Topic Modelling with Latent Dirichlet Allocation Method in Social Sciences – Case Study of Web of Science

Maja Buhin Pandur

Faculty of Organization and Informatics Varaždin, University of Zagreb

Topic modelling is an unsupervised machine learning technique that automatically analyses text data to create topics for a set of documents. Presently, Latent Dirichlet Allocation (LDA) is the most popular technique for topic modelling. One application of LDA in Natural Language Processing (NLP) is to discover “hidden” topics from different documents based on words or expressions with similar meanings.

The research aims to investigate research topics in social sciences through the number of terms from titles, abstracts, and authors’ keywords. We applied Structural Topic Model with LDA on the Web of Science (WoS) Core Collection database by searching articles containing phrase social network* in the WoS Social Science research area from 1999 to 2019 to extract topics from 3,664 scientific papers. We also compared the topics with centroids of WoS categories by using cosine similarity. The results show that an optimal number of topics coincides with the existing number of research areas defined in Social Science or its integer multiple. From cosine similarity between vectors based on word probability distribution from topics and centroids from WoS categories, we can see that some topics have something in common with WoS categories. These results open an area for research into comparing the existing taxonomy and the taxonomy proposed by the LDA model and for the future identification of interdisciplinarity.