Comprehensive Documentation of Clustering of Constitutions Using LLM Embeddings

Project Overview

The project aimed to cluster a set of constitutional documents using state-of-the-art embeddings generated by Large Language Models (LLMs). The primary objective was to explore how documents could be grouped based on their semantic content, as understood by advanced natural language processing models.

Methodology

Dataset:

- The dataset comprised 17 constitutional documents from various countries, converted from PDF to text format for processing.

Embedding Generation:

- Two different LLMs were used to generate embeddings: sentence transformers and BERT (Bidirectional Encoder Representations from Transformers).
- Sentence transformers were initially used, known for capturing broad semantic themes.
- BERT, specifically the 'bert-base-uncased' model, was later employed to capture deeper contextual relationships within text.

Clustering:

- K-Means clustering algorithm was applied to the embeddings from both models to group the documents.
- The number of clusters was set to 5, based on the dataset size and preliminary analysis.

Results

Sentence Transformers Clustering:

- Resulted in clusters that appeared to group documents more on regional and religious lines.
- Fewer single-document clusters, suggesting a broader interpretation of semantic similarity.

BERT Embeddings Clustering:

- Produced more evenly distributed clusters across documents.
- Some constitutions, like those of the United States and Argentina, formed their own unique clusters, indicating distinct features.
- Other clusters contained a mix of countries, not strictly along regional or religious lines.

Analysis

- BERT vs. Sentence Transformers:

    - BERT's deep contextual understanding led to more nuanced clusters, potentially more suitable for analyses requiring fine distinctions.
    - Sentence transformers' broader semantic capture resulted in clusters formed along more general themes.
- Content Analysis:
    - A detailed review of the documents within each cluster is recommended to understand the basis for their grouping better.
- Model Limitations:
    - BERT's token limit might affect its representation of longer documents.
    - The training corpus and design of each model influence their understanding and representation of legal texts.

Conclusion

- The clustering of constitutional documents using LLM embeddings revealed meaningful insights into how these advanced models perceive and group legal texts.
- The project highlighted the importance of choosing an appropriate LLM based on the analysis goals – whether focusing on detailed legal nuances or broader thematic similarities.
- Future explorations could involve fine-tuning LLMs on legal-specific corpora or experimenting with other models for potentially more accurate clustering in the legal domain.

Recommendations for Future Work

- Fine-Tuning on Legal Texts: Applying models that have been fine-tuned on legal corpora might yield more accurate clustering.
- Experimentation with Other Models: Exploring other LLMs, like GPT-3 or legal-specific BERT variants, could provide different perspectives.
- Incorporation of Expert Analysis: Collaborating with legal experts could enhance the interpretation of clustering results.