Hierarchical Topic Modeling over Financial Documents

Abstract

The objective of this project is to leverage unsupervised learning models to unveil the structure of the Enron email dataset. We focus on two tasks: hierarchical topic modeling and topic evolution in email threads. For this end, term frequency models like LDA, word embedding approaches like BERTopic, and an ensemble of the two are proposed. The ensemble model is achieved by combining BERT embeddings and LDA topic proportions to generate new word embeddings. These models are then compared with coherence and diversity metrics, as well as interpretability, finding that different models are better suited for different tasks.

BERTopic is selected for the hierarchical analysis due to its higher metrics and contextual nature and LDA is selected for topic evolution due to its better interpretability. Exploration beyond these models is also performed by estimating a hierarchical version of LDA. The results of the models are promising, all of them show consistent words among their topics. Additionally, it was found that the interpretability of topics strongly depends on the data cleaning process: a bag of words approach for LDA, and embeddings approach for BERTopic.