CIND820:
Big Data Analytics Project

Literature Review to Support Topic Modelling of the Parliament of Canada Hansard Debate Records (2006-2023)

Toronto Metropolitan University

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

Date of Submission: 22 February 2024

Table of Contents

Abstract

This literature review aims to explore what information can be gained through applying text mining and 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 an LDA topic model, as well as considerations for evaluation of topic models and the validation of the labels for the extract topic. The literature review also helped to identify comparable alternatives to LDA that include HDP and the more recent BERTopic. While the literature review has identified possible advantages for these alternative topic models, there is a potential for these alternatives to be more computationally intensive. Initial exploratory data analysis was conducted on the chosen dataset using an LDA topic model, where the challenge of a high occurrence of the same frequently used terms associated with parliamentary protocol, was used across all debate records.

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 literature review 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? Which model has the potential to perform the best and produces the more meaningful groupings with respect to the

Hansard debate records?

Literature Review

To better understand how text mining and topic modelling approaches can support the characterization and evaluation of a large body of text, I have reviewed 11 research articles published in reputable journals from 2003 to 2023. 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, the literature review attempts to explore the benefits and limitationa of the Latent Dirichlet Allocation (LDA) for modelling textual corpora, which will be the initial model used to summarize and evaluate the selected data source for the CIND820 capstone project, the Parliament of Canada Hansard debate records. Rounding out the literature review will be discussion on two alternative topic models algorithms, the Hierarchical Dirichlet Process (HDP) and BERTopic, which will be investigated further as part of the final report and compared against the performance of the LDA model.

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 literature review and eventually the final 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 literature review 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 corupus 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 literature review and eventual final 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 this literature review and conducting 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. 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 initial exploratory analysis of the dataset to be investigated as part of this literature review 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 literature review, 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 literature review, 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.

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 from this literature review 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 that are part of this literature review, 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 after

the literature review 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 literature review 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 literature review 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). In this literature review, 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 hierarchial 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 literature review is also considering investigating 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 process the documents through the pretrained language model may be more computational intensive.

Project Data Source:

This project will be evaluating 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 (2006-2023).
- 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.

Exploratory Data Analysis

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%

Word Cloud: Preprocessed Text

For the LDA model, the ideal number of topics with a coherence value of 0.31 was 7 topics. Through the processing of the data and the creation of LDA model, the following data frame was generated with the following attributes:

- Document (file name)
- Date
- Parliament and Session values
- Original Word Count
- Final Word Count
- Topic Coherence
- Dominant Topic

- Dominant Topic Value
- Topic Keywords (all values)
- Topic Keywords (individual columns 0-9)
- Preprocessed Tokens

Table: Hansard Debate Dataframe

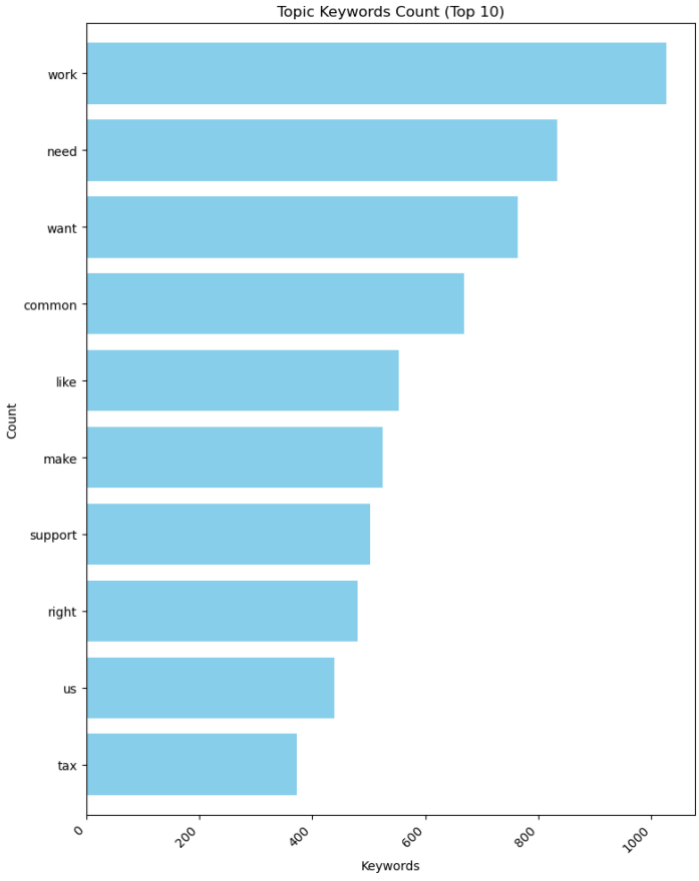
	Document	Date	Parliament-Session	Original Word Count	Final Word Count	Topic Coherence	Dominant Topic	Dominant Topic Value	Topic Keywords	Topic Keyword 0	Topic Keyword 1	Topic Keyword 2	Topic Keyword 3	Topic Keyword 4	Topic Keyword 5	Topic Keyword 6	Topic Keyword 7	Topic Keyword 8	Topic Keyword 9	Preprocessed 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...

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.

Table: Topic Keywords vs N-Grams

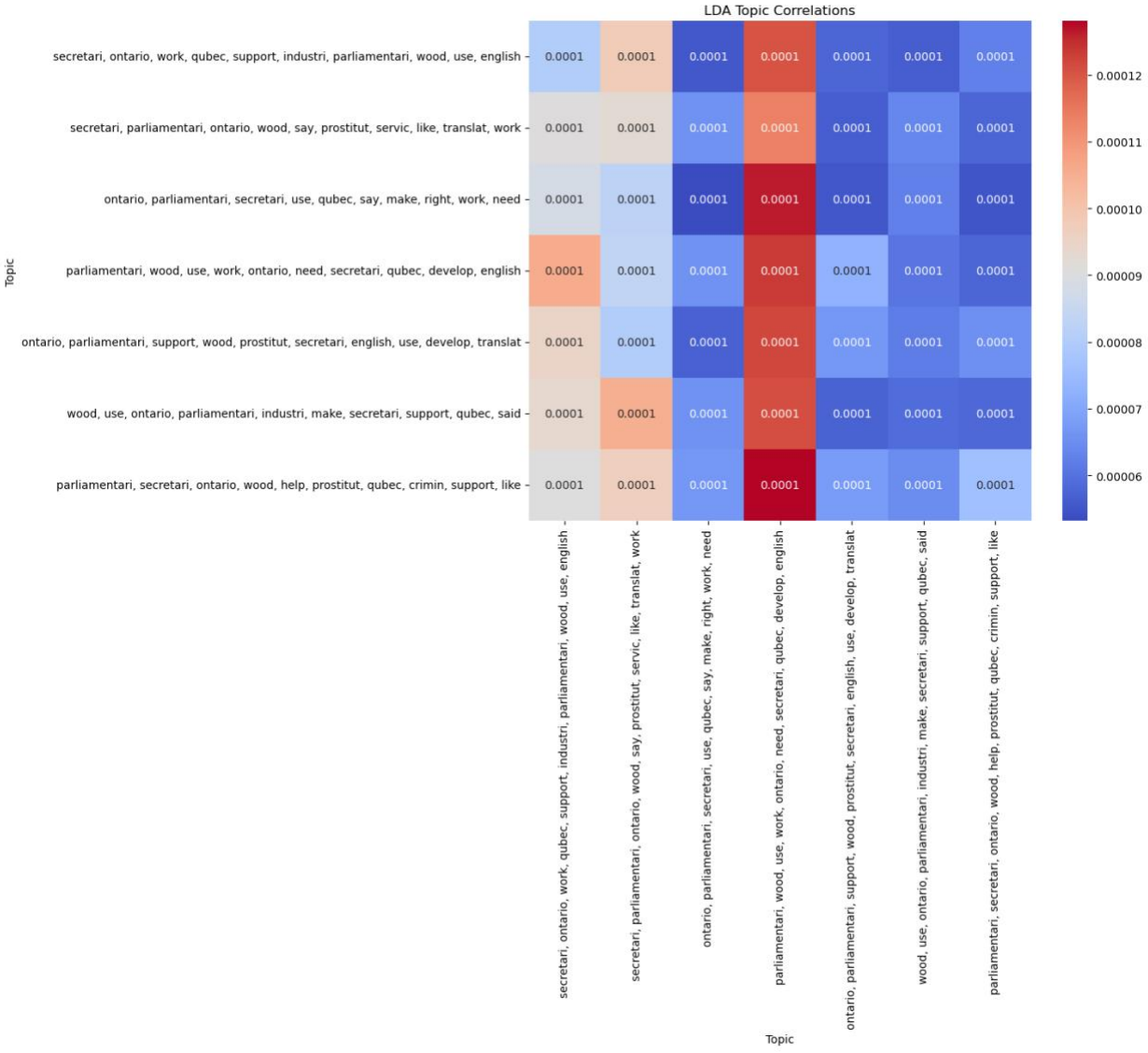
	Document	Dominant Topic	Topic Keywords	Unigrams	Bigrams	Trigrams
0	20170130-HAN129-E.pdf	6	regard, inform, statist, question, tabl, inclu...	aa, aandc, aandcinac, aban, abandon, abdic, ab...	aa amount, aandc indigenousand, aandcinac iden...	aa amount iap, aandc indigenousand northern, a...
1	20200420-HAN034-E.pdf	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,...

Histogram: Top 10 Topic Keywords (Frequency)



LDA Topic Correlations:

The different topics identified are weakly correlated. As identified in the literature review (Weston et al, 2023), identifying a representative label for each topic should be explored for the final paper.

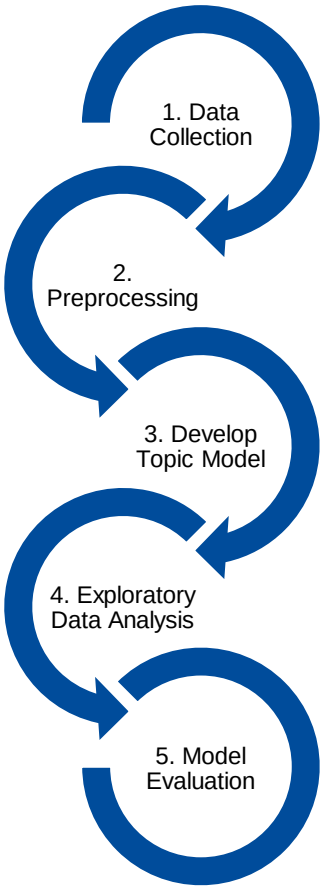


Word Clouds: Keywords from the 7 Topics

Research Methodology

The research methodology leveraged for this literature review 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 topic model
4. Exploratory Data Analysis including visualization and creation of a data frame with relevant attributes.
5. Model Evaluation



Github Repository

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

References

1. 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.

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

3. 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

4. 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

5. Grootendorst, M. (March 2022). “BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure”, Cornell University. https://arxiv.org/abs/2203.05794

6. 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

7. 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

8. 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

9. 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

10. 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.

11. 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

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