Topic Modeling der Hugo-Schuchardt-Korrespondenz – Möglichkeiten und Grenzen

Saric, Sanja
sanja.saric@uni-graz.at
Universität Graz, Österreich

Scholger, Martina
martina.scholger@uni-graz.at
Universität Graz, Österreich

Einleitung

Digitale Analyseverfahren verändern immer intensiver die Forschungsweise der GeisteswissenschaftlerInnen und mit dem wachsenden Spielraum der Methoden wächst auch die Anzahl an Fragen, die sich vor allem an den Grad der Genauigkeit und wissenschaftliche Relevanz dieser Methoden richtet. Das Topic Modeling gewinnt als eine Methode für automatische Erkennung von versteckten thematischen Strukturen in großen Textmengen (Blei 2012: 8) immer mehr an Beliebtheit, erweckt aber auch Unsicherheiten. Daher beschäftigt sich diese Arbeit mit den Möglichkeiten und Problemen des Topic Modeling am Beispiel von Briefen und stellt unter anderem die Fragen, 1) wie Topic Modeling in der Analyse von Briefkorpora eingesetzt werden kann und 2) wie die Qualität der Ergebnisse dieses Prozesses beeinflusst werden kann.

Forschungsmaterial

Das Forschungsmaterial besteht aus Briefen des Grazer Sprachwissenschaftlers Hugo Schuchardt (1842-1927). Die umfangreiche und mehrsprachige Korrespondenz dieses schon seinerzeit sehr geschätzten Wissenschaftlers ist seit 2007 Teil des Digitalisierung-Projektes *Hugo Schuchardt Archiv* (Hurch 2019). Für die Topic-Modeling-Analyse werden 2261 Briefdateien im TEI-Format in Betracht gezogen, da die restlichen zurzeit noch in keinem entsprechenden Format vorhanden sind. Der Vorteil einer solchen Methode ist es aber, dass das gleiche Modell jederzeit auf eine erweiterte Menge an Daten anwendbar ist. Eine Besonderheit dieses Korpus ist, dass Schuchardt in mehreren Sprachen korrespondiert hat, von denen hier elf repräsentiert sind (Abbildung 1). Daher wird das Modell für einzelne Sprachen separat angewendet. Dies ist insofern eine Herausforderung, weil 1) Vorgänge den jeweiligen Sprachen angepasst werden müssen (wie etwa die Lemmatisierung), 2) der Textumfang bei vielen Sprachen nicht ausreichend ist und daher nicht auf alle Sprachen effektiv angewendet werden kann und 3) die verschiedenen Ergebnisse pro Sprache verglichen werden sollten. Ein weiteres Problem für das Topic Modeling ist die große Diskrepanz in den Textlängen der einzelnen Dateien (Abbildung 2), da die Korrespondenz auch kürzere Formen wie Postkarten und Telegramme beinhaltet. So enthalten etwa die kürzesten deutschsprachigen Dateien etwa drei Tokens, die längste jedoch 3947. Dies ist aber ein Zustand, den viele Briefkorpora in der Realität begegnen, da wir als ForscherInnen selten einem ‚idealen' Korpus gegenüberstehen. Die Auseinandersetzung mit solchen Problemen ist ein fester Bestandteil unserer Arbeit.

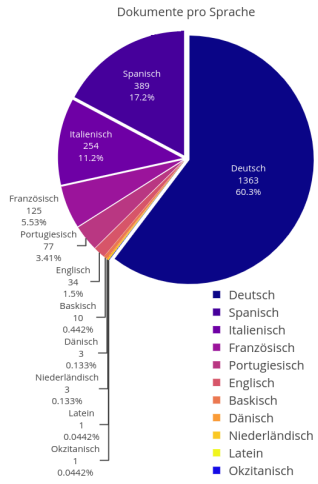

Abbildung 1: Anteil der einzelnen Sprachen im Briefkorpus

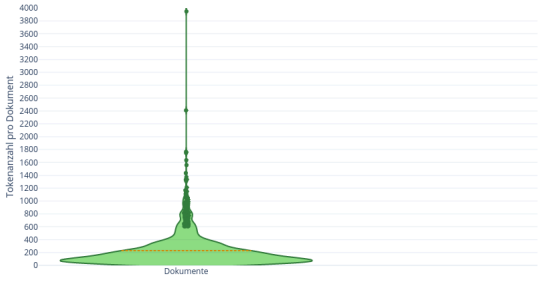

Abbildung 2: Menge der deutschsprachigen Briefdateien nach ihrer Anzahl der Tokens

Methode

Für die Beantwortung der Forschungsfragen war zuerst die Literaturrecherche nötig, und zwar erstens zum Topic Modeling, zweitens zur Textsorte Brief und drittens zu dieser Korrespondenz. Um eine genauere Vorstellung zum Forschungsstand des Topic Modeling

zu bekommen, wurden wissenschaftliche Aufsätze und Anwendungsbeispiele in Betracht gezogen, wie etwa Blei 2010, Jagarlamudi/Daumé 2010, Boyd-Graber/Blei 2012, Riddell 2015, Vuli# et al. 2015, Bock et al. 2016, Andorfer 2017, Fechner/Weiß 2017, Schöch 2017, Murakami et al. 2017 und Arora et al. 2018. Zudem wird am genannten Korpus Topic Modeling mit Hilfe der Programmiersprache *Python* (Python Software Foundation 2001-2019), der Software MALLET (McCallum 2002-2019) und der Anweisungen der Jupyter-Notebooks von DARIAH-DE (DARIAH-DE 2019) vollzogen. Darüber hinaus werden verschiedene Tools zur Vorverarbeitung evaluiert – z. B. *spaCy* (Explosion AI 2019) und DTA::CAB (Berlin-Brandenburgische Akademie der Wissenschaften 2011-2018) für die Lemmatisierung – sowie verschiedene Tools und Parameter für die Topic-Modellierung – z. B. *Topics Explorer* (DARIAH-DE 2018) – und die daraus resultierenden Ergebnisse und Erfahrungen verglichen.

Ergebnisse

Obwohl es sich um ein laufendes Projekt handelt, gibt es bereits einige relevante Ergebnisse und Schlussfolgerungen.

1) Die Vorverarbeitung stellt einen wichtigen Schritt in der Topic-Modellierung dar und beeinflusst die Ergebnisse. Dabei spielen nicht nur die eingesetzten Tools eine Rolle, sondern auch die gewählte Vorgehensweise.

2) Die Lemmatisierung, auf die beim Topic Modeling oft verzichtet wird, ermöglicht mehr semantische Differenz in den Topics.

3) Der unterschiedliche Textumfang von einzelnen VerfasserInnen kann zu falschen Ergebnissen führen, wenn die Topics pro VerfasserIn analysiert werden.

4) Entscheidungen über Parameter wie Optimierungsintervall, Topic- und Iterations-Anzahl können die Ergebnisse beeinträchtigen und müssen immer projektspezifisch getestet werden, bis ein sinnvolles Resultat vorliegt. Das ‚Sinnvolle' zu erkennen ist eine Herausforderung, die fachwissenschaftliches Verständnis verlangt.

Die Inkonsistenz der Topics und manchmal verwirrende Ergebnisse zeigen, dass die naive Anwendung eines Topic-Modeling-Tools nicht immer befriedigend sein kann. Intensivere Beschäftigung mit den einzelnen Schritten und Ergebnissen kann sich jedoch positiv auf den Erfolg der Analyse auswirken. Die weitere Arbeit wird zeigen, ob und welchen Mehrwert Topic Modeling bei der Analyse der Schuchardt-Korrespondenz leisten kann, die durch *close reading* nicht erreicht werden können.

Bibliographie

Andorfer, Peter (2017): "Turing Test für das Topic Modeling. Von Menschen und Maschinen erstellte inhaltliche Analysen der Korrespondenz von Leo von Thun-Hohenstein im Vergleich", in: *Zeitschrift für digitale Geisteswissenschaften* 2. http://zfdg.de/2017_002 [letzter Zugriff 27. September 2019].

Arora, Sanjeev / Ge, Rong; Halpern, Yoni / Mimno, David / Moitra, Ankur / Sontag, David / Wu, Yichen / Zhu, Michael (2018): "Learning topic models - provably and efficiently", in: *Communications of the ACM* 61 / 4: 85–93. 10.1145/3186262.

Berlin-Brandenburgische Akademie der Wissenschaften (ed.) (2011-2018): *Das DTA-Basisformat*. http://www.deutschestextarchiv.de/doku/basisformat/ [letzter Zugriff 27. September 2019].

Blei, David M. (2010): "Introduction to Probabilistic Topic Models", in: *Semantic Scholar*. https://pdfs.semanticscholar.org/5f10/38ad42ed8a4428e395c96d57f83d201ef3b [letzter Zugriff 27. September 2019].

Blei, David M. (2012): "Topic Modeling and Digital Humanities", in: *Journal of Digital Humanities* 2 / 1: 8–11. http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/ [letzter Zugriff 27. September 2019].

Bock, Sina / Du, Keli / Huber, Michael / Pernes, Stefan / Pielström, Steffen (2016): *Der Einsatz quantitativer Textanalyse in den Geisteswissenschaften. Bericht über den Stand der Forschung.* (= DARIAH-DE working papers 18). Göttingen: GOEDOC – Dokumenten- und Publikationsserver der Georg-August-Universität Göttingen. http://webdoc.sub.gwdg.de/pub/mon/dariah-de/dwp-2016-18.pdf [letzter Zugriff 27. September 2019].

Boyd-Graber, Jordan / Blei, David (2012): *Multilingual Topic Models for Unaligned Text*. http://arxiv.org/pdf/1205.2657v1 [letzter Zugriff 27. September 2019].

DARIAH-DE (2018): *Topics Explorer*. V. 2.0.1. https://github.com/DARIAH-DE/TopicsExplorer [letzter Zugriff 27. September 2019].

DARIAH-DE (2019): *DARIAH Topics. Easy Topic Modeling in Python*. V. 2.0.1. https://github.com/DARIAH-DE/Topics [letzter Zugriff 27. September 2019].

Explosion AI (2019): *spaCy*. V. 2.1.6. https://github.com/explosion/spaCy [letzter Zugriff 27. September 2019].

Fechner, Martin / Weiß, Andreas (2017): "Einsatz von Topic Modeling in den Geschichtswissenschaften: Wissensbestände des 19. Jahrhunderts", in: *Zeitschrift für digitale Geisteswissenschaften* 2. http://zfdg.de/2017_005 [letzter Zugriff 27. September 2019].

Hurch, Bernhard (2019): "Hugo Schuchardt Archiv". Institut für Romanistik, Karl-Franzens-Universität Graz (ed.). https://schuchardt.uni-graz.at [letzter Zugriff 27. September 2019].

Jagarlamudi, Jagadeesh / Daumé, Hal (2010): "Extracting Multilingual Topics from Unaligned Comparable Corpora", in: Gurrin, Cathal (ed.): *Advances in information retrieval. Proceedings* 444–456. (= Lecture notes in computer science 5993). Berlin / Heidelberg / New York: Springer.

McCallum, Andrew Kachites (2002-2019): *MALLET. A Machine Learning for Language Toolkit.* V. 2.0.8. http://mallet.cs.umass.edu [letzter Zugriff 27. September 2019].

Murakami, Akira / Thompson, Paul / Hunston, Susan / Vajn, Dominik (2017): "'What is this corpus about?': using topic modelling to explore a specialised corpus", in: *Corpora* 12 / 2: 243–277. https://www.euppublishing.com/doi/10.3366/cor.2017.0118 [letzter Zugriff 27. September 2019].

Python Software Foundation (2001-2019): *Python*. V. 3.7.4. https://github.com/python [letzter Zugriff 27. September 2019].

Riddell, Allen (2015): *Text Analysis with Topic Models for the Humanities and Social Sciences — Text Analysis with Topic Models for the Humanities and Social Sciences*. DARIAH-DE Initiative (ed.). https://liferay.de.dariah.eu/tatom/ [letzter Zugriff 27. September 2019].

Schöch, Christof (2017): "Topic Modeling Genre. An Exploration of French Classical and Enlightenment Drama", in: *Digital Humanities Quarterly* 11 / 2. http://www.digitalhumanities.org/dhq/vol/11/2/000291/000291.html [letzter Zugriff 27. September 2019].

Vuli#, Ivan / Smet, Wim de / Tang, Jie / Moens, Marie-Francine (2015): "Probabilistic topic modeling in multilingual settings. An overview of its methodology and applications", in: *Information Processing & Management* 51 / 1: 111–147. https://www.sciencedirect.com/science/article/pii/S0306457314000739 [letzter Zugriff 27. September 2019].