Project Proposal

Analyzing UN Speeches' Transcripts (1971 – 2018)

Ritesh Kumar

2024SP_MS_DSP_453-DL_SEC61: Natural Language Processing

Module 6

Project P.3

Nethra Sambamoorthi and Sudha BG

June 1, 2024

Project Objective:

The objective of this project is to analyze the evolution of topics in United Nations General Assembly (UNGA) speeches from 1971 to 2018. By applying Latent Dirichlet Allocation (LDA), a widely used topic modeling technique, we aim to uncover the underlying thematic structure within the speeches. This analysis will reveal the dominant topics discussed over nearly five decades and track their prominence over time.

Data Sources:

The data for this project comprises transcripts of UNGA speeches from over 100 countries, spanning the years 1971 to 2018. These speeches have been downloaded as plain text files from the United Nations website, totaling 68.6 MB of text.

Data Preprocessing:

Preprocessing is a crucial step to ensure that the text data is clean, consistent, and suitable for analysis. The following steps will be undertaken for preprocessing the UN speech transcripts:

1. Text Cleaning:
    - Remove Line Numbers: Strip out any line numbers to ensure a smooth, continuous flow of words.
    - Remove Extra Whitespace: Eliminate extra spaces, tabs, and line breaks to maintain consistent formatting.
    - Remove Punctuation: Strip out punctuation marks to avoid interference with text analysis.
    - Lowercase Conversion: Convert all text to lowercase to maintain consistency.
    - Remove Special Characters and Numbers: Eliminate special characters, numbers, and any non-alphanumeric symbols.

2. Tokenization: Split text into individual words (tokens) to analyze the frequency and distribution of words.
3. Stop Words Removal: Remove common stop words that do not add significant meaning to the text.
4. Stemming and Lemmatization: Apply stemming or lemmatization to convert words to their base or root forms.
5. Named Entity Recognition (NER): Identify and extract entities such as countries, organizations, and important figures mentioned in the speeches.
6. Document Segmentation: Segment longer speeches into smaller, coherent parts to improve the granularity of the analysis.

Methodology:

The primary method for this project is Latent Dirichlet Allocation (LDA), which treats each document as a mixture of topics and each topic as a mixture of words. The steps for implementing LDA include:

1. Corpus Creation: Build a text corpus from the pre-processed text data, typically a bag-of-words or term-document matrix.
2. Determine Number of Topics: Use methods such as perplexity scores or coherence measures to determine the optimal number of topics.
3. Run LDA: Apply LDA to the corpus to extract topics, where each topic is represented as a distribution of words, and each document as a distribution of topics.
4. Interpret Topics: Analyze the top words in each topic and assign descriptive labels.
5. Temporal Analysis: Track the evolution of topics over time, visualizing trends using line charts or heatmaps.

Metrics for Performance Measurement:

The performance of the LDA model will be evaluated using the following metrics:

- Perplexity: Measures how well the model predicts a sample.
- Coherence Score: Measures the semantic similarity between high-scoring words in the topic.
- Topic Interpretability: Evaluates how meaningful the topics are to human interpreters.

Novelty of the Methods

This project is novel in its application of LDA to a comprehensive dataset of UNGA speeches spanning nearly five decades. While previous works have utilized LDA for topic modeling in various contexts, this project uniquely focuses on tracking the evolution of global discourse over an extended period. By visualizing the temporal changes in topics, this analysis will provide insights into the shifting priorities and concerns of the international community, offering a valuable resource for policymakers, historians, and researchers.

Expected Outcomes

The expected outcomes of this project include:

- Identification of dominant topics in UNGA speeches from 1971 to 2018.
- Visualization of the evolution and prominence of these topics over time.
- Insights into significant trends and shifts in international discourse.

References:

1. Blei, David & Ng, Andrew & Jordan, Michael. (2001). Latent Dirichlet Allocation. The Journal of Machine Learning Research. 3. 601-608. (Link)
2. Rahul Kumar Gupta, Ritu Agarwalla, Bukya Hemanth Naik, Joythish Reddy Evuri, Apil Thapa, Thoudam Doren Singh, Prediction of research trends using LDA based topic modeling. (Link)
3. A. Goyal and I. Kashyap, "Latent Dirichlet Allocation - An approach for topic discovery," (Link)