CIND820: Big Data Analytics Project

PROJECT ABSTRACT SUBMISSION:

Exploring emerging themes in the unstructured historical debates from the Hansard Records of the Canada House of Commons

Student: Colin Lacey | Student ID: 501176114

Professor: Tamar Adbou

Date of Submission: January 22, 2024

Proposed Topic:

Exploring emerging themes in the unstructured historical debates from the Hansard Records of the Canada House of Commons.

Introduction:

The modernization efforts of many institutional and public sector organization's record keeping processes, including the digitization of historical paper records, is creating an opportunity to begin mining information from large volumes of unstructured historical data. The application of machine learning techniques to these datasets, from text mining, topic modelling and text summarization, could allow for the streamlining of the manual effort required to generate relevant metadata to make these records retrievable to wider audiences. Additionally, the ability to leverage topic modelling and text summarization for these previously paper records could open the door to highly targeted queries related to specific content, provide the ability for identifying patterns and trends within a dataset over time, as well as generally supporting more formal meta-analysis research.

Problem Statement:

This project aims to explore what information can be gained through applying text mining approaches to the historical Hansard Records of the Canada House of Commons. Challenges that this dataset may pose to text mining and topic modelling could include the variability in document lengths and content, interpretation and relevancy of content related to parliamentary procedures and traditions, and unknown impacts of filibuster (prolonged speech of sometimes questionable relevancy) on over topic modelling and the establishment of trends.

Overarching questions this project will seek to evaluate:

1. Will the application of topic modelling, an unsupervised learning text mining approach, produce meaningful insights into patterns and trends in the topics raised during government debates?
2. How relevant and meaningful are the identified topics from these models with respect to understanding and finding relevant records? Or do the records found in topic categories lack meaningful context?
3. Are there topic models that work more efficiently with this type of dataset? Does the Hansard record set pose any limitations for the use of topic modelling over other machine learning approaches?
4. Which topic modelling approaches would produce the best results and overall performance for this type of dataset?

Proposed Data Set:

This project will leverage the data set available from the House of Commons' Hansard archives. Documents will be downloaded as unstructured PDFs from the following locations:

- https://www.ourcommons.ca/documentviewer/en/39-1/house/hansard-index
    - Un-indexed 'debate' records are available for the 39th Parliament up to the current 44th Parliament (2006-2024).
    - Debate records are defined as 'the report—transcribed, edited, and corrected—of what is said in the House.'

- Individual PDFs are available for each day the house sat for debate in a given calendar year.

Techniques and Tools:

This project will seek to evaluate the following topic model techniques to determine which is a better fit for the Hansard record set:

- **Hierarchical Dirichlet Processes (HDP):** this process is considered an extension of a well-established technique called Latent Dirichlet Allocation (LDA), with HDP described as being more flexible and dynamic approach to topic modelling. HDP has the added benefit that the total number of topics does not need to be set prior to conducting the unsupervised learning.
- **BERTopic:** A newer technique, this approach leverages the Bidirectional Encoder Representations from Transformers (BERT) model to cluster documents into topics. It is considered to be a highly flexible model for a variety of uses cases and has the ability to processes multilingual documents, which may occur in the Hansard documents.

Some potential tools and Python packages to be used in the analysis include:

- **Gensim**: an open source library for topic modelling and document similarity analysis, which is includes and implementation of HDP.
- **Natural Language Toolkit (NLTK):** a library designed for natural language processing, including tokenization, stemming, and other preprocessing tasks needed prior to running the HDP topic model.

- **BERTopic:** this is the main package for topic model that uses the BERT embeddings.
- **Hugging Face Transformers:** BERTopic uses the Hugging Face Transformers library to work.
- **Umap-learn**: library often used with BERTopic for data visualizations.

REFERENCES

1. Chen, Y., Peng, Z., Kim, S-H., Choi, C.W.. (January, 2023). "What We Can Do and Cannot Do with Topic Modeling: A Systematic Review" in *Communication Methods and Measures*. DOI: 10.1080/19312458.2023.2167965

2. DeepLearning AI. (January 11, 2023). *"Natural Language Processing",* DeepLearning.AI Resources, https://www.deeplearning.ai/resources/natural-language-processing/

3. Benchimol, J., Kazinnik, S., Saadon, Y. (June 15, 2022). "Text Mining Methodologies with R: An Application to Central Bank Text" in *Machine Learning with Applications.* Volume 8.

4. Nath, M. (August 21, 2023). "*Topic Modeling Algorithms",* Medium.com. https://medium.com/@m.nath/topic-modeling-algorithms-b7f97cec6005

5. The, Y.W., Jordan, M.I., Beal, M.J., Blei, D.M. (November 15, 2005). *" Hierarchical Dirichlet Processes",* Berkeley University. https://people.eecs.berkeley.edu/~jordan/papers/hdp.pdf

6. Python Package Index (Jan 2024). "*Find, Install and Publish Python Packages with the Python Package Index",* PyPI.org. https://pypi.org/