Seminar FS 23
“Grand Tour gone digital”
Dozentin: Ina Serif

Stefan Heinen
Nora Holländer
Anna Scognamiglio

Arbeitsbericht zur Tf-idf-Analyse

Ausgangslage

Im Rahmen dieses Projekts haben wir uns mit Tf-idf auseinandergesetzt. Tf-idf steht für «term frequency–inverse document frequency» und erlaubt es, die relative Häufigkeit von Wörtern in Dokumenten ausfindig zu machen. Als Ergebnis für eine Tf-idf-Analyse erhält man einen Wert pro Wort, der angibt, wie spezifisch dieses Wort für ein Dokument ist, verglichen mit anderen Dokumenten im Korpus. Kommt also ein Wort in einem Dokument sehr häufig vor, im ganzen Korpus jedoch nur sehr selten, hat es einen hohen Tf-idf-Wert im jeweiligen Dokument. Kommt ein Wort in einem Dokument zwar sehr häufig vor, aber auch in allen anderen Dokumenten des Korpus, dann hat es einen tieferen Tf-idf-Wert. Somit kann herausgefunden werden, welche Wörter charakteristisch für ein Dokument sind, während häufig vorkommende, unspezifische Wörter, wie Funktionswörter oder Partikel, nicht ins Gewicht fallen.

Mit dieser Technologie können umfangreiche Textbestände durch Distant Reading verglichen werden, ohne den ganzen Korpus vollständig lesen zu müssen (Close Reading). Als Ausgangskorpus wählten wir die Reiseberichte von Philipp Hainhofer. Dieser besteht aus 11 Einzeltexten, die jeweils einer Station seiner Reisen entsprechen. Da es sich um Berichtsammlungen verschiedener Reisen handelt, kommen manche Orte mehrfach vor.

Im weiteren Verlauf wurden weitere Texte zu Vergleichszwecken herangezogen. Im Sinne der Tf-idf-Analyse wären Berichte über dieselben Orte besonders interessant gewesen. Da solche Vergleichstexte zur Zeit jedoch leider nicht digitalisiert zur Verfügung stehen, wurden Reiseberichte gewählt, die sich in ihrem Umfang weitestgehend gleichen. Schliesslich bestand das Korpus aus vier Quellen, von denen jeweils zwei bezüglich ihrer Entstehungszeit relativ nahe beieinander liegen:

1. Hainhofer, Philipp: Reiseberichte und Sammlungsbeschreibungen 1594–1636. Edition und Datensammlung zur Kunst- und Kulturgeschichte der ersten Hälfte des 17. Jahrhunderts [Wolfenbütteler Digitale Editionen, Nr. 4], hrsg. und

eingeleitet von Michael Wenzel, Transkription und Kommentar von Ursula Timann und Michael Wenzel, Wolfenbüttel: Herzog August Bibliothek 2020ff.

2. Neitzschitz, Georg Christoph von: Sieben-Jährige und gefährliche WeltBeschauung Durch die vornehmsten Drey Theil der Welt Europa/ Asia und Africa. Bautzen, 1666. In: Deutsches Textarchiv https://www.deutschestextarchiv.de/neitschitz_reise_1666.
3. Sulzer, Johann Georg: Tagebuch einer von Berlin nach den mittäglichen Ländern von Europa in den Jahren 1775 und 1776 gethanen Reise und Rückreise. Leipzig, 1780. In: Deutsches Textarchiv https://www.deutschestextarchiv.de/sulzer_reise_1780.
4. Schulz, Friedrich : Reise eines Liefländers, Bd. 1-3, 1795.

Alle Texte stehen als TXT-Dateien online zur freien Verfügung. Während die letzten drei bereits in normalisierter und lemmatisierter Form vorliegen, mussten diese Bearbeitungsschritte bei den Hainhofer-Texten noch durchgeführt werden.

Vorgehen

Zur Vorbereitung haben wir als Erstes den Blog zu Tf-idf von Programming Historian[1] gelesen und uns mit dem dort zur Verfügung gestellten Code vertraut gemacht. Nachdem wir das Jupyter Notebook und Python installiert hatten, haben wir die nötigen Funktionen programmiert, um den Korpus zusammenzustellen. Nun haben wir ausgehend vom Code von Programming Historian einen auf unsere Quellen angepassten Code geschrieben und damit Tf-idf Werte für jedes Wort von allen elf Texten des Korpus errechnet. Als Ergebnis erhielten wir eine Liste/CSV-Tabelle je Dokument, welche die Wörter nach Tf-idf Wert geordnet aufgelistet hat. Diese Listen waren, wie erhofft, für alle Reiseberichte unterschiedlich und es schien sich ein Profil pro Dokument abzubilden. Es gab jedoch auch einzelne Wörter, welche nicht aussagekräftig waren wie «euer» oder «bei». Diese Wörter lassen sich jedoch relativ einfach mit Stoppwortlisten herausfiltern.

Wir haben uns nun etwas genauer mit den Quellen selbst und deren Inhalt auseinandergesetzt, also das Distant Reading durch ein Close Reading ergänzt. Hierfür wurden für jedes Dokument die 20 Begriffe mit den höchsten Tf-idf Werten extrahiert und mit der online zur Verfügung stehenden Kurzzusammenfassung des

jeweiligen Reiseberichtes von Hainhofer verglichen. Wir wollten herausfinden, wie repräsentativ die errechneten Tf-idf-Werte und die ausgegebene Spezifität der Begriffe für jedes Dokument sind. Dabei konnte festgestellt werden, dass die ermittelten Begriffe durchaus als semantisch richtungsweisend anzusehen sind. Im Reisebericht von München 1613 beispielsweise geht Hainhofer nach München, um eine Hochzeit zu besuchen. Unter den ersten sieben Wörtern befinden sich «ehestand, braut, bräutigam, breutigam, hochzeiterin» Auch wenn die Liste die Zusammenfassung nicht ersetzt, wird das Thema also in der Tf-idf-Liste semantisch gut fassbar. Ein solches Profil ist bei allen Wortlisten erkennbar, auch wenn die Qualität unterschiedlich zu sein scheint. Zum Teil tauchen Wörter auf, die wenig bedeutungstragend sind.

Schliesslich haben wir die drei weiteren Reiseberichte hinzugefügt, um Vergleichswerte ermitteln zu können. Dies sind die Reiseberichte Nietzschitz (1666), Schulz (1795) und Sulzer (1780). Sie sind alle im deutschen Textarchiv online zugänglich.[2]

Zunächst mussten die neuen Texte bereinigt werden, um eine sinnvolle Vergleichbarkeit zu ermöglichen. Ausserdem wurde für den Hainhofer Korpus mithilfe des Modells «de_core_news_sm» von SpaCy eine Lemmatisierung erstellt. Wir haben Stoppwortlisten errechnet und für jedes Wort (Lemma) die Wortgruppe bestimmen lassen. Auf diese Weise konnten wir anschliessend alle Adjektive sowie Nomen herausfiltern und einer Ähnlichkeitsanalyse unterziehen. Hierfür haben wir Tf-idf Listen nur mit den Adjektiven bzw. Nomen für jedes der vier Dokumente erstellt und diese mit Balkendiagrammen visualisiert.

Parallel dazu haben wir den gesamten Korpus mithilfe von Voyant Tools etwas genauer betrachtet. Um die Bearbeitung in Voyant zu erleichtern, haben wir dafür auch manuell eine Stoppwortliste angelegt. Auch in Voyant haben wir eine Tf-idf Liste und andere interessante Werkzeuge entdeckt, die zur Verfügung gestellt werden. Bei den Stoppwortlisten ist jedoch nicht transparent, wie genau die Werte zustandekommen und die Liste scheint ungenauer zu sein. Auch die weiteren Tools von Voyant halfen dabei, einen groben Überblick zu erhalten, aber auch sie verhalfen uns nicht zu einer konkreten Fragestellung.

Stefan Heinen
Nora Holländer
Anna Scognamiglio

Probleme

Während der Projektarbeit haben sich ~~uns~~ verschiedene Probleme/ Herausforderungen ergeben, welche aber größtenteils lösbar waren. Der erste Themenkomplex war die Aufarbeitung der Texte. Beispielsweise kam manchmal dasselbe Wort mit unterschiedlicher Schreibweise in den Listen mehrmals vor. Als Lösung für dieses Problem haben wir eine Lemmatisierung durchgeführt.

Ein anderes Problem dieser Kategorie war, dass auch viele irrelevante Wörter wie «einem» oder «bej» im ersten Teil der Tf-idf-Listen vorkamen. Um diesem Problem entgegenzuwirken, haben wir Stoppwortlisten erstellt.

Die Qualität der einzelnen Listen hängt stark vom Gesamtkorpus ab. Eine der Listen war unbrauchbar, da das Dokument auch einen italienischen Brief enthielt und Italienisch in den anderen Berichten gar nicht vorkam. Somit wurden diese italienischen Begriffe nicht als «irrelevant» identifiziert und daher nicht herausgefiltert. Die ermittelten Werte sind in diesem Fall unbrauchbar, da hier die relative Häufigkeit äquivalent ist mit der absoluten Häufigkeit der Wörter. Dies kann damit gelöst werden, dass das Dokument übersetzt oder ausgelassen wird.

Fragestellung

Wir gingen von der folgenden Fragestellung aus: *Wie werden die verschiedenen Städte in den verschiedenen Reiseberichten dargestellt?* Bzw. *Wird die gleiche Stadt zu unterschiedlichen Zeiten anders dargestellt?* Und: *Werden die Städte unterschiedlich dargestellt?* Wir wollten herausfinden, ob für die verschiedenen Städte andere Gruppen von Vokabularien verwendet werden. Beispielsweise für eine Stadt mehr politische Begriffe und für eine andere Stadt eher religiöse etc. Auch wollten wir herausfinden, ob manche Städte positiver oder negativer bewertet werden. Wir konnten diese Fragestellung teilweise beantworten. Einerseits haben sich auch in den Stoppwortlisten gewisse Profile für die verschiedenen Berichte abgezeichnet. Im Bericht «München 1603» scheint es beispielsweise eher um den kirchlichen Aspekt der Stadt zu gehen und in «München 1613» mehr um den gesellschaftlichen Aspekt der Hochzeit.

Wir haben uns erhofft, im Verlauf der Arbeit eine noch konkrete Forschungsfrage zu finden, welche anhand der Listen beantwortet werden könnte. Dies war uns jedoch nicht möglich.

~~Reflektion~~

Es stellte sich als sehr schwierig heraus, von einer Methode zu einer Fragestellung zu gelangen. In einem nächsten Projekt würden wir von einer Fragestellung ausgehen und uns dann in einem zweiten Schritt überlegen, mithilfe von welchen Techniken und Methoden diese Fragestellung beantwortet werden kann.

Trotzdem haben wir im Verlauf dieses Projekts sehr viel gelernt und sind nun viel vertrauter mit Tf-idf und den notwendigen Parametern und Aufbereitungsschritten. Wir können uns nun besser vorstellen, für welche Projekte Tf-idf sinnvoll sein kann. Es ist sicherlich ein gutes Tool, um einen ersten Eindruck eines Textkorpus zu erhalten. Allenfalls würde dafür auch die Tf-idf-Funktion von Voyant Tools ausreichen. Für eine weitergehende Arbeit mit Tf-idf ist es sicherlich sinnvoll, selbst Code zu schreiben, um Parameter individuell und gezielt anpassen zu können.

Ausblick

Im weiteren Vorgehen könnten die Texte noch weiter aufbereitet und bereinigt werden. Die errechneten Stoppwortlisten und die Erkennung der Adjektive sind beide bereits recht gut, könnten für eine grössere Arbeit jedoch wohl noch verfeinert werden. Das Korpus ist nun so weit aufgearbeitet, dass mithilfe von SpaCy verschiedene weitere Analysemethoden einfach möglich sind. In einem weiteren Schritt könnte auch ein eigenes Modell trainiert werden.

---

[1] Unter folgendem Link findet sich eine gute Einführung in Tf-idf von Programming Historian: https://programminghistorian.org/en/lessons/analyzing-documents-with-tfidf (Zuletzt abgerufen am 16.05.2024)

[2] Nietzschitz: https://www.deutschestextarchiv.de/book/show/neitschitz_reise_1666 Schulz: https://www.deutschestextarchiv.de/book/show/schulz_reise0101_1795 Sulzer: htt333ps://www.deutschestextarchiv.de/book/show/sulzer_reise_1780