Acquire Data

Analyzing UN Speeches' Transcripts (1971 – 2018)

Ritesh Kumar

2024SP_MS_DSP_453-DL_SEC61: Natural Language Processing

Module 4

Project P.2

Nethra Sambamoorthi and Sudha BG

May 30, 2024

Project:

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 Collection:

The transcripts of UNGA speeches, by over 100 the countries, from 1971 to 2018 have been downloaded from the United Nations Website. These speeches have been downloaded as as plain text files. The total size of the text files is 68.6MB.

Data Preprocessing:

Preprocessing is a crucial step in any Natural Language Processing (NLP) project. It ensures that the text data is clean, consistent, and suitable for analysis. Here are the detailed steps for preprocessing the UN speech transcripts:

1. Text Cleaning:

- Remove Line Numbers: Strip out any line numbers that may be present in the text 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, which can interfere with text analysis.
- Lowercase Conversion: Convert all text to lowercase to maintain consistency and avoid treating words with different cases as separate entities.

- Remove Special Characters and Numbers: Eliminate special characters, numbers, and any non-alphanumeric symbols that do not contribute to the textual content.
- Tokenization: Split Text into Tokens: Break down the text into individual words (tokens). This process helps in analyzing the frequency and distribution of words in the text.
- Stop Words Removal: Filter Out Common Words: Remove common stop words (e.g., "and," "the," "is") that do not add significant meaning to the text. This step helps in focusing on the more informative words in the dataset.
- Stemming and Lemmatization: Reduce Words to Their Root Forms: Apply stemming or lemmatization techniques to convert words to their base or root forms (e.g., "running" to "run"). This step helps in reducing the vocabulary size and improving the accuracy of topic modeling.
- Named Entity Recognition (NER): Identify Key Entities: Use NER techniques to identify and extract entities such as countries, organizations, and important figures mentioned in the speeches. This can provide additional context for topic analysis.
- Document Segmentation: Split Documents into Segments: Depending on the length of the speeches, segment them into smaller, coherent parts to improve the granularity of the analysis. This can involve dividing long speeches into paragraphs or sections.

Using LDA for Topic Modeling:

Once the data is pre-processed, the next step is to apply LDA for topic modeling. LDA will help identify the underlying themes in the UNGA speeches and track their changes over time. Here are the steps for implementing LDA:

1. Corpus Creation: Build a Text Corpus: Create a corpus from the pre-processed text data. This involves converting the cleaned text into a format suitable for LDA, typically a bag-of-words or term-document matrix.

2. Determine Number of Topics: Optimal Topic Selection: Use methods such as perplexity scores or coherence measures to determine the optimal number of topics for the LDA model. This helps in balancing the granularity and interpretability of the topics.
3. Run LDA: Train the LDA Model: Apply LDA to the corpus to extract topics. Each topic will be represented as a distribution of words, and each document will be represented as a distribution of topics.
4. Interpret Topics: Label Topics: Analyze the top words in each topic and assign descriptive labels. This step involves interpreting the themes based on the most prominent words in each topic.
5. Temporal Analysis: Track Topic Evolution: Aggregate the topic proportions for each year or decade to visualize how the prominence of different topics has changed over time. Use visualization techniques such as line charts or heatmaps to illustrate these trends.

Expected Outcomes and Insights:

By applying LDA to the UNGA speeches, the project will reveal significant trends and shifts in international discourse. For example, topics related to the Cold War may dominate the early decades, while issues like globalization, terrorism, and climate change become more prominent in later years. The analysis can also uncover shifts in geopolitical focus, such as the rise of emerging economies and changing dynamics in international relations.

These insights can provide valuable information for policymakers, historians, and researchers interested in understanding the historical context and future trends in international relations. The project will contribute to a deeper understanding of the collective concerns and responses of the international community to global challenges.