Beyond LDA

How to integrate Topic Modeling into the Historian's Toolbox

Melanie Althage, M.A., Professur für Digital History, Humboldt-Universität zu Berlin

Digital History
Humboldt-Universität zu Berlin

HUMBOLDT-UNIVERSITÄT ZU BERLIN

Ausgangslage

Die Quellen der Gegenwart sind digital

Digitale Quellen legen als maschinell prozessierbare Daten den Einsatz digitaler Methoden nahe.

Dissertation

Titel: „Mining the Historian's Web – Methodenkritische Reflexion quantitativer Verfahren zur Analyse genuin digitaler Quellen am Beispiel der historischen Fachkommunikation"

Ziel:

i) Untersuchung der Adaptierbarkeit etablierter Textanalysemethoden der Digital Humanities und Computerwissenschaften für historische Quellen und Forschungsfragen

ii) Ausarbeitung eines Kriterienkatalogs zu ihrer Integration in den Werkzeugkasten der Historiker:innen

Bedarf

Methodenkritik

- ☐ Prüfung: Welches Verfahren passt zu meiner Fragestellung?
- ☐ Entwicklungskontext: Ursprüngliche Erkenntnisinteressen und Ziele
- ☐ Formalisierung welcher theoretisch-methodologischen Annahmen?
- ☐ Auf Basis welcher Art von Daten entwickelt und optimiert?
- ☐ Welche Parameter gibt es? Wie wirken sie sich auf den Output aus?

Beispiel aus dem digitalen Werkzeugkasten

Zur (diachronen) Untersuchung von Diskursen, Motiven, thematischen Trends, u.v.m. in den digitalen Geistes- und Geschichtswissenschaften etabliert:

Topic Modeling

… ist eine auf **Wahrscheinlichkeitsrechnung** basierende Methode aus dem Bereich des unüberwachten maschinellen Lernens zur **Gruppierung von Dokumenten** einer Textsammlung anhand ihrer **gemeinsamen Sprachgebrauchsmuster**.

Entwicklungskontext: Information Retrieval (IR) und Natural Language Processing (NLP)

Ursprüngliches Ziel: Beschreibung und Exploration umfangreicher Datenkollektionen zur Optimierung von Suchheuristiken

Intuition gemäß distributioneller Semantik: Über die gemeinsame Vorkommenshäufigkeit lexikalischer Einheiten (z.B. Wörter) in einem bestimmten Kontext (z.B. Dokumenten) lassen sich Aussagen über den Inhalt der Dokumente formulieren.

Der Quasi-Standard für Topic Modeling:

Latent Dirichlet Allocation (Blei et al. 2003)

Leicht zu implementieren dank gebrauchsfertiger Tools und niedrigschwelliger Tutorials

Herausforderungen für geschichtswissenschaftliche Anwendungsfälle

Die **Kontextabhängigkeit** und insb. **Historizität** der Daten wird bei der Modellierung nicht berücksichtigt:

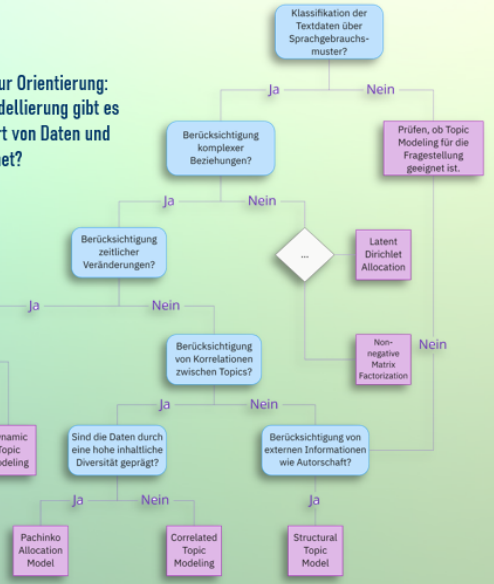
- → Topics sind über das gesamte Korpus statisch
- → keine Berücksichtigung von Inter-Topic-Beziehungen
- → Indifferenz ggü. Ordnung des Korpus
- → keine Berücksichtigung externer Informationen

Populäre Verfahren wie LDA stammen aus Forschungsbereichen mit je eigenen theoretisch-methodologischen Annahmen und Erkenntnisinteressen, die im Kontrast zu geschichtswissenschaftlichen Forschungstraditionen stehen können.

Problem

Work in Progress:

Guidelines und Best Practices zur Orientierung: Welche Verfahren zur Topic-Modellierung gibt es über LDA hinaus? Für welche Art von Daten und Fragestellungen sind sie geeignet?



Literatur (Auswahl): Andorfer, Peter: Turing Test für das Topic Modeling. Von Menschen und Maschinen erstellte inhaltliche Analysen der Korrespondenz von Leo von Thun-Hohenstein im Vergleich, in: ZfdG 2 (2017), DOI: 10.17175/2017_002; Bergamini, Enrico / Mourlon-Druol, Emmanuel: Talking about Europe: Exploring 70 Years of News Archives, in: Bruegel Working Paper 4 (2021), URL: https://www.bruegel.org/2021/03/talking-about-europe-exploring-70-years-of-news-archives/; Blei, David M. / Ng, Andrew Y. / Jordan, Michael I.: Latent Dirichlet Allocation, in: Journal of Machine Learning Research 3 (2003), S. 993–1022; Fechner, Martin / Weiß, Andreas: Einsatz von Topic Modeling in den Geschichtswissenschaften: Wissensbestände des 19. Jahrhunderts, in: ZfdG 2 (2017), DOI: 10.17175/2017_005; Maier, Daniel et al.: Applying LDA Topic Modeling in Communication Research: Toward a Valid and Reliable Methodology, in: Communication Methods and Measures 12, 2–3 (2018), S. 93–118, DOI: 10.1080/19312458.2018.1430754; Uglanova, Inna / Gius, Evelyn: The Order of Things. A Study on Topic Modeling of Literary Texts, in: CEUR Workshop Proceedings 2723 (2020), S. 57–76; Vayansky, Ike / Kumar, Sathish A. P.: A review of topic modeling methods, in: Information Systems 94 (2020), DOI: 10.1016/j.is.2020.101582.