Talk at the Text Encoding Initiative 2022 conference at Newcastle University, Friday 16 September 2022.

How to Represent Topic Models in Digital Scholarly Editions

Ulrike Henny-Krahmer (University of Rostock) and Frederike Neuber (Berlin-Brandenburgische Akademie der Wissenschaften)

Keywords: text mining, topic modeling, digital scholarly editions, data modeling, data integration

Abstract
Topic modeling (Blei et al. 2003, Blei 2012) as a quantitative text analysis method is not part of the *classic* editing workflow as it stands for a way of working with text that in many respects contrasts with critical editing. However, for the purpose of a thematic classification of documents, topic modeling can be a useful enhancement to an editorial project. It has the potential to replace the cumbersome manual work that is needed to represent and structure large edition corpora thematically, as has been done for instance in the projects *Alfred Escher Briefedition* (Jung 2022), *Jean Paul – Sämtliche Briefe digital* (Miller et al. 2018) or the *edition humboldt digital* (Ette 2016).
We apply topic modeling to two edition corpora of correspondence of the German-language authors Jean Paul (1763-1825) and Uwe Johnson (1934-1984), compiled at the Berlin-Brandenburg Academy of Sciences and Humanities (BBAW) and the University of Rostock (Miller et al. 2018, Helbig et al. 2017). In our contribution, we discuss how the results of the topic modeling can be usefully integrated into digital editions. We propose to integrate them into the TEI corpora on three levels: (1) the topic model of a corpus, including the topic words and the parameters of its creation, is modeled as a taxonomy in a separate TEI file, (2) the relevance of the topics for individual documents is expressed in the text classification section of the TEI header of each document in the corpus, and (3) the assignment of individual words in a document to topics is expressed by links from word tokens to the corresponding topic in the taxonomy. Following a TEI encoding workflow as outlined above allows for developing digital editions that include topic modeling as an integral part of their user interface.

Bibliography

Blei, David M., Andrew Y. Ng, and Michael I. Jordan. 2003. “Latent dirichlet allocation.” *The Journal of Machine Learning Research* 3: 993–1022. http://dl.acm.org/citation.cfm?id=944919.944937.

Blei, David M. 2012. “Probabilistic Topic Models.” *Communications of the ACM* 55 (4): 77–84. https://doi.org/10.1145/2133806.2133826.

Ette, Ottmar (ed.). 2016–2021. “edition humboldt digital” (version 7). Berlin: Berlin-Brandenburgische Akademie der Wissenschaften. https://edition-humboldt.de.

Helbig, Holger, Ulrich Fries, and Katja Leuchtenberger (eds.). 2017–2022. *Historisch-kritische Ausgabe der Werke, Schriften und Briefe Uwe Johnsons.* Die digitale Ausgabe verantwortet von Fabian Kaßner, Marc Lemke und Christian Riedel. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften. http://www.uwe-johnson-werkausgabe.de/.

Miller, Norbert, Markus Bernauer and Frederike Neuber. 2018–2022. “Jean Paul – Sämtliche Briefe digital”. Berlin: Berlin-Brandenburgische Akademie der Wissenschaften. https://www.jeanpaul-edition.de.

Jung, Joseph. 2022. “Digitale Briefedition Alfred Escher”. https://briefedition.alfred-escher.ch.