiteren Wörter hinzu (pro Wort eine Zeile).

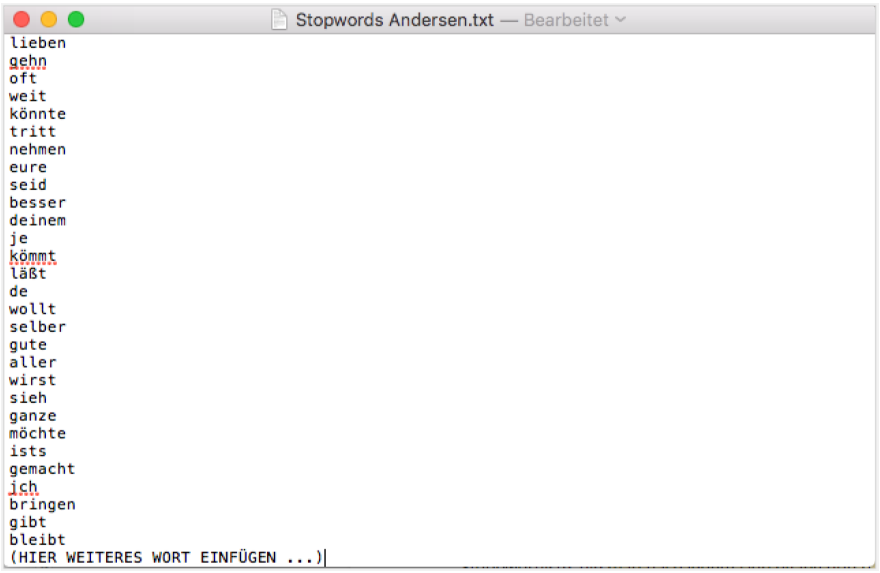

Abb. 8: Andersen-Stoppwortliste

Ein Blick in die zur Verfügung gestellte Stoppwortliste lohnt sich außerdem, wenn Sie eine Fragestellung haben, für deren Beantwortung bestimmte Wörter wichtig sein könnten (so stehen auf der Liste beispielsweise Wörter wie „müssen“ oder „sollen“ oder sogar „hören“, „wissen“ oder „sohn“).
Klicken Sie nun auf die Schaltfläche „**Train Topic Model**“ (siehe Abb. 9), beginnt das Programm mit der statistischen Auswertung. Mit nur 100 Iterationen geht das sehr schnell; in späteren Durchgängen mit sehr vielen Iterationen oder Topics werden Sie feststellen, dass die Berechnung teilweise ziemlich lange brauchen kann.

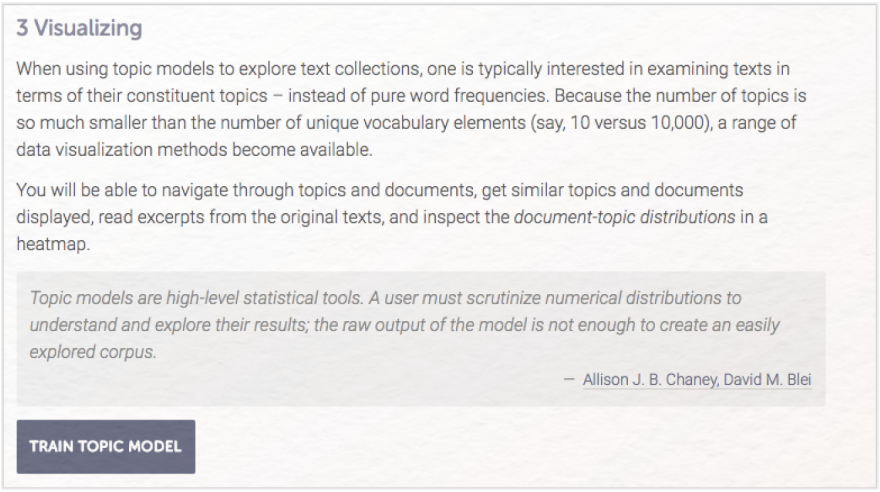

Abb. 9: Topic Model trainieren

Aufgabe 1: Führen Sie ein Topic Modeling mit den Standardeinstellungen durch und ergänzen im Anschluss Ihre Stoppwortliste. Welche Wörter schreiben Sie auf die Liste?

Nachdem das System das Topic Modeling durchgeführt hat (vgl. Text Mining), werden Ihnen die **Topics** in absteigender Häufigkeit als Balken angezeigt. Von jedem Topic sehen Sie in dieser Ansicht die ersten drei Worte. Im oberen blauen Balken finden Sie Module zur Navigation, Exportoptionen und unter „Reset“ können Sie einen neuen Topic-Modeling-Durchgang mit veränderten Einstellungen starten.

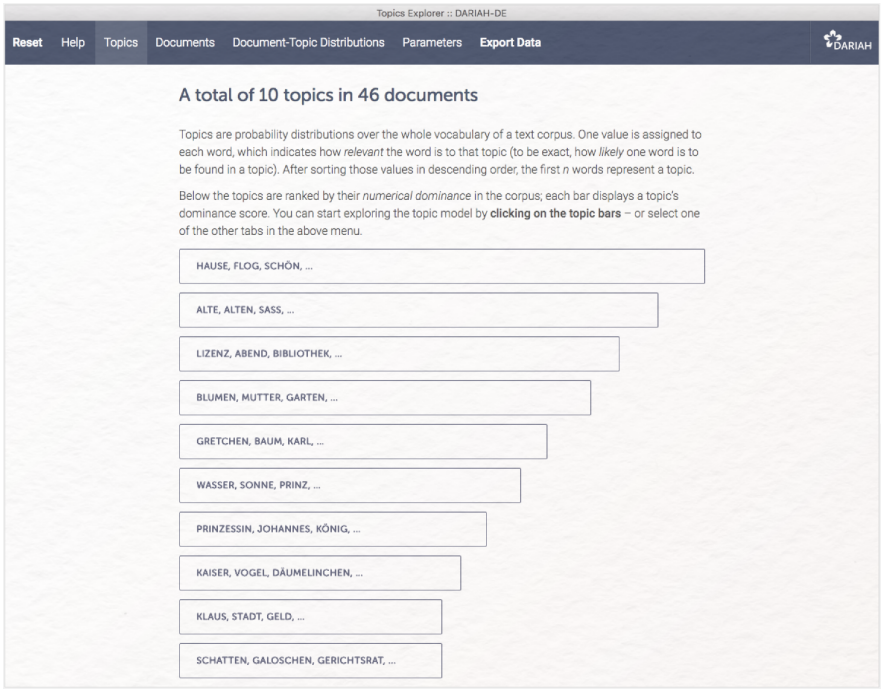

Abb. 10: Ergebnisse des Topic Modelings

Die Topic-Balken sind klickbar. Wählen Sie beispielsweise das Topic „BLUMEN, MUTTER, GARTEN, …“ aus, erhalten Sie eine sortierte Liste der 15 häufigsten Wörter in diesem Topic und wiederum horizontal angeordnete Balken daneben. In diesem Fall werden die 10 Dokumente aufgelistet, in denen das jeweilige Topic am ausgeprägtesten vorkommt. Unter dem angezeigten Topic werden außerdem die drei statistisch (nicht inhaltlich!) vergleichbarsten Topics gelistet.

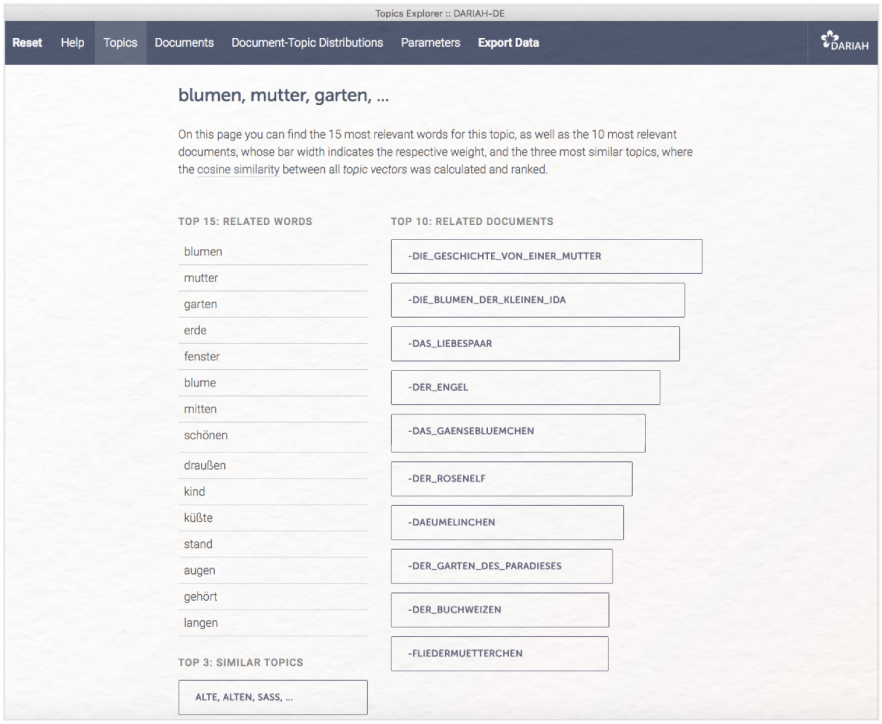

Abb. 11: Topics auswählen

Dass alle Wörter eines Topics klein (bzw. in der Gesamtübersicht in Versalien) geschrieben sind, zeigt, dass der Topic-Modeling-Algorithmus im Preprocessing sämtliche Buchstaben in Kleinbuchstaben verwandelt. Dieses Vorgehen ist Standard im Topic Modeling - einem Verfahren, dass an der (thematischen) Bedeutung von Wörtern

interessiert ist. Der große Vorteil ist, dass Wörter am Satzanfang und die gleichen Wörter in der Satzmitte auf diese Weise nicht als unterschiedliche Wörter gerechnet werden. Gerade im Deutschen geht damit jedoch auch eine Schwierigkeit einher, die man sich bei der Auswertung von Topics vergegenwärtigen sollte: Es gibt sehr viele Substantive, die klein geschrieben eine völlig andere Bedeutung erhalten, z.B. „spiel“ (als Imperativ), „spinnen“ (als Tätigkeit oder Charaktereigenschaft), oder auch das im hier gezeigten Topic auftauchende Wort „stand“, das gerade im thematischen Zusammenhang mit Blumen ebenso einen Stand bezeichnen als einen Hinweis dafür liefern könnte, dass die Erzählungen, in denen dieses Topic vorkommt, mit großer Wahrscheinlichkeit im Präteritum geschrieben sind.

Aufgabe 2: Schauen Sie sich das ausgewählte Topic und die dazugehörenden Dokumenttitel genau an. Anhand welcher Kriterien können Sie die „Korrektheit“ des Topics bewerten? Und warum sind im Topic alle Wörter klein geschrieben? Was kann das für Schwierigkeiten bei der Topic-Berechnung mit sich bringen? Und schließlich: Können Sie anhand dieses einen Topics bereits Interpretationshypothesen formulieren, die das Œuvre Andersens (oder einen Teil davon) charakterisieren?

Wenn Sie in der Navigationsleiste auf „**Documents**“ klicken, erhalten Sie eine Liste all der an dem jeweiligen Topic-Modeling-Durchgang beteiligten Texte. Auch hier sind die Balken interaktive Schaltflächen. Ein Klick auf ein Topic zeigt nicht nur den gesamten Primärtext an, sondern auch die Top 10 der vertretenen Topics sowie darauf aufbauend die drei statistisch am meisten vergleichbaren Texte der Sammlung.

Aufgabe 3: Welches Märchen Andersens ist das zweitlängste, welches das zweitkürzeste? Welche Topics sind in diesen beiden Texten besonders prominent?

Das Modul „**Document-Topic Distributions**“ zeigt eine sog. Heatmap aller Topics (auf der x-Achse) und aller Dokumente (auf der y-Achse). Eine grundlegende Annahme im Topic Modeling ist, dass jedes Topic mit einer bestimmten Wahrscheinlichkeit in jedem Dokument vorkommt. Die besonders geringen Wahrscheinlichkeiten werden in diesem Fall vom Topics Explorer auf Null abgerundet. Je stärker ein Topic in einem Dokument vertreten ist, desto dunkler ist das jeweilige Feld in der Heatmap visualisiert. Rechts oben gibt es in dieser Visualisierung noch eine gesonderte Möglichkeit des Downloads: Als Bilddatei (PNG) oder als Vektorgrafik (SVG).

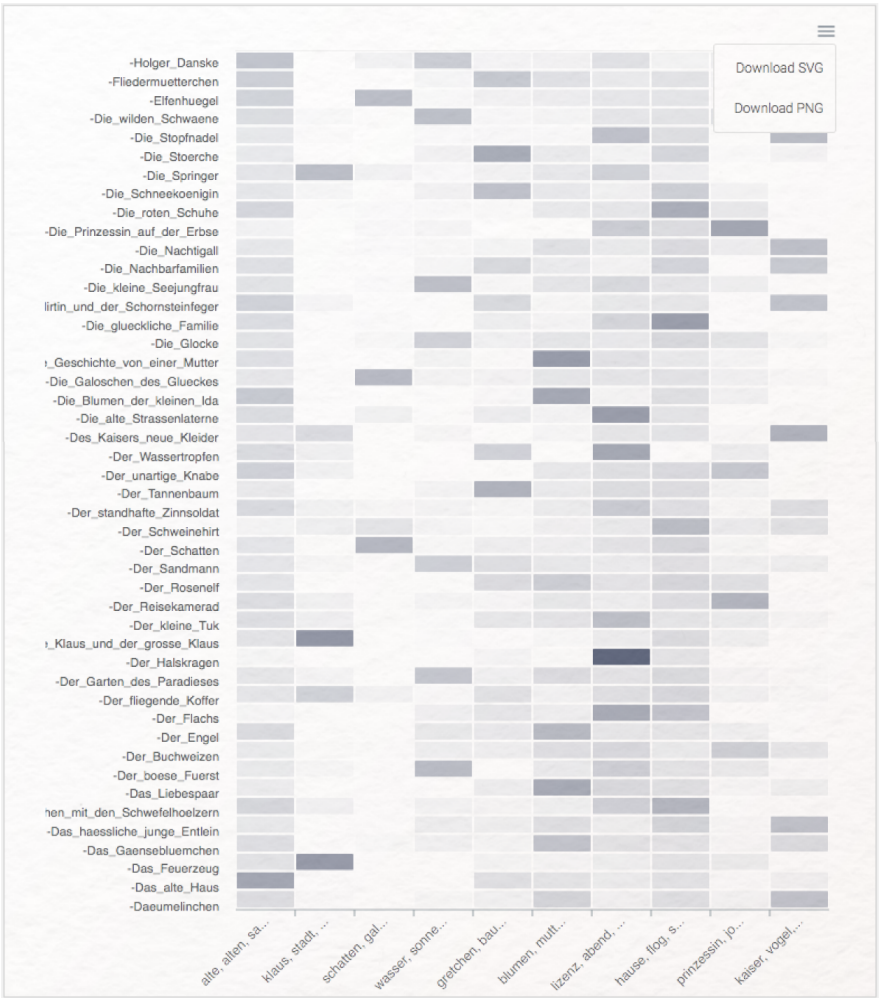

Abb. 12: TopicsExplorer Heatmap

Aufgabe 4: Über welche Texte lassen sich bereits nach diesem ersten Topic-Modeling-Durchgang relativ klare „thematische" Aussagen treffen? Bei welchen Texten haben wir bislang kaum einen Einblick in virulente Themen?

Unter „**Parameters**" können Sie jederzeit die Einstellungen Ihres jeweiligen Topic-Modeling-Durchgangs nachschauen. Hier erfahren Sie außerdem, wie ein Topic Modeling evaluiert und verbessert werden kann: Durch die Erhöhung der Iterationen werden sich die Qualität und das sog. „log-likelihood" (eine logarithmierte Angabe zur statistischen Wahrscheinlichkeit) bis zu einem gewissen Punkt erhöhen. Diesen Punkt gilt es durch Erhöhung der Iterationen und anschließenden Vergleich der jeweiligen log-likelihood-Zahlen herauszufinden.
Das Modul „**Export Data**" bietet die Möglichkeit, diverse Listenansichten Ihrer Topic-Modeling-Daten gebündelt als ZIP-Datei herunterzuladen. Wenn bei Ihnen das Kürzel „.zip" noch nicht hinter der Downloaddatei steht, schreiben Sie es einfach selbst dazu, damit Ihr Computer die Exportdateien entpackt (siehe Abb. 13).

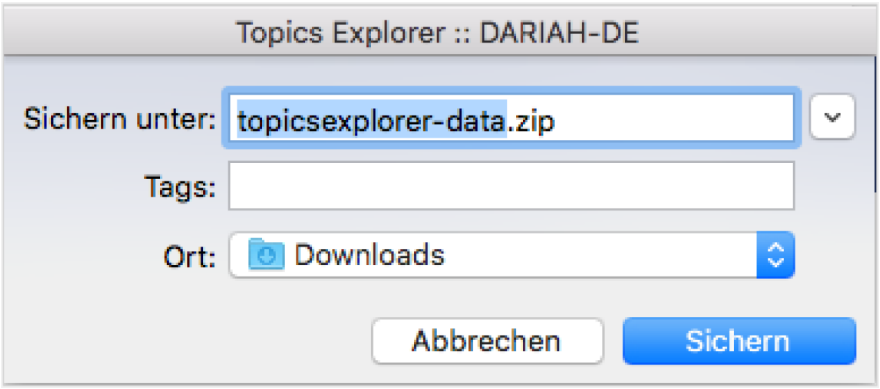

Abb. 13: TopicsExplorer Daten exportieren

In den Exportdateien werden Ihnen neben der aktuellen Stoppwortliste mehrere CSV-Dateien (vgl. CSV) (öffnen mit *Numbers* oder importieren in *Excel*) geboten: eine Liste der Dokumentähnlichkeiten, der Topicähnlichkeiten, eine zahlenbasierte Document-Topic-Verteilung und eine Liste der erstellten Topics mit den 99 statt nur 15 häufigsten Wörtern jedes Topics.
Ein Klick auf das Modul „**Reset**" beendet Ihren derzeitigen Topic-Modeling-Durchgang und führt Sie auf die Startseite zurück. Hier können Sie nun eine erweiterte Stoppwortliste verwenden und die Einstellungen (Menge der Topics und Iterationen) verändern. In der derzeitigen Version lässt es sich leider nicht vermeiden, dass auch die Textsammlung bei jedem Durchgang neu in den Topics Explorer hochgeladen werden muss.

Aufgabe 5: Erweitern Sie Ihre Stoppwortliste und erhöhen Sie die Zahl der Iterationen. Mit wie vielen Iterationen werden die besten Ergebnisse erzielt? Welche Märchen sollte man sich vermutlich näher anschauen, interessiert man sich für Metanarration?

4. Lösungen zu den Beispielaufgaben

Beachten Sie bei den hier angegebenen Lösungen, dass der Algorithmus zum Topic Modeling auf einer wiederholt zufälligen Auswahl der Textsegmente basiert. Die Ergebnisse sind bei gleichen bzw. guten Einstellungen daher zwar meistens sehr ähnlich, eine hundertprozentige Reproduktion von Ergebnissen ist mit diesem Verfahren jedoch nicht möglich. Die hier beschriebenen Lösungen können daher leicht von Ihren persönlichen Ergebnissen abweichen.

Aufgabe 1: Führen Sie ein Topic Modeling mit den Standardeinstellungen durch und ergänzen im Anschluss Ihre Stoppwortliste. Welche Wörter schreiben Sie auf die Liste?

„Thematisch" wenig aussagekräftig sind Wörter wie „sodaß", „elisa", „klaus", „johannes", vor allem aber Wörter wie „digitalen" „lizenz", „bibliothek", „textgrid", „andersen".

Aufgabe 2: Schauen Sie sich das ausgewählte Topic und die dazugehörenden Dokumenttitel genau an. Anhand welcher Kriterien können Sie die „Korrektheit" des Topics bewerten? Und warum sind im Topic alle Wörter klein geschrieben? Was kann das für Schwierigkeiten bei der Topic-Berechnung mit sich bringen? Und schließlich: Können Sie anhand dieses einen Topics bereits Interpretationshypothesen formulieren, die das Œuvre Andersens (oder einen Teil davon) charakterisieren?

Auch wenn man die Texte Andersens nicht kennt, kann man anhand der Dokumenttitel bereits darauf schließen, dass das Topic korrekt zusammengestellt und mit den jeweiligen Dokumenten identifiziert wurde: Blumen, Mütter, Gärten, Erde, Kind, Fenster zeigen nicht nur selbst eine Zusammengehörigkeit (man könnte hier das Thema „Familie und Garten" konstruieren), sondern auch die Zuordnung zu Dokumenten wie *Die Geschichte von*

einer Mutter, *Die Blumen der kleinen Ida* oder *Das Gänseblümchen* erscheinen auch ohne Textkenntnis sehr passend. Eine Beispielhypothese auf Grundlage dieses einen Topics der Texte Andersens wäre, dass Weiblichkeit in den Märchen dieses Autors mit Mutterschaft, Blumen, Gärten, Schönheit und Kindern verknüpft wird, womit die vorliegenden Texte sowohl für die Literatur des 19. Jahrhunderts als auch für das Genre Märchen als paradigmatisch eingestuft werden könnten.

Aufgabe 3: Welches Märchen Andersens ist das zweitlängste, welches das zweitkürzeste? Welche Topics sind in diesen beiden Texten besonders prominent?

Die Dokumente sind unter „Documents" der Textlänge nach angeordnet. Das zweitlängste Märchen ist *Die Galoschen des Glückes* mit dem virulenten Thema „schatten, galoschen, gerichtsrat, …" und das zweitkürzeste *Der Wassertropfen* mit dem wenig aussagekräftigen Topic „lizenz, abend, bibliothek, …".

Aufgabe 4: Über welche Texte lassen sich bereits nach diesem ersten Topic-Modeling-Durchgang relativ klare „thematische" Aussagen treffen? Bei welchen Texten haben wir bislang kaum einen Einblick in virulente Themen?

Die Texte *Die Prinzessin auf der Erbse*, *Die glückliche Familie*, *Die Geschichte von einer Mutter*, *Der kleine Klaus und der große Klaus* und *Das Feuerzeug* haben stärker vertretene Topics vorzuweisen. Vorsicht ist geboten bei der vermeintlich stärksten ausgeprägten Themenzuordnung zum Text *Der Halskragen*. Der Text selbst ist äußerst kurz und das hier als stark vertreten gekennzeichnete Thema beinhaltet sehr viele Wörter, die sich auf die Metadaten des Dokumentes beziehen (aus dem sog. TEI-Header (vgl. TEI)).
Wenige Aussagen kann man über diejenigen Texte treffen, in denen die einzelnen Topics entweder alle sehr ähnlich oder kaum vertreten sind, wie z.B. *Die Glocke*, *Der Buchweizen* oder auch *Die Hirtin und der Schornsteinfeger*.

Aufgabe 5: Erweitern Sie Ihre Stoppwortliste und erhöhen Sie die Zahl der Iterationen. Mit wie vielen Iterationen werden die besten Ergebnisse erzielt? Welche Märchen sollte man sich vermutlich näher anschauen, interessiert man sich für Metanarration?

Kontrolliert und vergleicht man die Werte des log-likelihoods nach jedem Durchgang, zeigt sich, dass bei mehr als 2000 Iterationen die Zahl nur noch sehr wenig zunimmt. Der höhere Zeitaufwand wird dadurch ab einem gewissen Punkt nicht mehr gerechtfertigt und auch aus der Nutzer*innenperspektive haben die Topics bei über 2000-5000 Iterationen eher wieder eine geringere Aussagekraft. Das folgende Diagramm visualisiert diese Entwicklung beispielhaft (siehe Abb. 14).

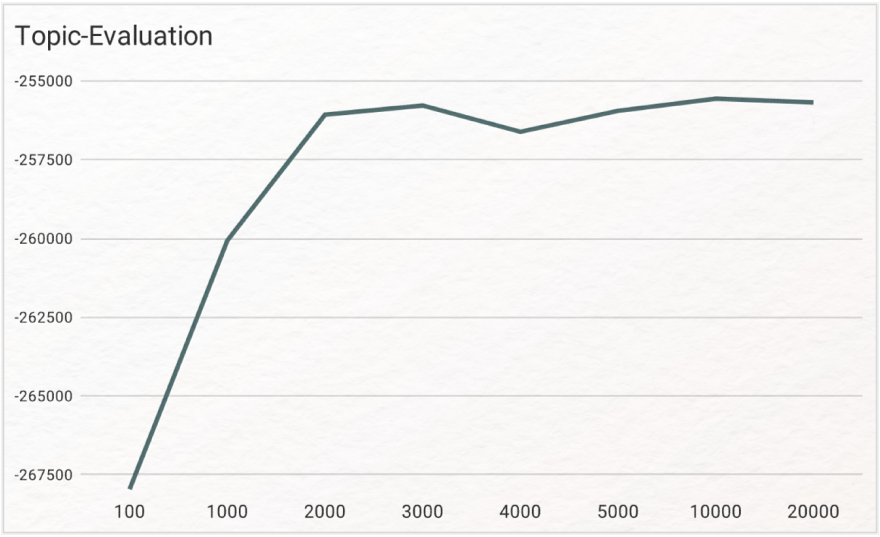

Abb. 14: TopicExplorer Evaluation

Die Stoppwortliste wurde für diese Tests mit den folgenden Wörtern erweitert: „andersen", „bibliothek", „danske", „digitalen", „elisa", „friedrich", „gretchen", „holger", „ida", „johannes", „karl", „klaus", „leipzig", „lizenz", „marie", „sodaß", „textgrid" und „tuk".
Besonders metanarrativ scheinen die Texte *Der Tannenbaum*, *Der Flachs*, *Der Buchweizen* und *Der Wassertropfen* zu sein, denn beispielsweise in dem hier virulenten Topic „standen, Baum, kamen, …" (bei 2000 Iterationen) kommen sehr viele Worte vor, die das Erzählen und die Sprache selbst betreffen.

Externe und weiterführende Links

- DARIAH TopicsExplorer: https://web.archive.org/save/https://github.com/DARIAH-DE/TopicsExplorer/releases/tag/v2.0 (Letzter Zugriff: 19.09.2024)
- DARIAH TopicsExplorer installieren: https://web.archive.org/save/https://dariah-de.github.io/TopicsExplorer/#getting%20started (Letzter Zugriff: 19.09.2024)
- TextGrid Repository: https://web.archive.org/save/https://textgridrep.org/ (Letzter Zugriff: 19.09.2024)
- Tutorialvideo Sicherheitseinstellungen (MacOS): https://doi.org/10.5281/zenodo.11074222 (Letzter Zugriff: 19.09.2024)
- Tutorialvideo Sicherheitseinstellungen (Windows): https://doi.org/10.5281/zenodo.11074232 (Letzter Zugriff: 19.09.2024)

Bibliographie

forTEXT. 2019a. Tutorial: Sicherheitsaufnahme für Internetprogramme Hinzufügen (Mac). 19. Januar. https://doi.org/10.5281/zenodo.11074232.

———. 2019b. Tutorial: Sicherheitsausnahme für Internetprogramme Hinzufügen (Windows). 25. Januar. https://doi.org/10.5281/zenodo.11074222.

Horstmann, Jan. 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.

———. 2024b. 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.

Schumacher, Mareike. 2024a. 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.

———. 2024b. Methodenbeitrag: Named Entity Recognition (NER). Hg. von Evelyn Gius. *forTEXT* 1, Nr. 9. Named Entity Recognition (30. Oktober). doi: 10.48694/fortext.3765, https://fortext.net/routinen/methoden/named-entity-recognition-ner.

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).

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.

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.

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.

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.

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.

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.

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.

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.