Topic Modeling

Sarić, Sanja; sanja.saric@uni-graz.at / Krušić, Lucija; lucija.krusic@uni-graz.at

Topic Modeling ist eine Methode des unüberwachten maschinellen Lernens und dient zur automatischen Erkennung semantischer Strukturen in großen Textmengen. Dabei wird eine aus der Statistik stammende Hypothese befolgt, dass häufig zusammen vorkommende Wörter thematisch verwandt sind. (Blei 2012)

Als Ergebnis entstehen dabei Topics – geordnete Listen von zusammenhängenden Wörtern, die mit Wahrscheinlichkeitswerten versehen sind, welche auf die Häufigkeit des Vorkommens hinweisen. Ein Topic wird meistens als ein Thema interpretiert, kann aber statt aus dem thematischen auch aus einem strukturellen Zusammenhang zwischen Wörtern resultieren. So kann einerseits ein Topic zum Beispiel genre- oder autorenspezifische Wörter repräsentieren, häufig zusammen genannte Eigennamen oder für einen bestimmten Sachverhalt typische Adjektive (Völkl et al. 2022). Andererseits können Themen, die nur implizit im Text vorhanden sind, für dieses Verfahren unsichtbar bleiben (Horstmann 2018, S. 10). Deswegen sind Topics nicht immer mit Themen gleichzusetzen.

Neben dem bekanntesten probabilistischen Modell LDA (*Latent Dirichlet Allocation*) und dem NMF-Zugang (*Non-negative Matrix Factorization*) wird in den letzten Jahren zunehmend auch das transformerbasierte BERT-Framework (Devlin et al. 2018) angewendet, das sich als ein neues State-of–the-Art-Modell im Bereich von Natural Language Processing etabliert. Während konventionelle Modelle wie LDA und NMF Dokumente als Bag-of-Words (ungeordnete Listen von Wörtern) behandeln, berücksichtigen BERT-Modelle wie etwa das BERTopic die semantische Relation zwischen Wörtern (Grootendorst 2022, S. 1) und können flexibler auf Texte unterschiedlicher Längen oder Sprachen angewendet werden. Zu Nachteilen von BERTopic zählt aber, dass nur ein Topic pro Dokument vergeben (Grootendorst 2022, S. 8) und eine höhere rechnerische Leistung erfordert wird.

Als Verfahren der digitalen Textanalyse wird Topic Modeling in den Digitalen Geisteswissenschaften seit 2011 verstärkt eingesetzt (Du 2019). Es kann eine korpusübergreifende Perspektive schaffen, weshalb es sich gut für explorative Analysen eignet. Die Anwendungsfälle sind so vielfältig wie der Fachbereich selbst: Dazu zählen unter anderem die Erkennung von Sub-Genres in Dramen (Schöch 2017), die Identifizierung vom genderspezifischen Diskurs in Zeitschriften der Aufklärungszeit (Völkl et al. 2022) sowie die Auffindung von Themen in Briefen (Bleier 2015), Gedichten (Navarro-Colorado 2018) oder Romanen (Haverals/Geybels, 2021).

Als probabilistisches Bag-of-Words-Modell liefert die Methode bei einzelnen Durchläufen leicht voneinander abweichende Ergebnisse, was sich in der un-

terschiedlichen Gewichtung einzelner Wörter pro Topic äußert. Dies lässt sich über Preprocessing (z. B. über das Erstellen einer Stoppwortliste, Segmentieren langer Texte oder Lemmatisieren) und Anpassen der Parameterwerte (wie Anzahl der Topics) so weit stabilisieren, dass immer die gleichen Topics erkennbar bleiben. Trotzdem stößt Topic Modeling auch auf Kritik in Bezug auf dessen Funktionsweise und Plausibilität, wie etwa bei Shadrova (2021, S. 21), die Topics als unvorhersehbar und reduktionistisch bezeichnet.

Zusammenfassend lässt sich jedoch sagen, dass Topic Modeling zwar einige Einschränkungen aufweist und als 'Black-Box'-Verfahren manchmal für Skepsis sorgt, aber dennoch ein nützliches Werkzeug für die Analyse großer Korpora ist. Es liegt weiterhin an den Forscherinnen und Forschern, sich mit den Texten selbst auseinanderzusetzen, sinnvolle und projektspezifische Entscheidungen über das *Preprocessing* zu fällen und ihr Fachwissen beim Interpretieren der Ergebnisse anzuwenden. Bei richtiger Anwendung kann Topic Modeling wertvolle Einblicke in die Schlüsselkonzepte und -ideen eines Korpus liefern und als Ausgangspunkt für weitere Untersuchungen und Interpretationen dienen.

Literatur:

- Blei, David M.: Topic Modeling and Digital Humanities. In: Communications of the ACM 55: 2012, S. 77-84.
- Bleier, Roman: Topic Modelling des Letters of 1916 Briefkorpus. Graz: 2015, URL: `https://gams.uni-graz.at/o:dhd2015.v.005`.
- Devlin, Jacob; Chang, Ming-Wei; Lee, Kenton; Toutanova, Kristina: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding BERT. In: NAACL-HLT 2019 NAACL-HLT 2019. Minneapolis, Minnesota,US: 2019, S. 4171-4186.
- Du, Keli: A Survey On LDA Topic Modeling In Digital Humanities. In: Book of Abstracts DH2019 DH2019. Utrecht University: 2019.
- Grootendorst, Maarten: BERTopic: Neural topic modeling with a class-based TF-IDF procedure BERTopic: 2022, S. 10.
- Haverals, Wouter; Geybels, Lindsey: Putting the Sorting Hat on J.K. Rowling's Reader: A digital inquiry into the age of the implied readership of the Harry Potter series Putting the Sorting Hat on J.K. Rowling's Reader. In: Journal of Cultural Analytics 6: 2021.
- Topic Modeling. URL: `https://fortext.net/routinen/methoden/topic-modeling`
- Navarro-Colorado, Borja: On Poetic Topic Modeling: Extracting Themes and Motifs From a Corpus of Spanish Poetry On Poetic Topic Modeling. In: Frontiers in Digital Humanities 5: 2018, S. 1-15.

- Schöch, Christof: Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama. In: Digital Humanities Quarterly 11: 2017, S. 1-20.
- Shadrova, Anna: Topic models do not model topics: epistemological remarks and steps towards best practices Topic models do not model topics. In: Journal of Data Mining amp; Digital Humanities 2021: 2021, S. 1-28.
- Völkl, Yvonne; Sarić, Sanja; Scholger, Martina: Topic Modeling for the Identification of Gender-specific Knowledge. In: Journal of Computational Literary Studies 1: 2022, S. 1-27 (online).

Software:

BERTopic, CLARIN-PL Topics, dfrtopics, DARIAH Topics Explorer, DARIAH-DE Topics, Gensim, jsLDA: In-browser topic modeling, MALLET, tmw - Topic Modeling Workflow, Topic Modeling Workflow of the Distant Spectators

Verweise:

Analysemethoden, NER, Distant Reading, Text Mining, Lemmatisierung

Themen:

Natural Language Processing

Zitiervorschlag:

Sarić, Sanja; Krušić, Lucija. 2021. Topic Modeling. In: KONDE Weißbuch. Hrsg. v. Helmut W. Klug unter Mitarbeit von Selina Galka und Elisabeth Steiner im HRSM Projekt "Kompetenznetzwerk Digitale Edition". URL: https://gams.uni-graz.at/o:konde.229