UNSUPERVISED MACHINE LEARNING APPROACH FOR HIERARCHICAL GRAPH-BASED REPRESENTATION OF NATURAL LANGUAGE TEXT COLLECTIONS

Jevgenijs Bodrenko[1]

[1] *Transport and Telecommunication Institute*
Lomonosova 1, Riga, LV-1019, Latvia
eugene.bodrenko@gmail.com

Keywords: hierarchical topic modeling, spectral clustering, natural language processing, visualization

In the modern digital era, the ever-growing volumes of data require advanced tools for efficient navigation and knowledge extraction. Documents written in human language to be consumed by human readers present an interesting dimension of this problem. Writing a comprehensive literature review is but one example where the need might arise to analyze collections on the merits of hundreds of texts.

Recent advances in natural language processing, particularly the development of Large Language Models (LLMs) allowed to create powerful tools to aid with this kind of tasks. The price of these advancements is the need for high amounts of labeled data, computational resources and specialized skills to train and/or fine-tune the model for the task at hand. Aiming to address the resource-related drawbacks of LLM-based tools, current work presents a natural language processing (NLP) pipeline to detect the potential presence of a topic hierarchy in a collection of human language texts, focusing specifically on full texts of scientific publications.

Three research questions were formulated:

1) How to discover the topic hierarchy in a collection of English texts using unsupervised machine learning methods, given that it was discovered by a human?

2) What model and quality metrics allow to ensure that topics are understandable, insightful about the structure of the collection and share interpretable connections between different levels of the hierarchy?

3) How can the output be visualized and used to explore the hierarchy structure and document similarity along and across the topic hierarchy?

The NLP pipeline was constructed to support an unsupervised, hierarchical topic model. The model was based on the additive regularization architecture proposed by (Chirkova, 2016) and additionally enhanced by introducing a variation of architecture suggested by (Khodorchenko *et al.*, 2020) for increased topic interpretability. It was discovered that the resulting pipeline allows to infer a hierarchy of human-interpretable topics from collections of texts, even without extensive hyperparameter tuning. The resulting model also allowed to generate probabilistic topic-based vector representation for each document at every level of the resulting hierarchy. Such soft clustering results were converted into hard clustering results by calculating pairwise document similarity as Bhattacharyya coefficient through Hellinger distance (HD) (Kitsos and Nisiotis, 2022), and applying the Spectral Clustering (SC) algorithm to the resulting similarity matrix.

In order to allow for exploration of document similarity both for a given level of the hierarchy, and between different levels, two types of visualizations were developed. The first one was based on visualizing HDs between documents by applying Multidimensional Scaling to the HD matrix for documents at a given level. This allowed to construct a scatter plot indicating HD-based document dissimilarity in terms of topic content in a 2-dimensional space. The plot was enhanced by adding connections between “topically-similar” documents, given a HD threshold value. The second one was based on assigning the sequence of SC-generated cluster labels at all levels of the topic hierarchy as attributes of each document in the collection.

This allowed to trace out the relatedness of documents and clusters between the layers of the hierarchy using Sankey plot.

The resulting pipeline coupled with the developed visualization was used to explore a potential topic hierarchy in two custom datasets consisting of 50 open-access scientific publication full texts each. The results suggest that the pipeline can be useful to search for groups of topically-related texts, estimate the degree to which a given collection can be meaningfully represented by a topic hierarchy, and to generate an informative visual topical summary of the collection. Potential applications include creating a collection-based topic map for a self-study process, a literature review or a meta-analysis both in academia and industry.

The research is supervised by Dr. sc. ing., Professor Jackiva Irina

References

1. Chirkova, N.A. (2016) Additive Regularization for Hierarchical Multimodal Topic Modeling, *Machine Learning and Data Analysis*, 2, pp. 187–200. Available at: https://doi.org/10.21469/22233792.2.2.05.
2. Khodorchenko, M. *et al.* (2020) Optimization of Learning Strategies for ARTM-Based Topic Models, in. Available at: https://doi.org/10.1007/978-3-030-61705-9_24.
3. Kitsos, C.P. and Nisiotis, C.-S. (2022) Considering distance measures in Statistics, *Biometrical Letters*, 59(1), pp. 65–75. Available at: https://doi.org/10.2478/bile-2022-0006.