Arnold Schönberg - Korrespondenz
Topics

Katharina Bleier

Machine Learning für digitale Editionen WS 2024/25

Universität Graz

Institut für Digitale Geisteswissenschaften

1. Projektquellen und Fragestellungen

Arnold Schönbergs Korrespondenz mit seinen Verlagen Universal-Edition (UE) und Dreililien umfasst den Zeitraum von 1902 bis zum Tod des Komponisten 1951. Der Verlag Dreililien in Berlin war der erste Verlag, der Schönbergs Werke publizierte, 1912 gingen diese ins Verlagsprogramm der UE über, weshalb diese Korrespondenz in das Korpus integriert wurde.[1] Die UE war Schönbergs Hauptverlag bis zu seiner Emigration in die USA 1933, der Briefwechsel ist das größte Konvolut der innerhalb der Schönberg-Korrespondenz und umfasst rund 1400 Dokumente. Sowohl innerhalb verlagstechnischer Publikations- und Vertriebsprozesse als auch darüberhinausgehend stellt die UE einen wichtigen Knoten im Netzwerk des Komponisten und der Wiener Schule dar. Dies spiegelt sich in einer großen Bandbreite von behandelten Themen in der Korrespondenz wider.

Das vorliegende Projekt soll dem Leseeindruck von Vielfalt, Wiederkehr und Relevanz von Themen den maschinellen Ansatz des Topic Modeling gegenüberstellen. Von besonderem Interesse ist hier zunächst, wieviele Topics eine Analyse bei "out-of-the-box"-Anwendung eines Modells definiert, welche Themen sich als besonders gewichtig erweisen und wie sinnvoll musikalische Sachverhalte in Topics abgebildet werden können. Das Projekt umfasst 3 Schritte:

- Präparation der Daten
- Topic Modeling
- Visualisierung und Interpretation

Die Projektdaten umfassen vorliegende Dokumentation und Interpretation, Code mit Kurzdokumentation (Machine_Bleier.ipynb), 10 xml-Dokumente zur Überprüfung des Präparations-Workflows[2], die Datei „sentences.csv“, welche die Grundlage für das Topic Modeling enthält sowie „sentences.mm“, das Modell, anhand dessen untenstehende Interpretation vorgenommen wurde.

2. Präparation der Daten

Die Datengrundlage bilden rund 1400 xml-Dateien, welche TEI codierte Briefe der beschriebenen Korrespondenz enthalten. Für die Topic-Analyse werden aus den TEI-Dateien lediglich jene Textteile benötigt, welche in Paragraphen innerhalb des <body> stehen sowie die ID der verwendeten Quellen um Zuordenbarkeit zu gewährleisten. Diese Informationen werden zur

[1] *Arnold Schönberg: Briefwechsel mit den Verlagen Universal-Edition und Dreililien.* Hrsg. von Katharina Bleier und Therese Muxeneder unter Mitarbeit von Jannik Franz und Philipp Kehrer, Universität für Musik und darstellende Kunst Wien und Arnold Schönberg Center, Wien. Version 1.1 vom 29.11.2024. URL: https://www.schoenberg-ue.at.
[2] Die XML-Dateien bedürfen noch einer Überarbeitung und werden daher erst zu einem späteren Zeitpunkt m obengenannten Projekt veröffentlicht.

Weiterverarbeitung zunächst briefweise in eine csv-Datei gespeichert. Für die Extraktion der Texte werden Module aus der Python-Standard-Library[3] verwendet:

- os — Miscellaneous operating system interfaces: handling file systems, creating output folder, listing files, constructing file paths
- csv – CSV File Reading and Writing: creating and opening and writing into csv file
- xml.etree.ElementTree — The ElementTree XML API: parsing XML, locating elements, extracting text and cleaning

In diesem Schritt erwies sich die Extraktion des Brieftextes unter Auslassung der Herausgeber:innen-Stellenkommentare als schwierig: Ein iterativer Ansatz zur Überprüfung aller <note> Elemente führte nicht zum gewünschten Ergebnis und wurde durch einen rekursiven Prozess durch den XML-Baum und Abbruch bei <note> Elementen ersetzt.[4]

Der briefweise abgelegte Text wird nun in kleinere Einheiten geteilt, in Kleinschreibung umgewandelt sowie Stopwörter entfernt. Von der nltk-Bibliothek wird das tokenizer-Modul[5] aus dem „Punkt"-Paket verwendet um den Text in Sätze zu splitten. Ein alternativ versuchter Ansatz, die Brieftexte mit dem BertTokenizer (BertTokenizer.from_pretrained("bert-base-german-cased")) aus der Hugging Face Transformers-Bibliothek in Chunks von max 300 Tokens zu splitten, sollte aufgrund des deutschsprachigen Trainings und der etwas größeren Textabschnitte besser geeignet sein. Es stellt sich aber heraus, dass aus bisher nicht klaren Gründen in der weiteren Verarbeitung lediglich vier, nicht aussagekräftige Topics identifiziert werden. Daher wird der satzbasierte Ansatz weiterverfolgt und eine weitere Bearbeitung vorgenommen.

Die nunmehr vorbereiteten Textteile verursachen eine Fehlermeldung im Ablauf des Topic Modeling, da vorkommende Aufzählungsnummerierungen (1.) und Daten (10.10.1917) als Datentyp „float" interpretiert werden, jedoch nur strings erlaubt sind. Die Untersuchung der entsprechenden Tabellenzeilen zeigt, dass diese lediglich Zahlen und Punkte enthalten und werden daher entfernt.

3. Topic Modeling

Das Topic Modelling erfolgt mit BERTopic[6], einem bidiretionalen Modell, welches nicht auf vordefinierte Wortlisten zurückgreift, sondern den vorhergehenden und nachfolgenden

[3] https://docs.python.org/3/library/index.html (Zugriff 10.02.2025)
[4] Code und ausführliche Erklärung vgl. e-Mail Martina Scholger an Katharina Bleier, 25.02.2025; herzlichen Dank für die Lösung dieses Problems.
[5] https://www.nltk.org/api/nltk.tokenize.html#nltk.tokenize.sent_tokenize (Zugriff 20.02.2025)
[6] https://maartengr.github.io/BERTopic/index.html (Zugriff 25.02.2025)

Wortkontext berücksichtigt um thematische Cluster zu bilden. Das Modell wird in der mehrsprachigen Variante verwendet, die Deutsch als Hauptsprache der Briefe inkludiert.

4. Visualisierung und Interpretation

Die Analyse ergibt 231 Topics. Übersicht über die Themen visualisiert eine Intertopic Distance Map, in welcher die relative Häufigkeit sowie Verwandtschaft von Themen veranschaulicht werden.

Exemplarische Detailbetrachtung ausgewählter Topics:

Topic -1 bezeichnet sog. „outliers" die üblicherweise ignoriert werden. Die genauere Betrachtung dieses Topics zeigt, dass das Wort „werk" enthalten ist, welches im Zusammenhang mit musikalischen Werken allerdings durchaus relevant sein könnte.

Topic 0: Typisch für diese Korrespondenz, die häufig Vertragsverhandlungen thematisiert, ist das hohe Ranking des entsprechenden Topics.

Topic 2 enthält spezifischere und damit aussagekräftigere musikalische Besetzungsbegriffe.

Topics 10 und 11 zeigen exemplarisch eine sinnvolle und für die vorliegende Korrespondenz relevante Differenzierung zwischen Korrektur von Aufführungsmaterial und Druckprozess.

Die Reduktion der Topics auf eine Anzahl von 30 bringt eine sinnvolle Verdichtung und bewirkt Übersichtlichkeit. Je nach Fragestellung kann eine weitere Verschmelzung von Topics sinnvoll sein, z. B. von Topic 12 und Topic 17, welche beide Aspekte musikalischer Aufführungen thematisieren.

Die abschließende Visualisierung in einer Heatmap ermöglicht den paarweisen Vergleich von Topics unter Angabe eines Ähnlichkeitswerts.

5. Fazit

Die „out of the box"-Anwendung bestätigt eine – gemessen an Anzahl und Umfang der untersuchten Quellen – große Themenanzahl. Die Treffsicherheit der Themenidentifizierung variiert und müsste ggf. einer möglichen Forschungsfrage entsprechend verfeinert werden, etwa durch Training anhand von annotierten Daten oder durch Mergen von Topics, die zusammengefasst werden können. Eine Verfeinerung wäre auch in der Präparation der Texte noch vorzunehmen um Rauschen und fehlerhafte Daten zu reduzieren.