CIND820:
Big Data Analytics Project

Final Results: Topic Modeling of the Parliament of Canada Hansard Debate Records (2006-2023)

Student: Colin Lacey
Student ID: 501176114
Supervisor: Professor Ceni Babaoglu

Date of Submission: 01 April 2024

Toronto Metropolitan University

Table of Contents

Abstract

This capstone project aims to explore what information that can be gained through applying topic modeling approaches to the historical Hansard debate records from the Parliament of Canada. The review of past research has highlighted the importance of a using a systematic approach to text mining techniques, the benefits of leveraging a Latent Dirichlet Allocation (LDA) topic model, as well as considerations for evaluation of topic models and the validation of the labels for the extract topic. This capstone project has compared the well-established LDA model against two alterative algorithms, Hierarchical Dirichlet Process (HDP) and the more recent BERTopic. While the literature review asserted that HDP and BERTopic were computationally intensive when compared to LDA, this research conducted as part of this capstone project encountered high computational requirements across all three models due to the size and complexity of the chosen dataset. Additionally, while the scripts for the LDA and HDP topic models could be run on a MacBook Pro with Apple M1 processor (8 CPU cores, 16 GB of RAM, not able to leverage GPU processing), both the sample dataset (385 files) and full dataset (1972 files) would take several hours to complete the building of these models and analysing the coherency values. The BERTopic model required more advanced processing capacity with the ability to leverage GPU processors. As such, leveraging of virtual machines and cloud processing services were required in order to complete the building and analysis of the BERTopic on the sample dataset (385). For BERTopic, Gradient Paperspace was used along with the virtual machine that could operate with the NVIDIA A6000x2 chip, 16 CPU cores, 90 GB of RAM and 48 GB of GPU processing capacity. Ultimately, with the possible adjustments to the training parameters to improve the LDA model, the LDA performance exceeded HDP with the LDA achieving an overall coherence value of 0.63, HDP achieved 0.32. With respect to the interpretability of the identified topics, the LDA model also produced clearly distinct topics, whereas HDP contained many topics with a high degree of similar keywords between topic

categories. As for BERTopic, it was able to identify 7 topics from the representative sample (385 documents), but were comprised almost exclusively of terms associated with parliamentary etiquette (Mr. Speakers, etc.) and were not generally meaningful.

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.

This capstone project will be using the Parliament of Canada Hansard debate records from 2006 until the end of 2023 to explore the approaches to text mining on these unstructured documents, following by exploring the advantages and limitations of topic modelling on the corpus of debate records. The investigation of the use of topic modeling to review the current research available for the latent Dirichlet Allocation (LDA) model, the hierarchical Dirichlet Processes (HDP), and finally the Bidirectional Encoder Representations from Transformers (BERT) model designed for topics, the BERTopic.

Overarching questions this project will seek to evaluate:

1. Given that debate records can cover a wide range of subjects over the course of a day and a large proportion of the language used in debates can best be described as

following parliamentary protocol, will the application of unsupervised topic modelling produce meaningful insights into patterns and trends in government debates?

2. How relevant and meaningful are the identified topics from these models? Are the topics sufficiently precise and non-overlapping enough to allow us to distinguish between closely related topics? Between LDA, HDP and BERTopic, which model has the greatest potential to provide the more meaningful context?

3. What are the general advantages and limitations of the use of the different topic model approaches, such as LDA, HDP and BERTopic?

Literature Review

To better understand how text mining and topic modelling approaches can support the characterization and evaluation of a large body of text, 11 research articles published in reputable journals from 2003 to 2023 were reviewed. Specifically, focusing on articles that covered aspects of text mining and topic modeling approaches to exploratory data analysis, including data cleaning, topic labeling, validation of results, and other qualitative and quantitative statical outputs from these processes. Additionally, this capstone project is interested in reviewing past research on the benefits and limitations of the Latent Dirichlet Allocation (LDA) for modelling textual corpora, which will be the initial model used to summarize and evaluate the selected data, the Parliament of Canada Hansard debate records. Once the LDA model has been trained, the capstone project will be comparing the output and perforce against two alternative topic models algorithms, the Hierarchical Dirichlet Process (HDP) and BERTopic.

Approaching the question of whether text mining and topic modeling could produce meaningful insights from the Parliament of Canada Hansard records, Salloum et al (2018) highlighted some best practices in their paper *Using Text Mining Techniques for Extracting Information from Research Articles* (2018) with respect to extracting data from unstructured and semi-unstructured information sources. That stressed that when moving intentionally through your research from preparation to text refining and eventually information distillation, the successful outcomes of the application of text mining ideally is to detect information in the data source that was not recognized before in the previous unstructured format, and normally would not be possible to achieve for very large datasets in a meaningful way if the work had been done manually (Salloum et al, 2018). The work done by Salloum et al (2018) has greatly informed the research methodology to be used in this capstone project, from the data collection and pre-processing approaches to the application and evaluation of the applicable models generated through the research.

The limitations addressed in the work from Salloum et al (2018) stemmed from the research concentrating primarily on text mining techniques and did not build on the insights obtained from the application of text mining to support topic modelling analysis. Additionally, in the part of their research focused on the data visualizations, the method of leveraging the association rule produced some interesting insights about which terms have strong connections to each other. However, as the research transitioned into exploring the cluster model, the outcomes of evaluating the connections between terms was less impactful and produced some visualizations that did not have evidently clear or impactful insights as to what Salloum et al (2018) were attempting to communicate.

Expanding on the theme of using systemic investigation approaches to extracting quantitative data from unstructured, text heavy data sources, Benchimol et al (2022) undertook similar research on the application of text mining techniques on documents related to central bank

communications. While Benchimol et al (2022) leveraged R programming language to conduct their investigation, the concepts of how they approached the data selection and steps for pre-processing are relevant to the intentions of this capstone project, which will be leveraging Python. The aspect of Benchimol et al (2022) that was of key interest was the exploration of document term frequency and the weighting of term frequency within and across the corpus of records through the use of Term Frequency-Inverse Document Frequency (TF-IDF). As Benchimol et al (2022) explain, TF-IDF offers the benefit of being able to to weight the importance of terms in a document while also comparing how frequently they appear across the selected record set. For example, if a term appears frequently across all documents, it is given a lower weighting than a term that may appear frequently in one or a smaller set of documents, where that term would be considered unique and important, raising the weighting of that term against others.

From the perspective of the dataset selected as part of this capstone project, comparing the outputs of a topic model that either leverages TF-IDF or a “bag of words” corpus dictionary may prove insightful. The Parliament of Canada Hansard debate records capture a lot of text that is repetitive and derived from protocol and parliamentary etiquette and may pose problems in identifying relevant topic keywords. For example, every member of parliament is required to address the Speaker and refer to other members of parliament indirectly through the use of terms such as “honourable member” and the mention of the specific geographic region that the member they are talking about represents. These examples of terms are expected to have a very high frequency within and across all documents and pose a challenge when extracting a relevant set of topics from the record set.

For the purposes of supporting some initial exploratory data analysis on the selected dataset along with generating some descriptive statistics, the intention will be to initially use the latent

Dirichlet allocation (LDA) model to baseline the comparison between the various topic model algorithms. As described by Blei et al (2003), LDA is able to address some of the short comings of using TF-IDF on its own as it is considered a flexible generative probabilistic model for the collection of data. The concept of LDA is based on an assumption of exchangeability of representations of a words and topics, and the LDA model can be scaled up to be leveraged by a significant volume of documents (Blei et al, 2003). This would be an advantage for the exploratory analysis of the dataset to be investigated as part of this capstone project and it contains 1972 files that collectively represent approximately 156,000 pages of text.

As stated by Mohammed & Al-augby (2020), and since Blei et al first published their paper back in 2003, the use of the LDA model has become one of the more popular topic modelling approaches and is now considered the standard to use for unsupervised modelling that has the ability to recognize the latent topic structure inherent to a document's contents. Unfortunately, there is an important factor to take into consideration when deciding to use the LDA model or not, and that consideration is related to the need to identify the optimal number of topics to use when processing a corpus of texts (Mohammed & Al-augby, 2020). The best approach to determine the appropriate number of topics for a specific data set is through the assessment of the values of coherence, where the highest coherence value is generally associated with the optimal number of topics (Mohammed & Al-augby, 2020). Unfortunately, there is no standard optimal number of topics that apply across different data sources. With the research by Mohammad & Al-augby (2020), they were able to confirm through the use of the coherence values that for a data set consisting of 100 eBooks, 20 topics produced the highest coherence value under an LDA model. For this capstone project, initial evaluation of the value of coherence, it appears that 7 topics produces the highest coherence values for 1972 records found within the Parliament of Canada Hansard debate records.

As outlined by Weston et al (2023), the challenge to identifying the correct number of topics for LDA models is inherent to what these latent topics represent – the clustering of co-occurring words in a document. Too little or too many topics may produce results that are not meaningful or representative of the collection of documents being evaluated. There is arguably no set or specific correct number of topics to extract for a given set of records but several options to at least consider (Weston et al, 2023). However, when done correctly, topic modelling provides the advantage of being to describe the broad themes of the corpus that can be scaled to thousands of documents, providing efficiencies and savings of resources when compared to the amount of manual effort that would be required to complete the same scope of work (Weston et al, 2023). While for the purposed of this capstone project, extracted topics were assigned a number value and their associated words visualized in a word cloud, Weston et al (2023) suggests that labelling topics with frequent and exclusive words generated from a given topic may aide the researcher quickly identify which topics may be important to certain research questions as well as those which may be a lower priority for certain analysis. As cited by Eggar and Yu (2022), one of the advantages of using the BERTopic model as compared to the LDA is that BERTopic leverages embeddings so no preprocessing of the body text is required, and the model will automatically find the ideal number of topics. This purported advantage of using BERTopic will be explored in this capstone project, including assessing if the identified topics are meaningful. In addition to meaningful topics, Atagun et al (2021) points out that while BERTopic has the advantage on not requiring preprocessing, the BERTopic requires a significant amount of processing power to complete training of the model.

Adding to the discourse on the challenges of determining the appropriate number of topics to extract from a data set, Greene et al (2014) attempt to address this issue by employing a term stability analysis in order to avoid the pitfalls of extracting a small set of topics that are overly broad or an over-clustering of the data set into many small and similar topics. This is an

important consideration to carry over into the final report for the capstone project with respect to evaluating the performance of the different topic models. Greene et al (2014) defined the stability of a clustering algorithm as the ability to consistently produce comparable results with each iteration. The approach taken by Greene et al (2014) involves initially generating a topic model on the complete data set which will be used as a reference point. They then propose to randomly sample a subset of records for the same number of topics and apply the topic model algorithm and assess the agreement between the sample and reference point. This is repeated for the range of total topic numbers to be tested and based on generated numerical values, the total suggested topic values can be identified. Similar to other points raised by articles reviewed for this capstone project, the stability analysis may reveal that there is more than one potential solution in terms of ideal number of topics and discretion will need to be used in order to decide if a more granular or fine-grained number of topics would be beneficial to the objectives of the research being undertaken (Greene et al, 2014). This stability analysis may prove to be a useful resource to leverage for the final report where the intent will be to evaluate and compare the different topic modelling algorithms of LDA, HDP and BERTopic.

Continuing on the subject of how to evaluate topic models, Wallach et al (2009) note that the unsupervised nature of topic models makes the selection of the applicable model for research difficult. For LDA specifically, this model has traditionally been evaluated based on the performance of outputs from secondary tasks, such as document classification (Wallach et al, 2009). Wallach et al (2009) raise an interesting point about the need to consider the methodology to assess and ultimate select the appropriate topic model for the Parliament of Canada Hansard debate records. The conclusions from Wallach et al (2009) point toward the use of the Chib-style estimator or the “left-to-right” algorithm as the better options to evaluate which is the more appropriate topic model for a data set and may prove useful to consider to be

used by the final report for CIND820 when comparing the model performance of LDA, HDP and BERTopic.

Beyond evaluation, the next logical step is considering validating the topics identified through the chosen model. As Ying et al (2021) stressed through their article *Topics, Concepts, and Measurement: A Crowdsourced Procedure for validating Topics as Measures,* certain applications of topic modelling require validation to ensure that the extracted topics accurately capture the subject matter it is based on, especially when used in political science context. The dataset used by Ying et al (2021) were text-based social media posts of US Senators, which bears many similarities with the Parliament of Canada Hansard debate records that are being analysed in this capstone project and will be further explored in the final report. As such, the emphasis by Ying et al (2021) on the need to consider validation as part of the research methodology for topic modelling of political and social science applications really stood out given the application in these contexts is a text-as-measure rather than the typical information retrieval topic modelling was originally designed to support. While the approach by Ying et al (2021) engaged people to confirm the interpretability of extracted topics, the main question that could be applied following this capstone project is related to the assessment of assigned topic labels and if they are sufficiently precise and non-overlapping to allow us to distinguish between closely related topics.

In the vein of considering the limitations of LDA with respect to identifying an appropriate set of topics along with the need to consider model evaluation and validating the topic labels, the natural next step is to review other possible topic models. In the paper by Teh et al (2005), hierarchical Dirichlet Processes is presented as an extension of the LDA model with notable differences related to allowing for an infinite number of topics. Meaning, a set number of topics does not need to be set in advance of conducting analysis. It also allows for the introduction of a

hierarchical structure for the identified topics which allows for the capturing of more complex relationship in the model (Teh et al, 2005). For this capstone project, while the promise of not needing to evaluate the appropriate number of topics to support the use of LDA from the Hansard dataset and offer potential efficiencies, the computational impact of introducing a hierarchical structure to the extracted topics may increase time required to generate a model. It will be interesting to compare the performance and overall fit of the HDP model against LDA.

In addition to HDP, this capstone project will also investigate the BERTopic model against the LDA model developed so far on the Parliament of Canada Hansard debate records. Grootendorst (2022) provides an excellent overview of what is to be expected when using the BERTopic model over the tried-and-true LDA. Notably, instead of extracting latent topics from a corpus, BERTopic will generate document embeddings based on pre-trained language models and it is these embeddings which are used to cluster into related groupings. The use of the pretrained language models might allow for the capture of more interesting semantic nuances, however for larger datasets the interpretability of the generated topics might be more difficult (Grootendorst, 2022). Similar to HDP, while BERTopic offers advantages over LDA, the need to process the documents through the pretrained language model may be more computationally intensive. As such, prior the running of the script to build and train the BERTopic model, opportunities to optimize the python code to better leverage any available graphics processing units (GPUs) should be considered, including leveraging libraries such as the Compute Unified Device Architectural (CUDA). (Holm et al, 2020).

Overview of Dataset:

This project has evaluated the data set available through the House of Commons' Hansard archives. Documents have been 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 (2006-2023).
- 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.

The Parliament of Canada Hansard debate record dataset from 2006 to 2003 contains:

- 1972 individual PDF files
- 155,385 pages in total
- A mean of 78.9 pages per file
- A median value of 80 pages.
- Total word count of 128,933,818 words across all files.

The preprocessing of the dataset, from removing stop word, numbers and other special characters resulted in the total word count changing from:

- Original word count: 128,933,818 words
- Final word count: 56,563,041 words
- Percent change: -56.13%

The preprocessing of the dataset proved to present its own unique challenges. A large portion of the document text in the Hansard debates relate to Parliamentary procedures and practices that are unique to the Canadian Parliament. All the documents also contain both English and French text. As a result, in addition to leveraging the existing stop words (English and French) from the NTLK package, the set of stop words leveraged in the analysis was augmented through a defined term additional_stopwords that comprised a list of frequently used procedural

language, such as addressing the Speaker of the House during a speech (i.e., Mr. Speaker, Mrs. Speaker). However, the remaining body of text still contains a significant portion of parliamentary etiquette. For example, members of parliament will often refer to other members of parliament by referencing the name of the geographic riding that they represent. Additionally, as the dataset to be processed and evaluated through the various topic modes would be a collection of PDFs containing unstructured data, a data frame was created to provide an overview of the representative information contained in the 1972 file dataset.

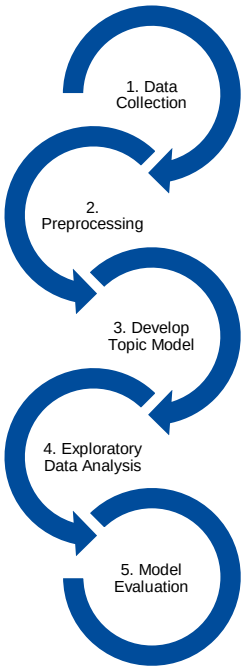
To adequately describe the information to be analyzed in this capstone project, the following data frame was generated with these attributes:

- Document (file name)
- Date of debate in the house of commons (YYYY-MM-DD)
- Parliament (PP) and Session values (SS): PP-SS
- Original Word Count
- Final Word Count
- Topic Coherence Score
- Dominant Topic
- Dominant Topic Value
- Topic Keywords (all values)
- Topic Keywords (individual columns 0-9)
- Preprocessed Tokens

Table: Hansard Debate Dataframe

	Document	Date	Parliament-Session	Original Word Count	Final Word Count	Topic Coherence	Dominant Topic	Dominant Topic Value	Topic Keywords	Topic Keyword 0	Topic Keyword 1	Topic Keyword 2	Topic Keyword 3	Topic Keyword 4	Topic Keyword 5	Topic Keyword 6	Topic Keyword 7	Topic Keyword 8	Topic Keyword 9	Preprocessed Tokens
0	20170130-HAN129-E.pdf	2017-01-30	42-1	94682	42004	0.250252	6	mani	regard, inform, statist, question, tabl, inclu...	regard	inform	statist	question	tabl	includ	mani	return	provid	nation	[common, debat, volum, number, st, session, nd...
1	20200420-HAN034-E.pdf	2020-04-20	43-1	74492	35285	0.268101	0	busi	busi, need, work, health, question, help, make...	busi	need	work	health	question	help	make	mani	covid	order	[rd, parliament, st, session, common, debat, o...
2	20230602-HAN205-E.pdf	2023-06-02	44-1	29410	14065	0.369540	5	question	point, mr, deputi, order, assist, question, ca...	point	mr	deputi	order	assist	question	carol	hugh	parliamentari	common	[th, parliament, st, session, common, debat, o...
3	20120307-HAN091-E.pdf	2012-03-07	41-1	31265	12533	0.286930	1	common	job, common, make, debat, question, want, elec...	job	common	make	debat	question	want	elect	last	countri	english	[common, debat, volum, number, st, session, st...
4	20131126-HAN024-E.pdf	2013-11-26	41-2	75148	30271	0.359525	5	duffi	question, say, offic, parti, know, duffi, ask,...	question	say	offic	parti	know	duffi	ask	wright	get	said	[common, debat, volum, number, nd, session, st...

Research Methodology



The research methodology leveraged for this capstone project involves a systematic approach to the text extraction and distilling information through the following steps:

1. The identification of the data source and the collection of raw data,
2. Pre-processing the information and extracting text,
3. Development of the LDA, HDP and BERTopic models
4. Exploratory Data Analysis including visualization and creation of a data frame with relevant attributes.
5. Model Evaluation

The approach to this capstone project began with setting out the parameters to collecting the scope of PDF files from the Hansard debate records. It was decided that all transcribed debates since the start of the 39th parliament until the end of Dec 2023 (part of the current 44th parliament) would be included in the dataset as this would provide a sufficient scope of recent

government debates that have spanned several election cycles and includes two Prime Ministers since 2006. To ensure consistency, a standardized file name convention was developed to make each individual debate record easily identifiable and retrievable. All file name names were structured as the calendar date plus the number of where this record falls within the series of total debates that occurred in a specific session. For example, the file format (YYYYmmdd-HAN###-E.pdf) would translate as 20230323-HAN172-E.pdf for the House of Commons debate held on March 23, 2023.

Prior to preprocessing the text found across all 1972 pdf files, the script was scaled to create standard samples to allow the testing and development of the final pre-processed text that would support would be used to train the LDA and HDP topic models. The BERTopic model does not require the use of preprocessed texts but accessed the same standardized samples.

A script was developed to randomly select PDFs from the full dataset and to create a copy in a designed folder – one sample set of 25 PDFs, and a second folder for a representative sample of 385 PDFs files.

Using the three dataset variations (25 files, 385 files and the full dataset of 1972 files), the datasets were preprocessed in order to extract the text from all files, then to tokenize the body of text into sentences and then into individual words. Typically, a standard set of stop words are removed from the final preprocessed text. However, the Hansard records posed interesting challenges of French words making it into the topic keywords (e.g., de, au, etc.) along with many keywords associated with Parliamentary etiquette, such as address the Speaker of the House and referring to other members of parliament by the geographic region that they represent. As such, the standard set of stop words was augmented to remove French stop words as well as a customized list of additional stop words unique to the Hansard records.

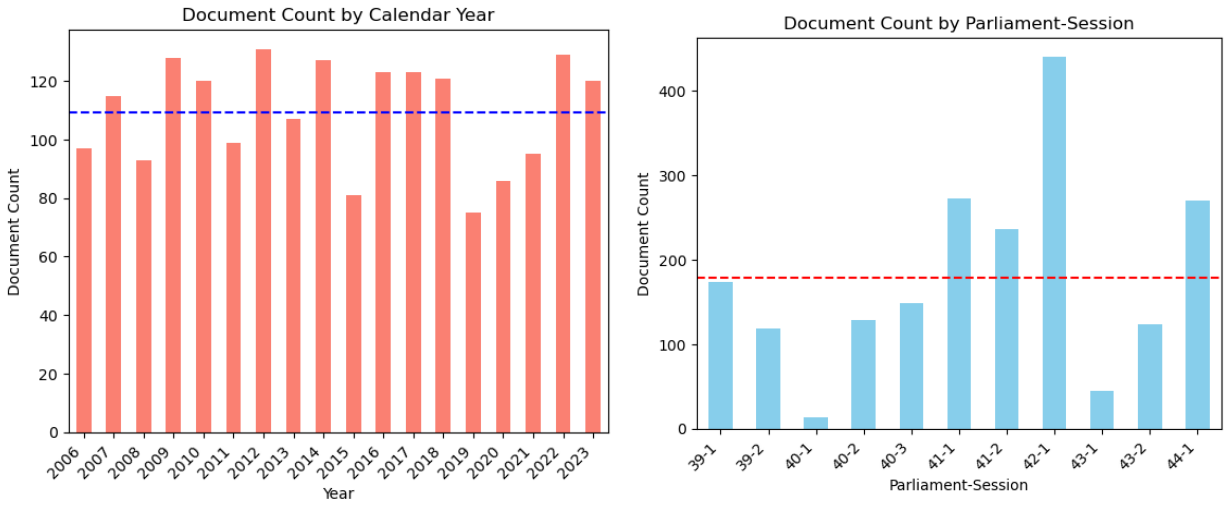
Finally, prior to developing the LDA and HDP topic models, the corpus of text was split into a testing and training set to support cross-validation, with an 80%/20% split between training and testing. The exploratory data analysis was conducted to better understand the scope of files within the full dataset and to support the preprocessing of the text to be used in the training of the models.

Three different topic models were developed and analysed in this capstone project, which include the Latent Dirichlet Allocation (LDA), Hierarchical Dirichlet Process (HDP) and the BERTopic (topic model for the bidirectional encoder representations from transformers, or BERT). Models were evaluated against the individual coherence values for each identified topic, the overall coherence value for the model, the efficiency of the model which included time and processing capacity required, and finally a review the of interpretability of the topics and their associated keywords. Part of model evaluations includes considerations for limitations and other obstacles when developing and evaluating the performance of the LDA, HDP and BERTopic models.

Exploratory Data Analysis

The exploratory data analysis of the Hansard debate records includes assessment of the distribution of files across calendar year as well as parliamentary session. The total documents per calendar year ranged from 79-120 files, with a mean of 109.4 files per calendar year. Looking at document distribution across parliamentary sessions, the distribution was highly variable, with a range of xx-yy files, with a mean of 179.1 files per parliamentary session.

Figure: Overview of total documents by calendar year and by Parliament-Session



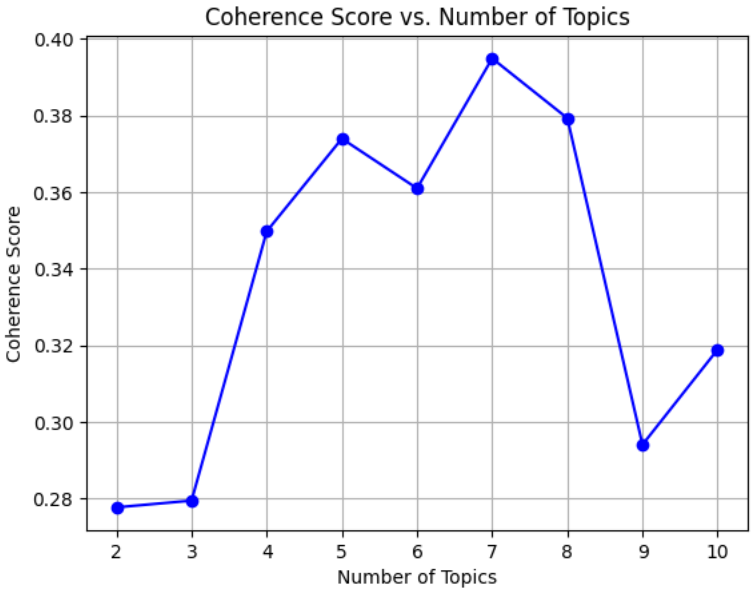
The analysis of the preprocessed text prior to running it through the LDA and HDP models (BERTopic does not require preprocessed text), some of the dominant keywords that make up the generated word cloud that stood out and were easily identifiable:

- **Regions:** Ontario, Québec, British Columbia, Alberta
- **Economy & Trade Issues:** forestry industry, wood, industry, job, product
- **Law Enforcement:** human trafficking, prostitution, women, criminal, victim
- **International:** Iraq

Figure: Word Cloud of Preprocessed Text

In order to support the building of the LDA model, cross validation of the number of total topics against the coherence scores was completed, using a range of 2-10 topic models in the dataset. As this was a heavy computing process, the cross-validation was conducted using a representative sample of 385 files out of the 1972 total. For the LDA model, the ideal number of topics with a coherence value of 0.39 was 7 topics and evaluating the sample set of 385 documents and assessing all coherence values between 2 and 10 topics.

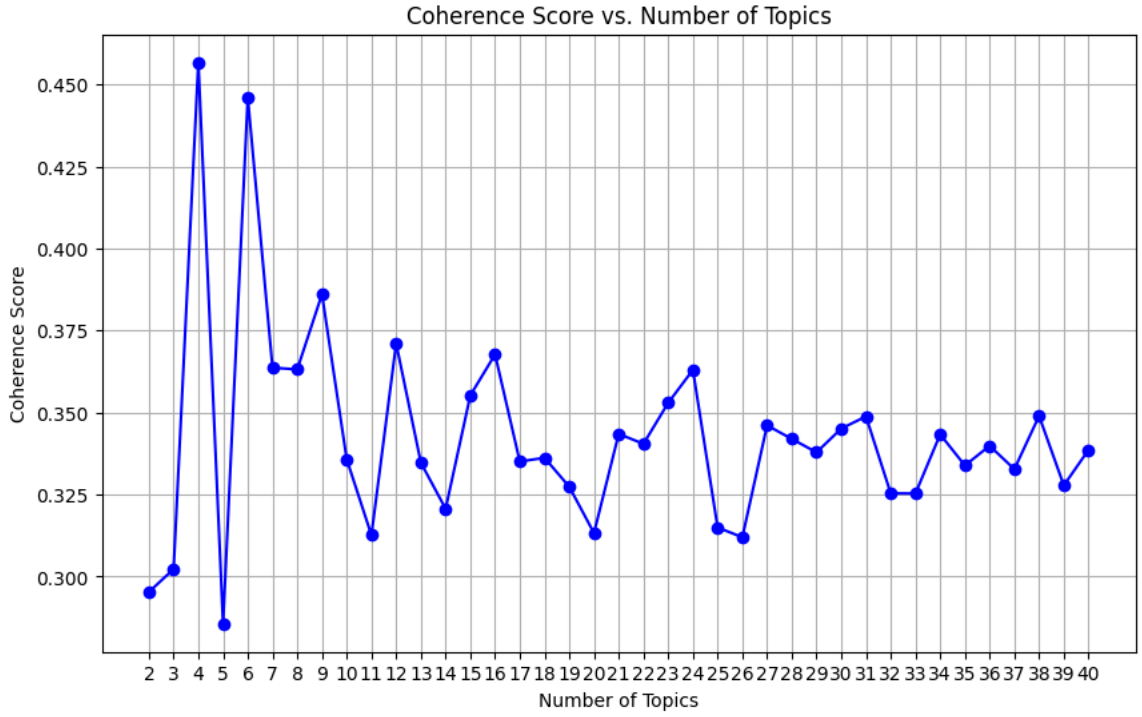
Figure & Table: LDA Coherence Score vs. Number of Topics (s=385)



Number of Topics	Coherence Score
2	0.2777207278487633
3	0.27948628535283787
4	0.3497679850116785
5	0.37398161141900993
6	0.36098889812859786
7	**0.3949243102093841**

8	0.3793170579177607
9	0.29408767522024126
10	0.3189270003203882

Figure: LDA Coherence Score vs Number of Topics (Assessment of 2-40 topics)



Topic Model Definitions

1. **Latent Dirichlet Allocation (LDA)** – LDA is an unsupervised and generative probabilistic approach for topic modelling. LDA is a three-level hierarchical Bayesian model where each identified topic is modelling on the underlying set of topic probabilities. In the context of text modeling, the topic probabilities provide an explicit representation of a document (Blei et al, 2003).

2. **Hierarchical Dirichlet Process (HDP)** – Is also a nonparametric Bayesian model for unsupervised machine learning. Unlike LDA, the HDP model does not need to have the maximum number of topics to be defined in advanced but is instead derived from the dataset (Teh et al, 2005).

3. **BERTopic** – Is a topic modelling algorithm that leverages transformers and term frequency-inverse document frequency (TF-IDF) to derive easily identifiable topics from a body of text. Similar to HDP, the maximum number of topics does not need to be defined in advance. Also, an advantage BERTopic has over LDA and HDP is that BERTopic is able to leverage embeddings, so no preprocessing of text is required (Egger and Yu, 2022).

Research Results

The first topic model that was built on the Hansard debate records was the LDA model. The first iteration of the LDA model was developed as part of the original literature review where a sample of 200 documents was used to train the model. However, all default settings were used when running this first iteration, including limited that number of passed the LDA model ran over the body of text. Additionally, the sample dataset was not split into a training and test sets.

As such, the results of the first iteration of the LDA model indicated that there was a relatively low coherence values for individual as well as the overall coherence value for the model of 0.321. The second iteration and third iteration improved upon the original LDA model by first creating a training and testing set (80%/20% split), as well as setting the number of passes the LDA would train over the body of text to 10 passes. As such, the overall coherence value for the LDA model increased – the second iteration (representative sample size of 385 documents)

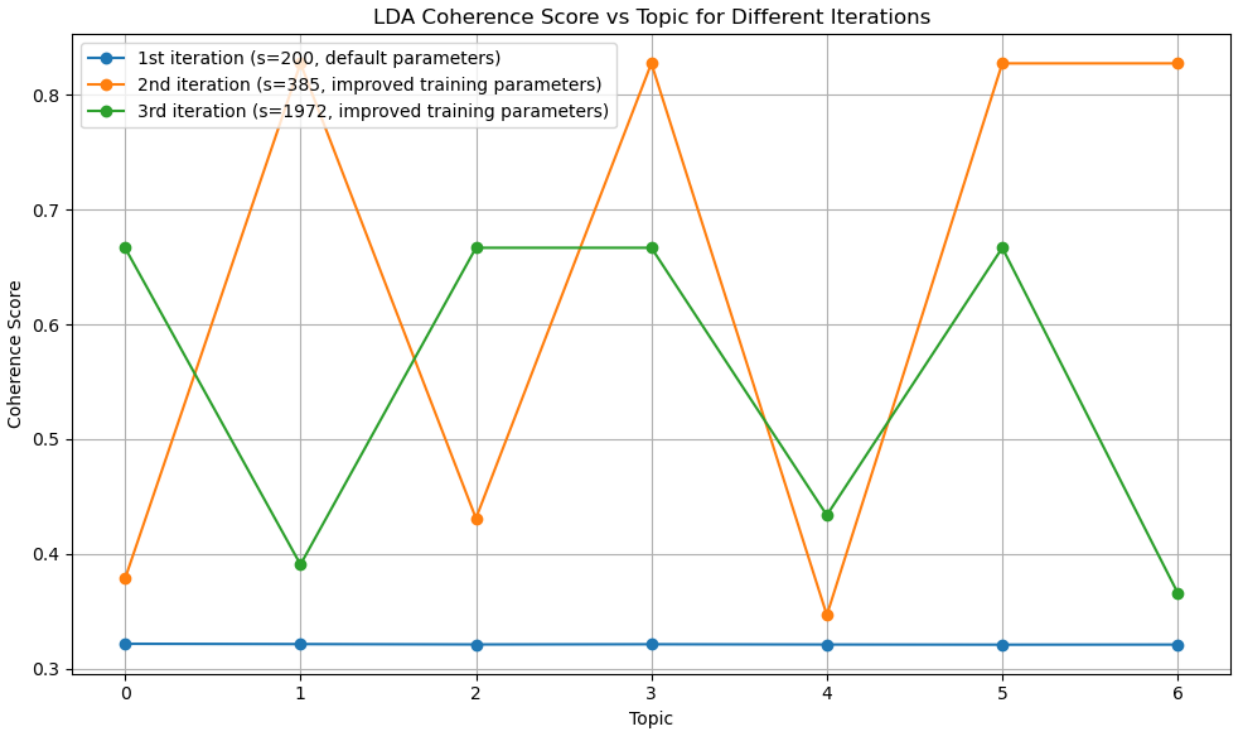
achieved a coherence value of 0.701, and the third iteration (used full dataset of 1972 documents) achieved a coherence value of 0.628.

Table: LDA Overall Coherence Values

Model	1st Iteration (s=200, default parameters)	2nd Iteration (s=385, improved training parameters)	3rd Iteration (s=1972, improved training parameters)
LDA	0.3209934375	0.7010199998587691	0.628447487531789

Table: LDA Individual Topics Coherence Values (7 Topics)

Topic	1st iteration (s=200, default parameters)	2nd iteration (s=385, improved training parameters)	3rd iteration (s=1972, improved training parameters)
0	0.321582	0.378322483	0.66674105
1	0.321323	0.827499487	0.39064673
2	0.321006	0.430924069	0.66674105
3	0.321173	0.827499487	0.66674105
4	0.320939	0.346593962	0.4336867
5	0.320810	0.827499487	0.66674105
6	0.320885	0.827499487	0.36582895



The building and training of the HDP model leveraged the same preprocessed text as well as the same training and test corpus that was used to run the LDA model. While the LDA model required that the maximum number of topics be define, the HDP model will interpret the total number of topics based on the body of text. Between the first iteration (used representative sample of 385 documents) and the second iteration (used the full dataset of 1972 documents), the HDP training parameters were improved to set a max chunk value that would see the HDP model pass over the body of text three times. However, this additional training of the HDP model did not result in any meaningful change in the overall coherence values (0.315 for the first iteration, 0.319 for the second iteration).

Table: HDP Overall Coherence Values

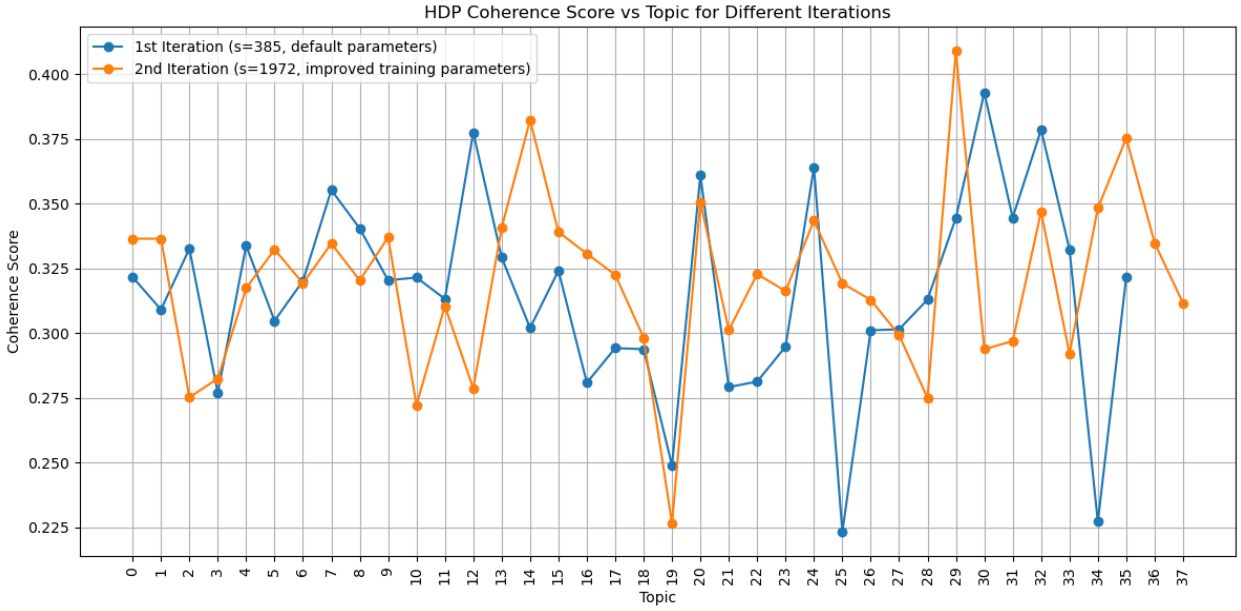
Model	1st Iteration (s=385, default parameters, 35 topics)	2nd Iteration (s=1972, improved training parameters, 38 topics)
HDP	0.31535182848347443	0.3194660065275221

Table: HDP Individual Topics Coherence Values

Topic	1st iteration (s=385, default parameters)	2nd Iteration (s=1972, improved training parameters)
0	0.32158900	0.33640348
1	0.30904022	0.33640348

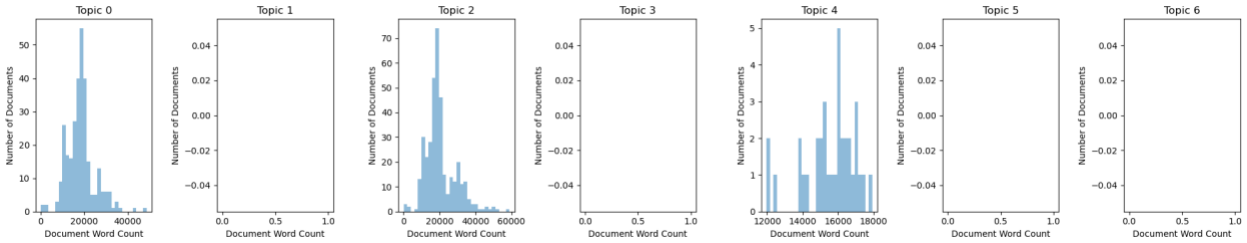
2	0.33253991	0.27515214
3	0.27681661	0.28234077
4	0.33370110	0.31732046
5	0.30470254	0.33226346
6	0.32009432	0.31918224
7	0.35522892	0.33455076
8	0.34036993	0.32032993
9	0.32034357	0.33691783
10	0.32141587	0.27214106
11	0.31315342	0.31036996
12	0.37742434	0.27861806
13	0.32919183	0.34067572
14	0.30194155	0.38221844
15	0.32406232	0.33902581
16	0.28075245	0.33064272
17	0.29416703	0.32245783
18	0.29371789	0.29819393
19	0.24887322	0.2263735
20	0.36090615	0.35053454
21	0.27910223	0.30109477
22	0.28118665	0.32276842
23	0.29476552	0.31615488
24	0.36405859	0.34339287
25	0.22312023	0.3191801
26	0.30104352	0.31279899
27	0.30140142	0.29914988

28	0.31289342	0.27487793
29	0.34451057	0.4089477
30	0.39280360	0.29375181
31	0.34455985	0.29683283
32	0.37850265	0.34686377
33	0.33229397	0.29199312
34	0.22703961	0.34823557
35	0.32158900	0.37526959
36	~	0.33474749
37	~	0.31153237



LDA Representative Text & Word Count by Total Number of Documents

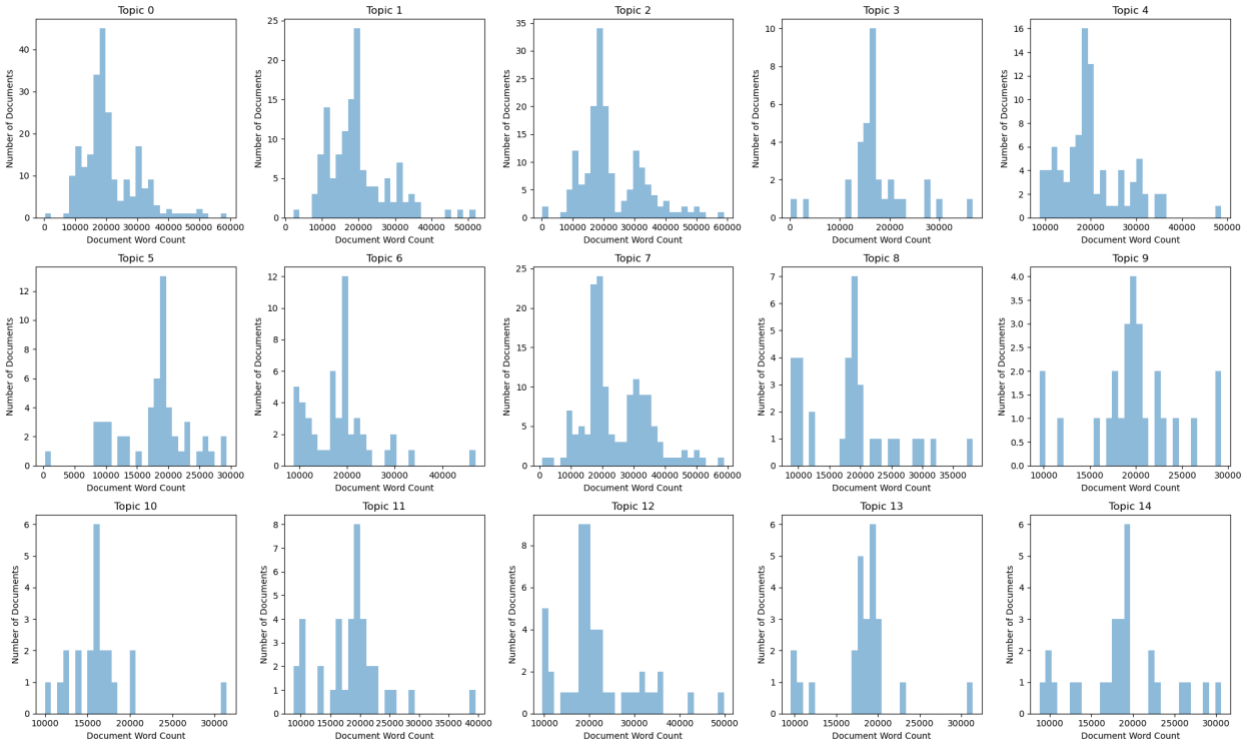
For the representative sample of 385 documents, when assessing the document word count against the number of documents for each identified topic, the graphs produced were consistent with the data frame generated of the representative text from the corpus, with Topics 0, 2 and 4 standing out from the 7 identified topics.



	Topic_Num	Topic_Perc_Contrib	Keywords	Representative Text
0	0	0.9774	commonsdeb, english, translat, order, work, nation, new, countri, like, howev	[common, debat, volum, number, session, parliament, offici, report, hansard, wednesday, march, h...
1	2	1.0000	need, work, want, support, make, like, get, know, common, question	[parliament, session, common, debat, offici, report, hansard, volum, thursday, septemb, honour, ...
2	4	0.4743	ontario, qubec, britishcolumbia, tom, theminist, deanallison, davidsweet, leonbenoit, chriswarke...	[common, debat, volum, number, session, parliament, offici, report, hansard, friday, decemb, hon...

HDP Representative Text & Word Count by Total Number of Documents

For the representative same of 385 documents, the results for the HDP model were different from LDP in that nearly all of the 38 identified topics contained representative text and it was difficult to distinguish between the keywords contained in each topic due to a higher degree of similarity.

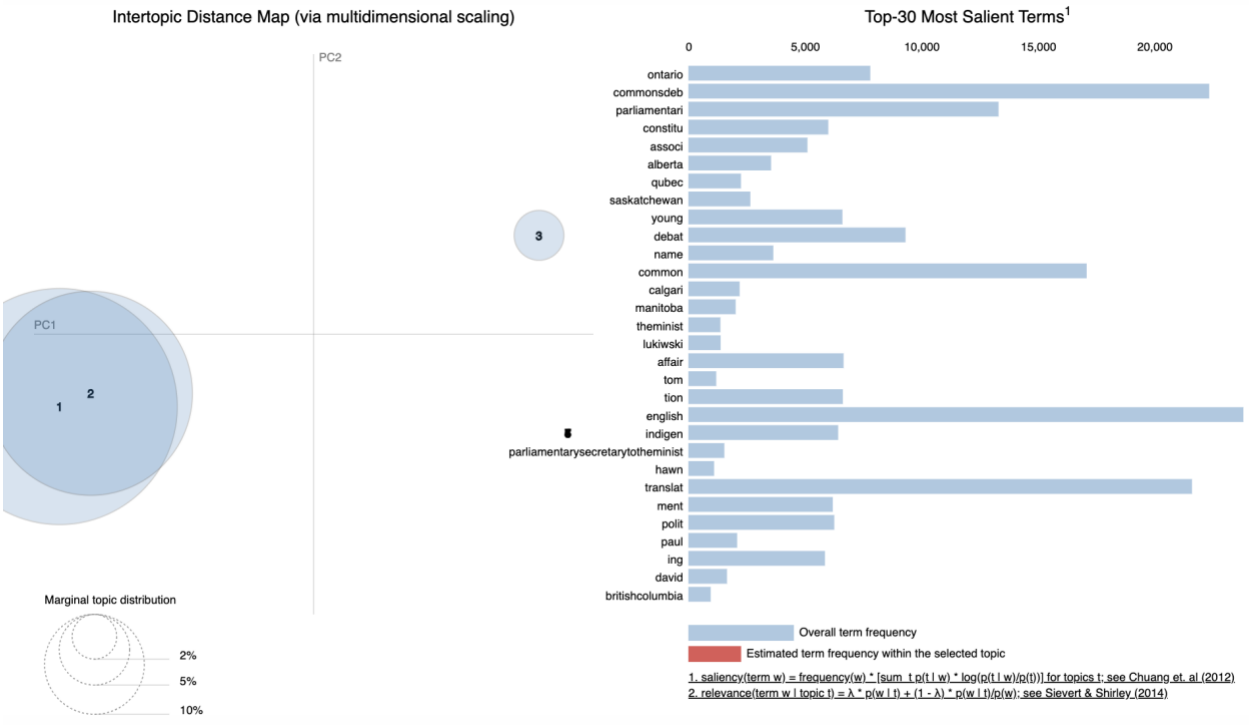


	Topic_Num	Topic_Perc_Contrib	Keywords	Representative Text
0	0	1.0000	work, need, support, make, like, want, know, order, right, commun, question, import, countri, ge...	[common, debat, volum, number, session, parliament, offici, report, hansard, tuesday, april, hon...
1	1	1.0000	work, order, right, want, support, need, like, make, question, commonsdeb, english, nation, know...	[common, debat, volum, number, session, parliament, offici, report, hansard, monday, june, honou...
2	2	0.9999	work, need, want, support, tax, make, like, countri, know, say, get, question, right, health, us...	[common, debat, volum, number, session, parliament, offici, report, hansard, thursday, may, hono...
3	3	0.9999	work, ontario, support, need, nation, english, want, countri, translat, make, commonsdeb, qubec,...	[common, debat, volum, number, session, parliament, offici, report, hansard, wednesday, may, hon...
4	4	1.0000	right, work, support, want, need, like, import, make, question, countri, english, order, commun,...	[common, debat, volum, number, session, parliament, offici, report, hansard, monday, june, honou...
5	5	0.9999	work, commonsdeb, countri, english, support, make, need, like, translat, import, want, order, ne...	[common, debat, volum, number, session, parliament, offici, report, hansard, tuesday, april, hon...
6	6	1.0000	work, order, right, make, need, support, want, import, like, know, commun, commonsdeb, english, ...	[common, debat, volum, number, session, parliament, offici, report, hansard, thursday, june, hon...
7	7	1.0000	work, need, want, right, make, care, support, like, countri, know, mani, say, get, common, impor...	[parliament, session, common, debat, offici, report, hansard, volum, thursday, novemb, honour, a...
8	8	1.0000	work, right, question, want, need, make, like, import, know, commonsdeb, countri, quebec, suppor...	[common, debat, volum, number, session, parliament, offici, report, hansard, tuesday, june, hono...
9	9	0.9999	need, work, support, make, right, countri, want, like, tax, know, mani, commonsdeb, new, english...	[common, debat, volum, number, session, parliament, offici, report, hansard, tuesday, januari, h...

LDA & HDP Topics: Intertopic Distance Map & Top 30 Terms

In terms of assessing whether the models could produce distinct topics that are meaningful, there is distinct difference in the performance of the LDA and HDP models when reviewing their Intertopic Distance Maps (IDMs) and top-30 most salient keywords. When reviewing the graphs below, it becomes clear that while LDA was able to produce a smaller number of distinct topics with meaningful keywords, the HDP model was able to find a relevant keywords across a greater population of documents in the dataset.

LDA Topics:



HDP Topics:

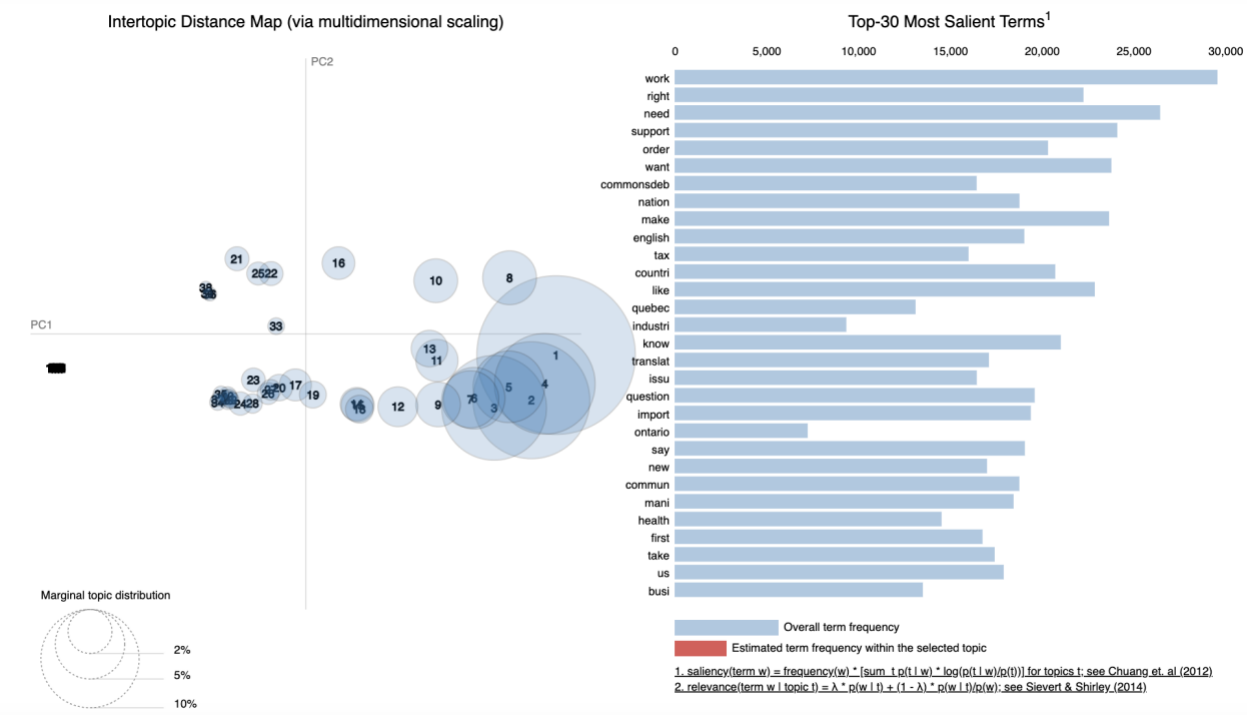


Table: Most Salient Terms by Model (Top 30)

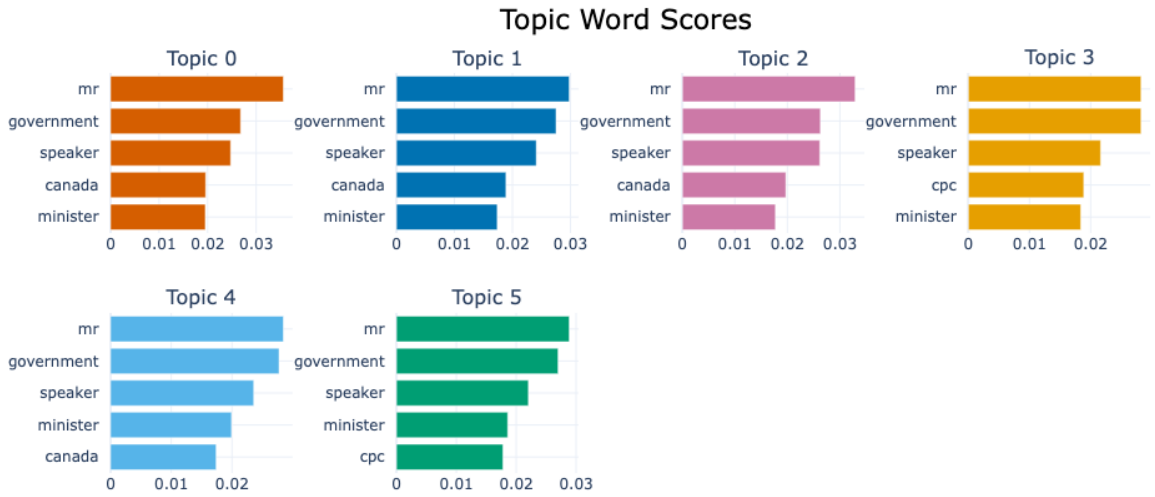
Topic Model	Top 30 Keywords
LDA	ontario, commonsdeb, parliamentari, constitu, associ, alberta, qubec, saskatchewan, young, debat, name, common, calgari, manitoba, theminist, lukiwski, affair, tom, tion, english, indigen, parliamentarysecretarytotheminist, hawn, translat, ment, polit, paul, ing, david, britishcolumbia
HDP	work, right, need, support, order, want, commonsdeb, nation, make, english, tax, country, like, quebec, industri, know, translat, issu, question, import, ontario, say, new, commun, mani, health, first, take, us, busi

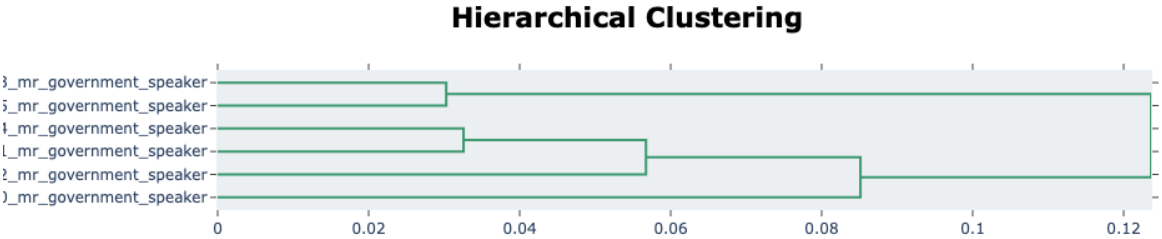
BERTopic Model Outcomes:

After navigating the obstacle of running the BERTopic model on a machine with the Apple M1 processor (8 cores, 16 GB of RAM and no accessible GPUs), there was limited success in training the BERTopic model on the Hansard debate records. Leveraging the guidance and research completed for configuring the BERTopic Model from Grootendorst (2022) and the associate information provided in Grootendorst's GitHub repository for BERTopic parameters, the model was able to identify 7 topics within the dataset. However, all of the identified topics were characterized by key terms that can are best described as being related to parliamentary etiquette.

While it is an accurate representation of the documents analyzed in that a significant proportion of the transcribed conversations in debate records are related to the protocols of how members of parliament are support to engage with one another. However, to better understand the issues raised and discussed in these debates, these parliamentary etiquette terms will need to be removed from the text. While LDA and HDP were able to leverage customized stop words during the preprocessing, a similar ability will need to be achieved for BERTopic. This runs contrary to one of the promoted advantages of using BERTopic as it typically does not require preprocessing of texts.

So while the BERTopic was successfully trained in principal, until further refinements of the source text and modification of model parameters is conducted, the outputs of the current BERTopic are not very meaningful.





Model Efficiency & Research Limitations:

In general, the experience of building and training each of the models (LDA, HDP and

BERTopic) was found to be resource intensive. While there were some differences in the

training of each model type, the amount of effort required to process supporting and dependant code, such as preprocessing the and calculating the ideal number of topics for the LDA, must be taken into consideration.

While the LDA model was relatively easier to train, it did require two processor heavy steps prior to training the model – defining the maximum number of topics as well as preprocessing the text. Compared to HDP which does not need maximum topics defined upfront, it still required the preprocessing of the text and did require more run time to train.

Finally, with respect to the BERTopic model performance, the advantage of not requiring the definition of number of topics in advance or preprocessing text is off set by the high computing processor resources required to even run the model. For example, the BERTopic could not be run locally on the same MacBook Pro used to successfully run the LDA and HDP models. As such, a virtual machine and cloud computing resource that had access to machines with higher CPU cores and GPU capacity was leveraged to train and compare the BERTopic model.

Computing power	**Preprocessing Text**	**Train LDA**	**Train HDP**	**Train BERTopic**
Apple M1 (8 CPU cores, 16 GB RAM, no GPU)	4 hrs, 17 mins, 8s	31 mins, 32s	35 mins, 29s	*Kernel Failed*
NVIDIA A6000x2 (16 CPU cores, 90 GB RAM, 40 GB GPU)	6hrs, 23mins, 35s	1hr, 3min, 24s	15mins, 10s	Gather text: 1hr, 12mins, 36s Fit model: 1min, 32s

In addition to preprocessing text and training the models, there was significant processing time required to evaluate the coherence values for each individual topic as well as the overall coherence value for the model in general. In terms of performance, there was greater

efficiencies in evaluating the coherence values for the LDA model, and more difficult to evaluate coherence for the HDP model.

Computing power	Evaluating LDA topic coherence values (7 topics)	Evaluating LDA overall coherence	Evaluating HDP topic coherence values (38 topics)	Evaluating HDP overall coherence
Apple M1 (8 CPU cores, 16 GB RAM, no GPU)	7 mins, 30s	1 min, 35s	54 mins, 8s	1 hr, 2mins, 48s
NVIDIA A6000x2 (16 CPU cores, 90 GB RAM, 40 GB GPU)	9 mins, 13s	1 min, 43s	1hr, 6mins, 30s	1 hr, 12min, 33s

Additionally, the computing processor constraints put limitations on the extent of analysis conducted. For example, the LDA model could have been further enhanced through the inclusion of bigrams to the overall corpus by addressing some of the limitations as discussed by Park et al (2015) regarding the “bag of words” concept that underpins the LDA model.

Table: Topic Keywords vs N-Grams

	Document	Dominant Topic	Topic Keywords	Unigrams	Bigrams	Trigrams
0	20170130-HAN129-E.pdf	6	regard, inform, statist, question, tabl, inclu...	aa, aandc, aandcinac, aban, abandon, abdic, ab...	aa amount, aandc indigenousand, aandcinac iden...	aa amount iap, aandc indigenousand northern, a...
1	20200420-HAN034-E.pdf	0	busi, need, work, health, question, help, make...	aaron, aarontuck, abandon, abandonedw, abandon...	aaron tuck, aarontuck greg, abandon parliament...	aaron tuck jolen, aarontuck greg jami, abandon...
2	20230602-HAN205-E.pdf	5	point, mr, deputi, order, assist, question, ca...	abil, abilityof, abit, abl, aboard, aboultaif,...	abil better, abil extern, abil feed, abil fina...	abil better review, abil extern depth, abil fe...
3	20120307-HAN091-E.pdf	1	job, common, make, debat, question, want, elec...	abandon, abdic, abil, abitibitmiscamingu, abl,...	abandon inshor, abandon veteran, abdic democra...	abandon inshor fisheri, abandon veteran first,...
4	20131126-HAN024-E.pdf	5	question, say, offic, parti, know, duffi, ask,...	aballot, abandon, abdic, abeauti, abet, abett,...	aballot sacrifici, abandon mental, abdic respo...	aballot sacrifici lamb, abandon mental health,...

Some early analysis regarding the topic keywords generated through the LDA were compared to n-grams (unigrams, bigrams, and trigrams) in the dataset. The n-grams results underscored the repetitive term usage across all files, particularly related to the geographic reference that members represent. However, the n-grams did reveal some keywords that did not previously

appears in the later analysis, such as “abandon mental health” and “abandon veteran” which warrants further investigation.

Conclusions

The outcomes of the topic modelling in this capstone project demonstrate that meaningful insights can be obtained from running a topic model on a large corpus of text. When it comes to performance of each algorithm tested, the LDA model produced a higher overall coherence value with a smaller number of distinct topics that were comprised of meaningful keywords. The general performance of the model was positively impacted by increased the number of passed the model would run over the corpus, raising it from the default single pass up to 10 passes.

The HDP model performed with a lower overall coherence value, and it identified a much higher number of total topics (38) that contained higher degree of keywords that overlapped with more than one topic. This higher number of topics with overlapping content made it difficult to distinguish between each identified topic. However, the HDP model did find a greater occurrence of the relevant keywords across a higher volume of documents within the dataset.

The BERTopic model presented its own set of challenges, requiring a machine with access to more advanced processing power such as a machine with access to GPUs. In order to train and test the BERTopic, a virtual machine with cloud computer processing services was identified as an alternative to running the scripts on the local terminal. Unfortunately, the topics identified through the BERTopic model were not meaningful due to the inherent characteristics of the dataset analyzed is comprised of a high degree of jargon and terms associated with protocol and parliamentary etiquette. Additionally, while the BERTopic could be run on the selected virtual machine, there was inconsistencies with how the LDA and HDP script ran in this environment (some aspects run fast, others such as the preprocessing of the text took longer).

Future use of a virtual machine to compare the different models will need to factor in potential optimization for cloud computing.

Recommendations:

Due to the limitations encountered with training the BERTopic model, there are still possibilities to continue comparing the advantages offered by this algorithm. Additionally, the HDP model performance would continue to be improved, such as increasing the max-chunk size so that the number of passes over the corpus is comparable to the 10 passes that the LDA model uses for training. While the LDA model generally showed the higher overall coherence values, further improvements could be achieved through the inclusion of bigrams and trigrams into the corpus, as well as other minor revisions to the “additional_stopwords” list. As for next steps, any of these topic model may benefit from being combined with the development of an Elasticsearch so that once topics and keywords have been identified, specific documents can be identified and retrieved after performing full-text keyword searches on the dataset. Finally, given LDAs proven model and flexibility of being able to run this topic model on a variety of currently available processing hardware, the initial assessment of any body of text may benefit from baselining outputs from LDA model before moving on to a more advanced topic model algorithms.

Github Repository

https://github.com/CDL-DataSci/CIND820

References

1. Atagun, E., Hartoka, B., Albayrak, A. (2021). “Topic Modeling Using LDA and BERT Techniques: Teknofest Example” from the 6th International Conference on Computer Science and Engineering. https://www.researchgate.net/publication/355226397

2. 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.

3. Blei, D.M., Ng, A.Y., Jordan, M.I. (2003) “Latent Dirichlet Allocation" in the Journal of Machine Learning Research 3.

4. 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

5. Eggar, R., Yu, J. (May 2022). “A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts” in Frontiers in Sociology. https://doi.org/10.3389/fsoc.2022.886498

6. Greene, D., O’Callaghan, D., Cunningham, P. (2014). “How Many Topics? Stability Analysis for Topic Models” from the Joint European Conference on Machine Learning and Knowledge Discovery in Databases. https://doi.org/10.1007/978-3-662-44848-9_32

7. Grootendorst, M. (March 2022). “BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure”, Cornell University. https://arxiv.org/abs/2203.05794 and associated git repo: https://maartengr.github.io/BERTopic/index.html

8. Holm, H. H., Brodtkorb, A. R., Saetra, M. L. (January 2020). “GPU Computing with Python: Performance, Energy, Efficiency and Usability” in Computation 2020, vol 8, no. 4. DOI: 10.3390/computation8010004

9. Mohammed, S.H., Al-augby, S. (June 2020) “LSA & LDA Topic Modeling Classification: Comparison Study on e-Books” in Indonesian Journal of Electrical Engineering and Computer Science, vol 19., no. 1. DOI: 10.11591/ijeecs.v19.i1.pp353-362

10. Park, Y., Alam, H., Ryu, W.-J., Lee, S.K. (2015). “BL-LDA: Bringing Bigram to Supervised Topic Model” from the 2015 International Conference on Computational Science and Computational Intelligence. https://american-cse.org/csci2015/data/9795a083.pdf

11. Salloum, S.A., Shaalan, K., Al-Emran, M. (January 2018) “Using Text Mining Techniques for Extracting Information from Research Articles” in Studies in computational Intelligence

12. Teh, 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

13. Wallach, H. M., Murray, I., Salakhutdinov, R., Mimno, D. (June 2009). “Evaluation Methods for Topic Models” in the Proceedings of the 26th Annual International Conference on Machine Learning. https://dl.acm.org/doi/10.1145/1553374.1553515

14. Weston, S.J., Shyrock, I., Light, R., Fisher, P.A., (April-June 2023) “Selecting the Number and Labels of Topics in Topic Modeling” in Advances in Methods and Practices in Psychological Science vol. 6, no. 2.

15. Ying, L., Montgomery, J.M., Stewart, B.M., (June 2021) “Topics, Concepts and Measurement: A Crowdsourced Procedure for Validating Topics as Measures” in Political Analysis. doi:10.1017/pan.2021.33

16. Python Package Index (Jan 2024). “Find, Install and Publish Python Packages with the Python Package Index”, PyPI.org. https://pypi.org/