Ein GUI-Topic-Modeling-Tool mit interaktiver Ergebnisdarstellung

Severin, Simmler
severin.simmler@stud-mail.uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Deutschland

Thorsten, Vitt
thorsten.vitt@uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Deutschland

Pielström, Steffen
pielstroem@biozentrum.uni-wuerzburg.de
Julius-Maximilians-Universität Würzburg, Deutschland

Zu den Verfahren der digitalen Textanalyse, die in den vergangenen Jahren in den textbasierten digitalen Geisteswissenschaften etabliert wurden, gehört das LDA (Latent Dirichlet Allocation) Topic Modeling (Blei 2012; Steyvers und Griffiths 2006). Diese Methode eignet sich zur Analyse der Verteilung semantischer Wortgruppen in Textsammlungen, und kann sowohl für die computergestützte Textklassifikation als auch für die explorative Betrachtung der Inhalte eines Corpus herangezogen werden. Um dem zunehmenden Interesse am Topic Modeling von Seiten der DH-Community Rechnung zu tragen, entwickelt DARIAH-DE seit 2017, basierend auf den Python-Bibliotheken "LDA" von Allan Riddell und "DARIAHTopics" (Jannidis et al. 2017), den DARIAH-TopicsExplorer, eine Software, mit der interessierte Forschende Topic Modeling auf ihren eigenen Rechnern an ihren eigenen Texten ausprobieren können, und die den gesamten Analyseprozesses, vom unverarbeiteten Text bis zum Ergebnis, durch eine graphische Nutzeroberfläche (GUI) unterstützt.

Der TopicsExplorer bietet nicht die Leistungsfähigkeit und vor allem die Flexibilität bisheriger Lösungen, die entweder, wie das weit verbreitete MALLET (McCallum 2002), als Kommandozeilenprogramme, oder aber, wie Gensim (Rehurek und Sojka 2010), als Bibliothek für eine Programmiersprache konzipiert wurden. Dafür kann er aber ohne Kenntnis irgendeiner Programmier- oder Skriptsprache eingesetzt werden und muss nicht einmal über die Kommandozeile aufgerufen werde. Damit schließt der TopicsExplorer eine wichtige Lücke: Forschende, die selbst nicht, oder noch nicht, programmieren, können sich hier einen Eindruck davon verschaffen, wie die Methode funktioniert und was sie theoretisch leisten kann. Das befähigt auf der einen Seite, Forschungsarbeiten, die auf Topic Modeling basieren, und ihre Ergebnisse informiert einzuschätzen. Auf der anderen Seite kann das Werkzeug für einfache Fragestellungen, die Topic Modeling erfordern, direkt eingesetzt werden, und bei komplizierteren Ansätzen eine informierte Entscheidung darüber ermöglichen, ob sich die Aneignung der notwendigen technischen Fähigkeiten für die Verwendung einer fortgeschrittenen Lösung für die jeweilige Forschungsfrage überhaupt lohnt.

Der TopicsExplorer wurde in einer ersten Version 2017 und 2018 im Rahmen mehrerer Workshops und Konferenzen verschiedenen Gruppen von Forschenden und Studierenden vorgestellt. Die Erfahrungen aus dem Umgang mit Nutzerinnen und Nutzern in solchen Workshops und vor allem ihr direktes Feedback sind seither umfangreich in die Weiterentwicklung eingeflossen und die daraus resultierenden Änderungen gehen weit über die Beseitigung von Bugs und die Sicherstellung der nachhaltigen Funktionalität hinaus. Aus einem anfänglichen "GUI-Demonstrator" (Simmler et al. 2018), der noch die Installation einer Python-Bibliothek erforderte, auf einem lokalen Server lief und im Browser angezeigt wurde, ist eine vollwertige Stand-Alone-Software geworden, die nach dem Herunterladen ohne weitere Vorbereitung auf gängigen Windows-, MacOS- und Linux-Systemen gestartet werden kann. Die Visualisierungen können interaktiv manipuliert, und die Ergebnisse im csv-Format exportiert werden. Zahlreiche kleinere, von der Testcommunity gewünschte Features, wie z.B. eine Fortschrittsanzeige mit Abbruchbutton (Abb. 1), haben die Usability wesentlich verbessert.

Zur Zeit wird eine Version in grundlegend überarbeitetem Design vorbereitet, deren Ziel es ist, auch komplizierteren, aus den Testcommunies heraus formulierten Ansprüchen an die Interaktivität der Software gerecht zu werden. Auf technischer Ebene wird dabei die Visualisierung der Ergebnisse statt mit der bisher verwendeten Python-Bibliothek "Bokeh" direkt in Javascript realisiert, um zusätzliche Flexibilität für die Umsetzung neuer Funktionalitäten zu gewinnen. Äußerlich wird es damit möglich, Ergebnisse auch nachträglich interaktiv umzusortieren und in mehreren Fenstern darzustellen. Mit dem Ziel, die User-Experience im explorativen Umgang mit den erzeugten Modellen zu verbessern, wird darüber hinaus ein völlig neues Visualisierungskonzept auf Basis der Vorschläge von Chaney und Blei (2012) umgesetzt. Dieses Konzept erlaubt es nicht nur, einzelne Dokumente im Corpus mitsamt den dazugehörigen Analyseergebnissen zu betrachten, es werden darüber hinaus automatisch andere Texte mit ähnlichen inhaltlichen Schwerpunkten vorgeschlagen (Abb. 2).

Wir hoffen, dass der TopicsExplorer mit den angestrebten Verbesserungen und Erweiterungen dazu beiträgt, eine mittlerweile doch recht verbreitete Forschungsmethode aus der Nische derjenigen DH-Verfahren heraus zu holen, die nur von Programmiererinnen und Programmierern verwendet, verstanden und kritisch diskutiert werden. Die neuen Features, die für das kommende Release entwickelt werden, sollten dazu einen wesentlichen Beitrag leisten.

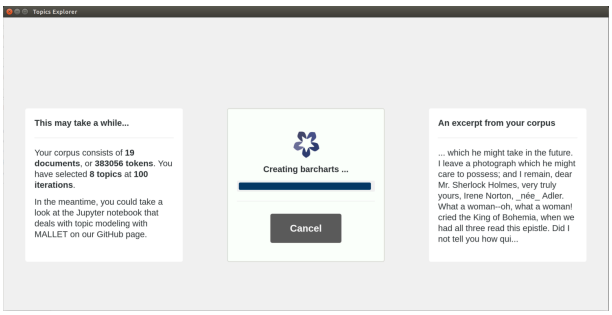

Abbildung 1. Fortschrittsanzeige für die laufende Modellierung im aktuellen TopicsExplorer

Abbildung 2. Übersicht für ein Dokument im Prototypen der Version 2

Bibliographie

Blei, David M. (2012): *Probabilistic Topic Models*, in: Communication of the ACM55, Nr. 4 (2012): 77–84. doi:10.1145/2133806.2133826.

Chaney, Allison J.B. / Blei, David M. (2012): *The Visualizing Topic Models*, in: Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media 419-422.

Jannidis, Fotis/ Pielström, Steffen / Schöch, Christof / Vitt, Thorsten (2017): *Making topic modeling easy: a programming library in Python*, in: Proceedings of the Digital Humanities 2017 Conference.

McCallum, Andrew K. (2002): *MALLET#: A Machine Learning for Language Toolkit. http://mallet.cs.umass.edu.*

Rehurek, Radim/ Sojka, Petr (2010): *Software framework for topic modelling with large corpora.* In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks.

Simmler, Severin / Vitt, Thorsten / Pielström. Steffen (2018): *LDA Topic Modeling über eine graphisches Interface*, in: Konferenzabstracts der 5. Tagung des Verbands Digital Humanities im deutschsprachigen Raum e.V. 428-429.

Steyvers, Mark/ Griffiths, Tom (2006): *Probabilistic Topic Models*, in: Latent Semantic Analysis: A Road to Meaning, herausgegeben von T. Landauer, D. McNamara, S. Dennis, und W. Kintsch. Laurence Erlbaum.