BERTopic: Neural topic modeling with a class-based TF-IDF procedure

Group Member Names :

1. Shijun Ju
2. Dhaval Vasant

Introduction

This project is a replication of the research paper titled "BERTopic: Neural topic modeling with a class-based TF-IDF procedure" (ArXiv, 2022 (https://arxiv.org/abs/2203.05794)). The goal is to reproduce the experiments and methodologies discussed in the paper, which presents BERTopic, a novel topic modeling technique that leverages embeddings and clustering algorithms to generate coherent topics.

BERTopic aims to improve topic modeling by using a combination of Transformers, Dimension Reduction, Clustering, Count Vectorizing, and TF-IDF procedure, providing a more accurate and interpretable set of topics from large text corpora.

AIM/PROBLEM STATEMENT/CONTEXT OF THE PROBLEM

- Replicate results in the original paper
- Use BERTopic to categorize and understand Economics multiple choice questions from CAIE past papers

Issues with Manual Categorization

- Time consuming - over 3000 questions
- Too many topics and subtopics to differentiate - over 100 economic concepts
- Inconsistency, Minor adjustments necessitating sequential modifications: as an educator gained experience, the understanding of categorization changes over time, thus requires constant adjustments.

Github Repo: url (https://github.com/shj37/BERTopic-paper-replication)

DESCRIPTION OF PAPER/Methodology

The BERTopic model combines several techniques to achieve its results:

1. **Embeddings**: Sentence embeddings are generated using pre-trained transformer models like BERT.
2. **Dimensionality Reduction**: UMAP (Uniform Manifold Approximation and Projection) is applied to reduce the dimensionality of the embeddings.
3. **Clustering**: HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) is used to cluster the reduced embeddings.
4. **Topic Representation**: A class-based TF-IDF procedure is employed to create a representative topic for each cluster.

This methodology allows BERTopic to efficiently and accurately model topics in a given text corpus, ensuring that the resulting topics are both coherent and interpretable.

SOLUTION

- All evaluations are carried out in Kaggle notebook thus were contrained by the Kaggle's 12-hour training limit, thus we could not finish running LDA sequence model. Similarly, CTM evaluations for 20NewsGroup and BBC News datasets are also skipped.
- We skipped all models using `Top2Vec` or `Doc2Vec` models due to unsolved conflicts caused by the older version of the `gensim` library.
- We did not test Wall Time of models as packages/libraries have changed significantly in the last two years, so the results won't be comparable with the original paper.

Extension

- Customized stop-words
- 1-4 n-grams
- Topic-word diversity

Background

Datasets:

- 20NewsGroup: https://github.com/MIND-Lab/OCTIS/tree/master/preprocessed_datasets/20NewsGroup (https://github.com/MIND-Lab/OCTIS/tree/master/preprocessed_datasets/20NewsGroup)
- BBC News: https://github.com/MIND-Lab/OCTIS/tree/master/preprocessed_datasets/BBC_News (https://github.com/MIND-Lab/OCTIS/tree/master/preprocessed_datasets/BBC_News)
- Trump tweets: https://www.thetrumparchive.com/faq (https://www.thetrumparchive.com/faq)

Implement paper code

- Model Evaluation: https://github.com/shj37/BERTopic-paper-replication/blob/main/bertopic-paper-replication.ipynb (https://github.com/shj37/BERTopic-paper-replication/blob/main/bertopic-paper-replication.ipynb)

- Result Summary: https://github.com/shj37/BERTopic-paper-replication/blob/main/replication_results.ipynb (https://github.com/shj37/BERTopic-paper-

Contribution Code

- https://github.com/shj37/BERTopic-paper-replication/blob/main/bertopic-extension.ipynb (https://github.com/shj37/BERTopic-paper-replication/blob/main/bertopic-extension.ipynb)

Results

Observations:

- For LDA, NMF, and BERTopic-MPNET, our results are almost the same as the orginal paper. The CTM for Trump data is also close to the original.
- The BERTopic-MPNET outperform LDA and NMF. Though CTM has higher diversity score than BERTopic-MPNET, it suffered the same issue of significantly longer training time as in the paper.
- Our results using the three embedding models are similar to the paper.
- Our results for the Trump dynamic topic modelling are similar to the paper.
- Extension:
  - mostly successfully categorizing questions into major economic topics

Conclusion and Future Directions

Learnings

The project demonstrates the effectiveness of BERTopic in generating coherent and interpretable topics from text corpora. Future work could involve exploring the impact of different embedding models or applying BERTopic to other domains such as social media analysis or customer feedback.

Result discussion and limitations

The analysis revealed a significant challenge with the categorization process. Approximately 20% to 25% of the questions could not be classified into any categories. This high rate of uncategorized questions suggests that our current classification system may not be comprehensive enough, and it highlights a need for refining our categorization criteria or expanding the classification schema.

Additionally, the data processing revealed that non-essential nouns, such as "cars" and "country," are frequently selected by the current stopword filter. This indicates that the stopwords list is not adequately tailored to our specific dataset, leading to irrelevant words being included in the analysis. To address this issue, a thorough review and modification of the stopwords for different

knowledge domain is necessary. This will involve identifying and excluding more context-specific terms that do not contribute meaningfully to the categorization process, thus improving the accuracy and efficiency of the data handling.

Future extension

To enhance the categorization system further, integrating Large Language Models (LLMs) could be a promising approach. By leveraging LLMs, we can generate new questions based on the summarized categories and example documents. This would involve training the model on the

References

M. Grootendorst, 2022, BERTopic: Neural topic modeling with a class-based TF-IDF procedure, https://arxiv.org/abs/2203.05794 (https://arxiv.org/abs/2203.05794)

Type *Markdown* and LaTeX: α^2