CIND820: Big Data Analytics Project

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

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

Date of Presentation: 01 April 2024

Toronto Metropolitan University

Parliament of Canada Hansard Debate Records: Project Objectives

Conduct topic modeling on the debate records using three different algorithms:

1. Will the application of unsupervised topic modeling produce meaningful insights into patterns and trends in Hansard debate records?

2. How relevant and meaningful are the identified topics from these models?

3. What are the general advantages and limitations of the different topic model algorithms, LDA, HDP and BERTopic?

Toronto Metropolitan University

Hansard Debate Records:
Data Source & Formatting Files

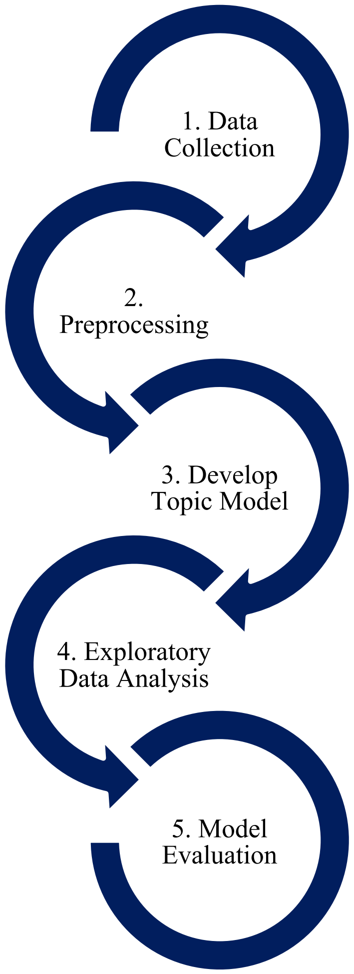
- House of Commons' Hansard archives

- Un-indexed debate records from the start of the 39th Parliament up to the current 44th Parliament (Apr 2006-Dec 2023)
  - Period of time covers several election cycles and two Prime Ministers.
  - (YYYYmmdd-HAN###-E.pdf) = 20230323-HAN172-E.pdf

- Dataset is comprised of:
  - 1972 PDF documents
  - 155,385 pages (median of 80p/doc)
  - 128,933,818 words

Toronto Metropolitan University

Research Methodology

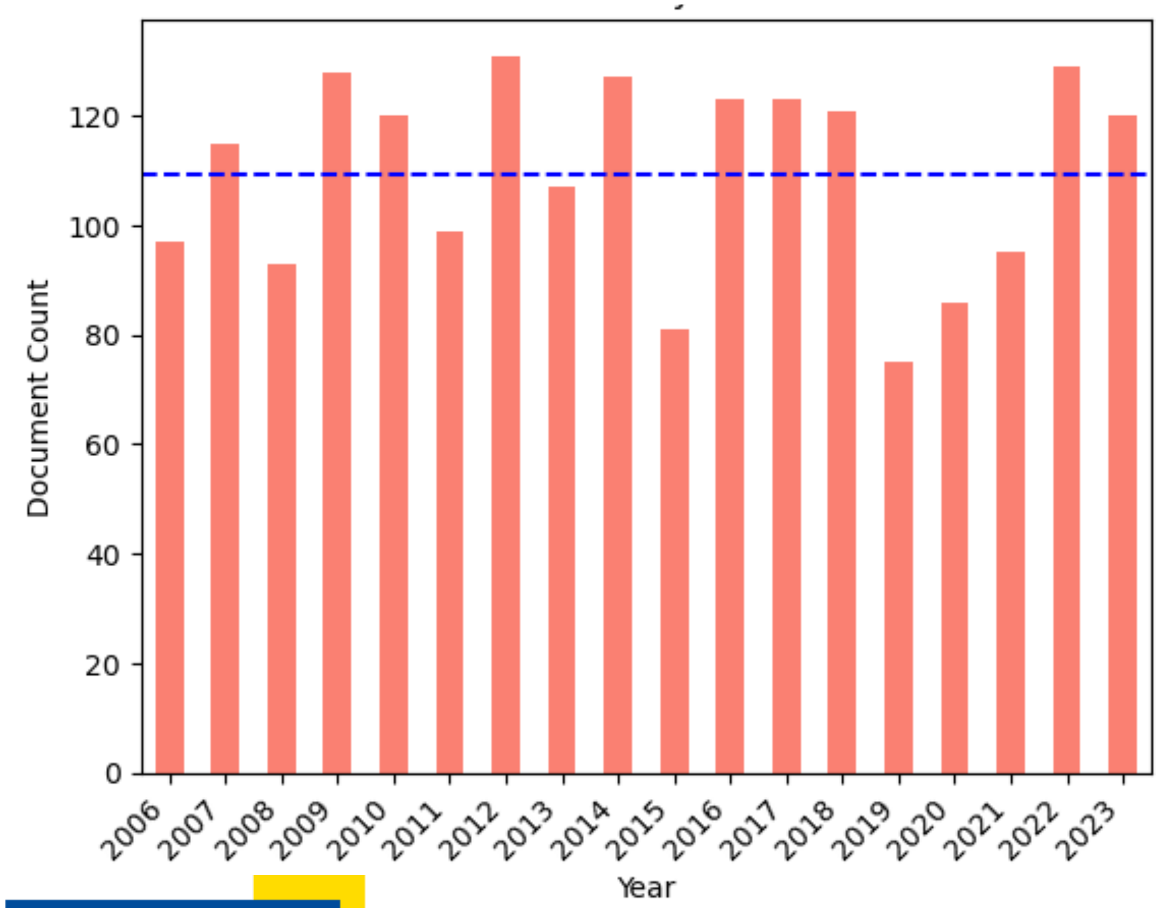


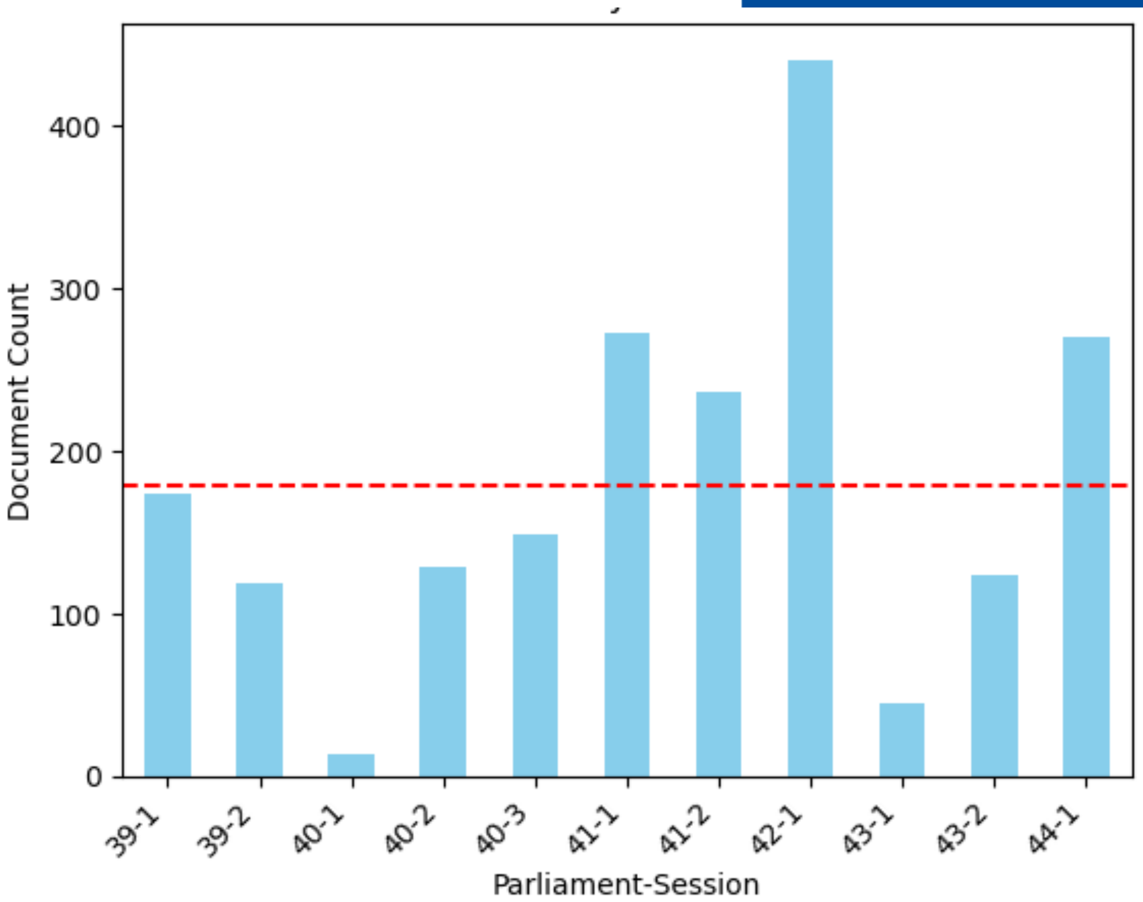
1. The identification of the data source and the collection of raw data (initial formatting)
2. Pre-processing the information and extracting text
3. Development of the LDA, HDP and BERTopic models
4. Exploratory Data Analysis including visualization and creation of a data frame with relevant attributes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

Completed many iterative cycles through methodology, including developing python script using samples of dataset to test out functionality

Toronto Metropolitan University

Exploratory Analysis

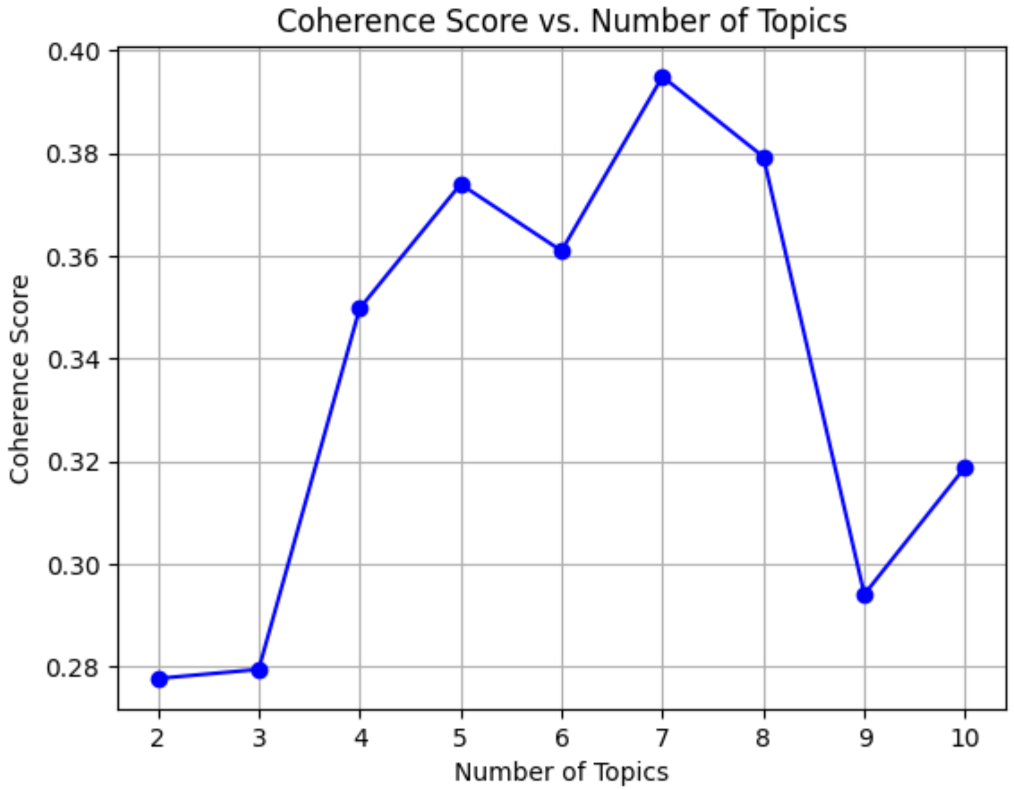


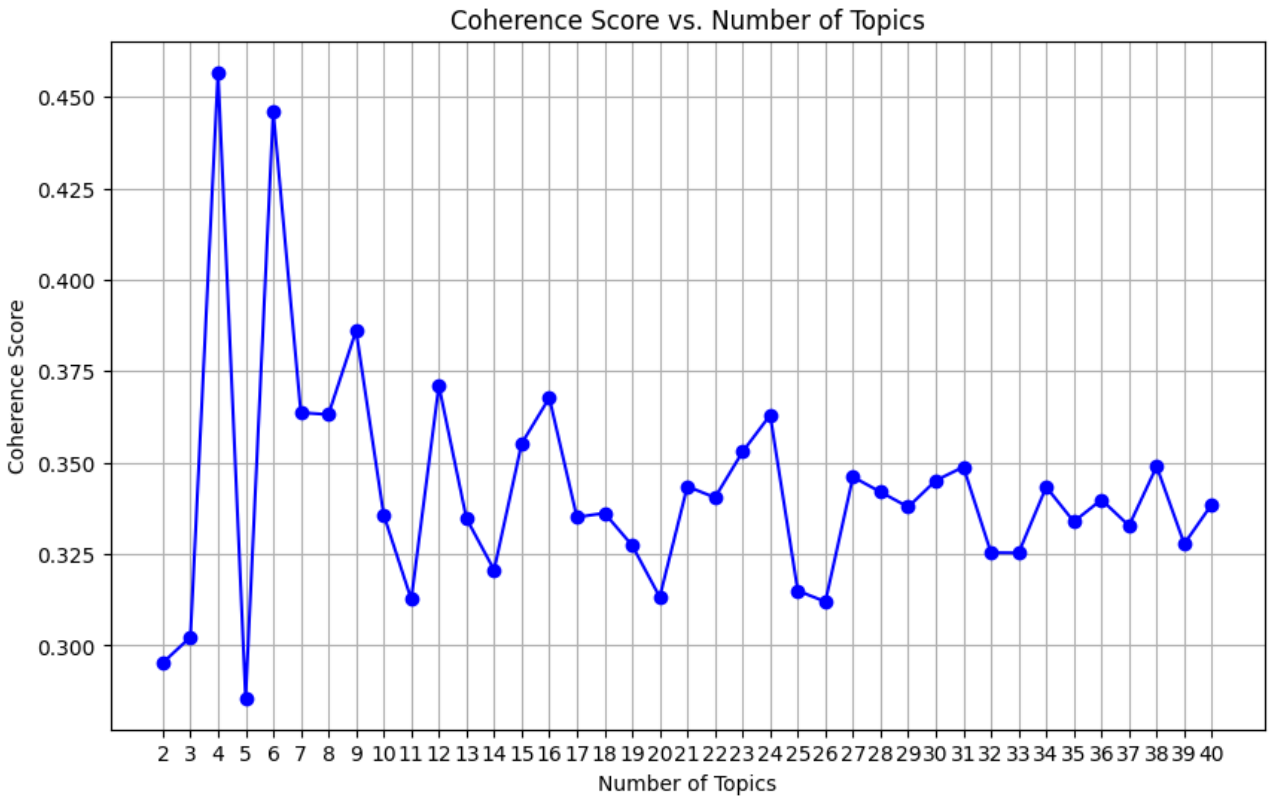


Toronto Metropolitan University

Preprocessed Text – Support LDA and HDP Models

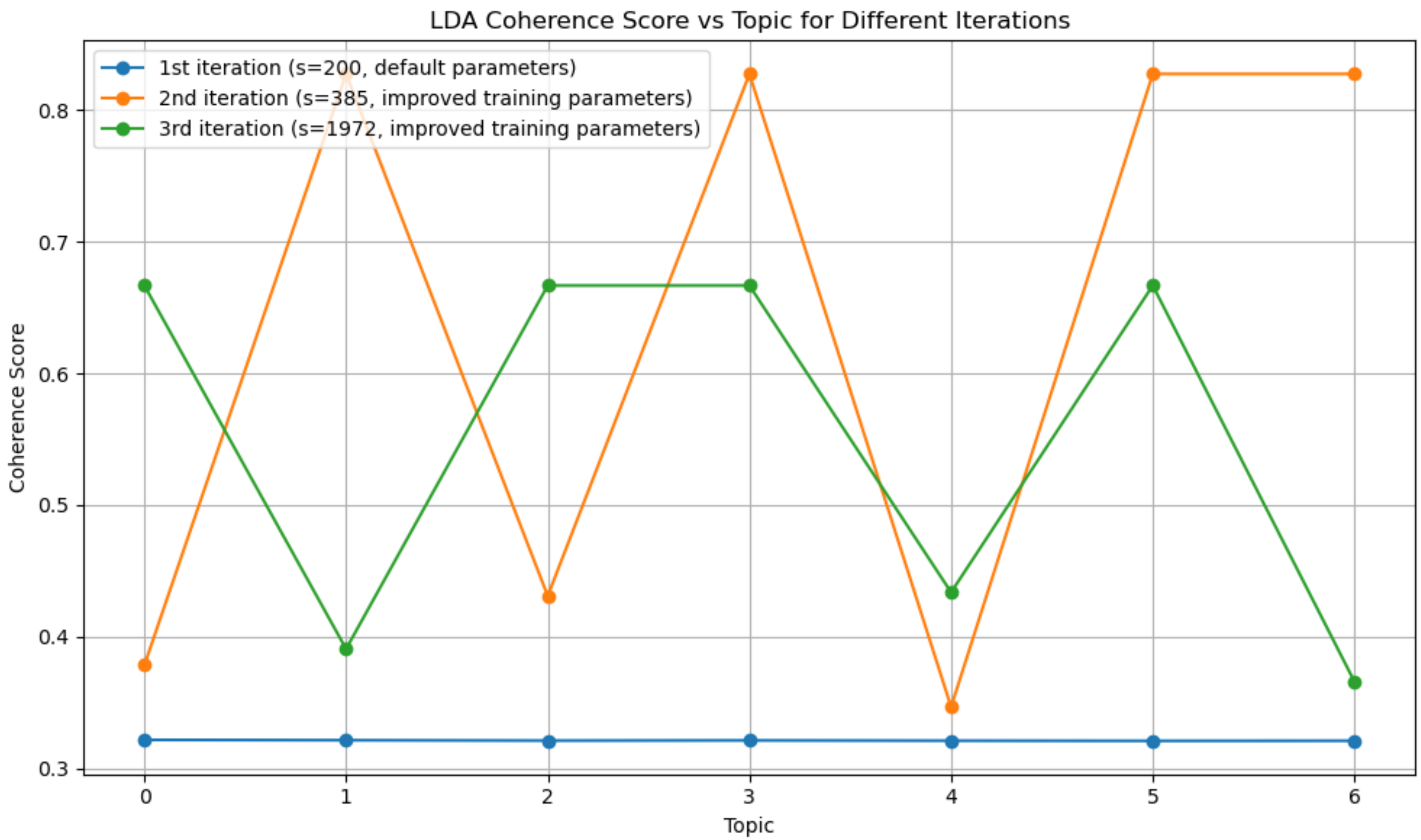
Assessment of ideal number of topics for LDA Model





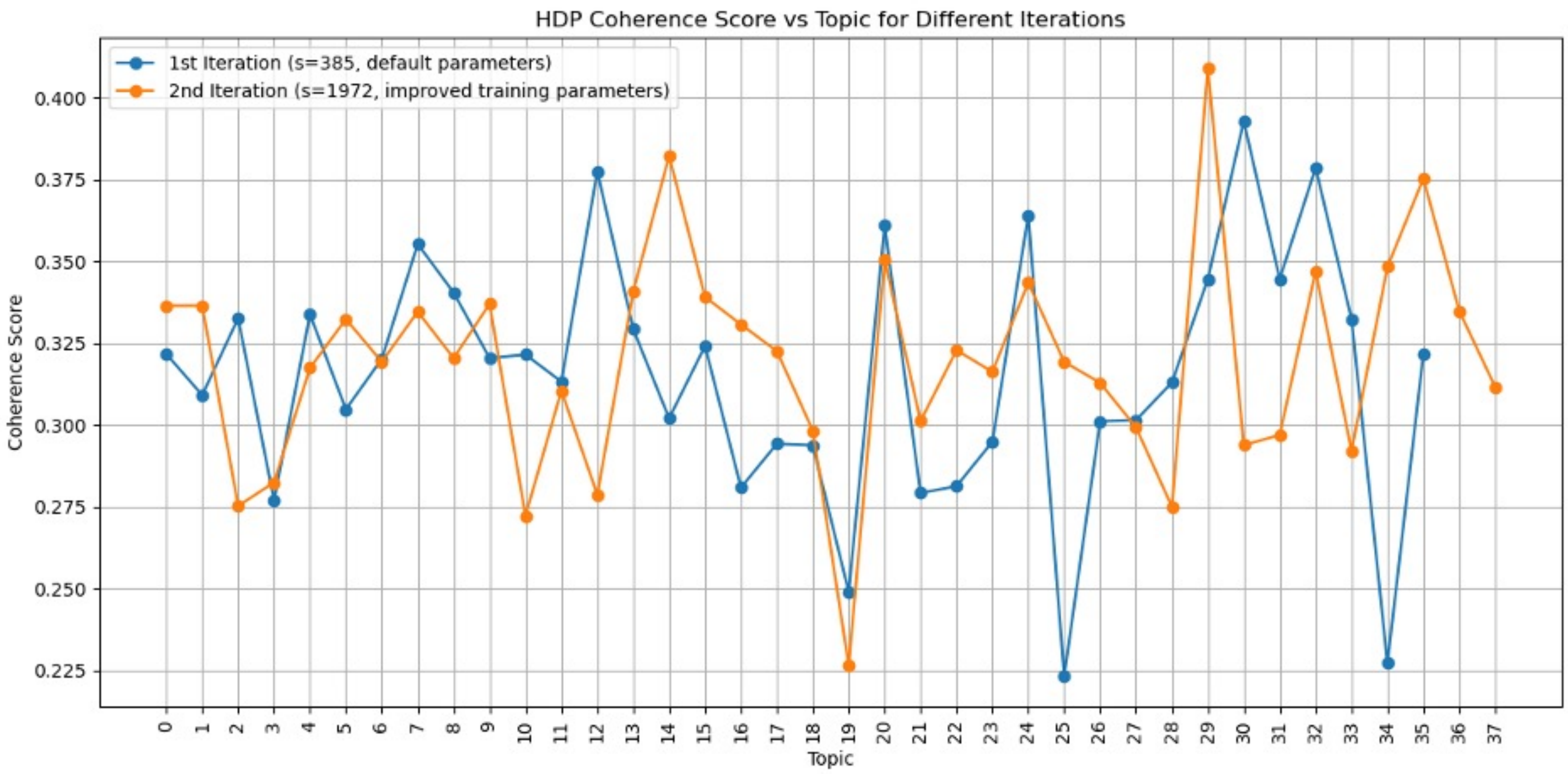
Toronto Metropolitan University

LDA Overall Coherence Values



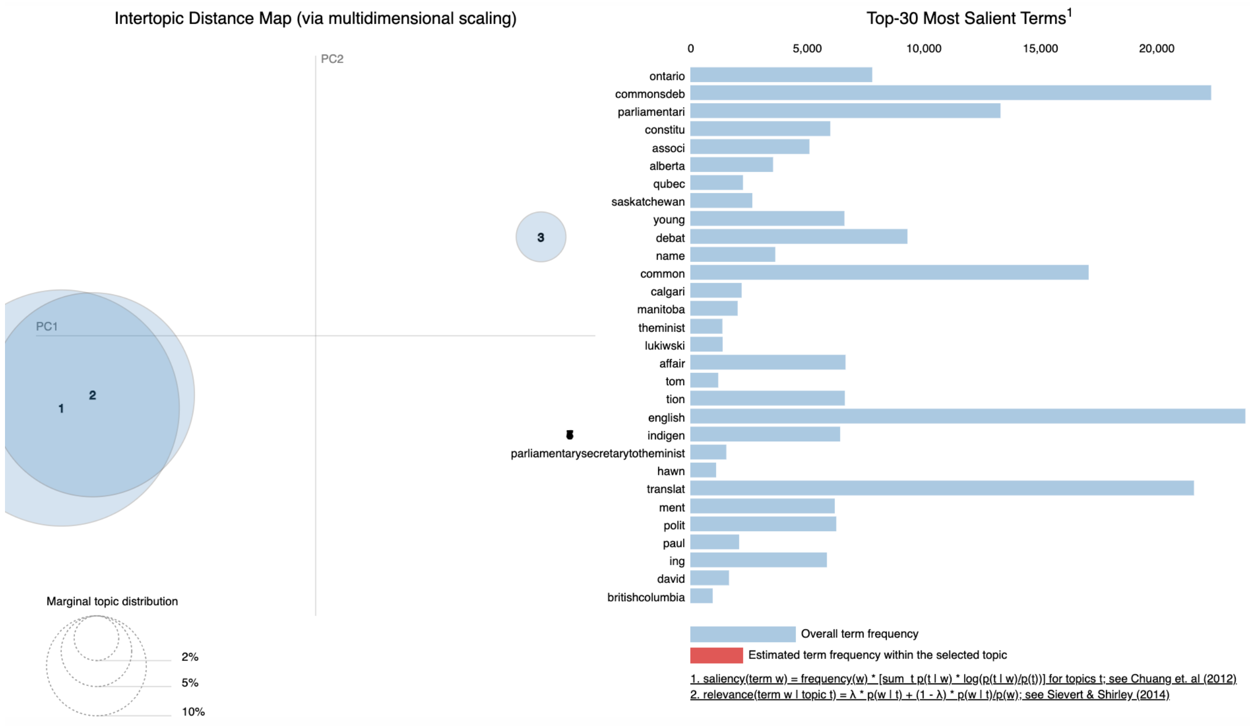
Model	1st Iteration (s=200, default parameters)	2nd Iteration (s=385, improved training parameters)	3rd Iteration (s=1972, improved training parameters)
LDA	0.3209934375	0.7010199998587691	0.628447487531789

HDP Overall Coherence Values

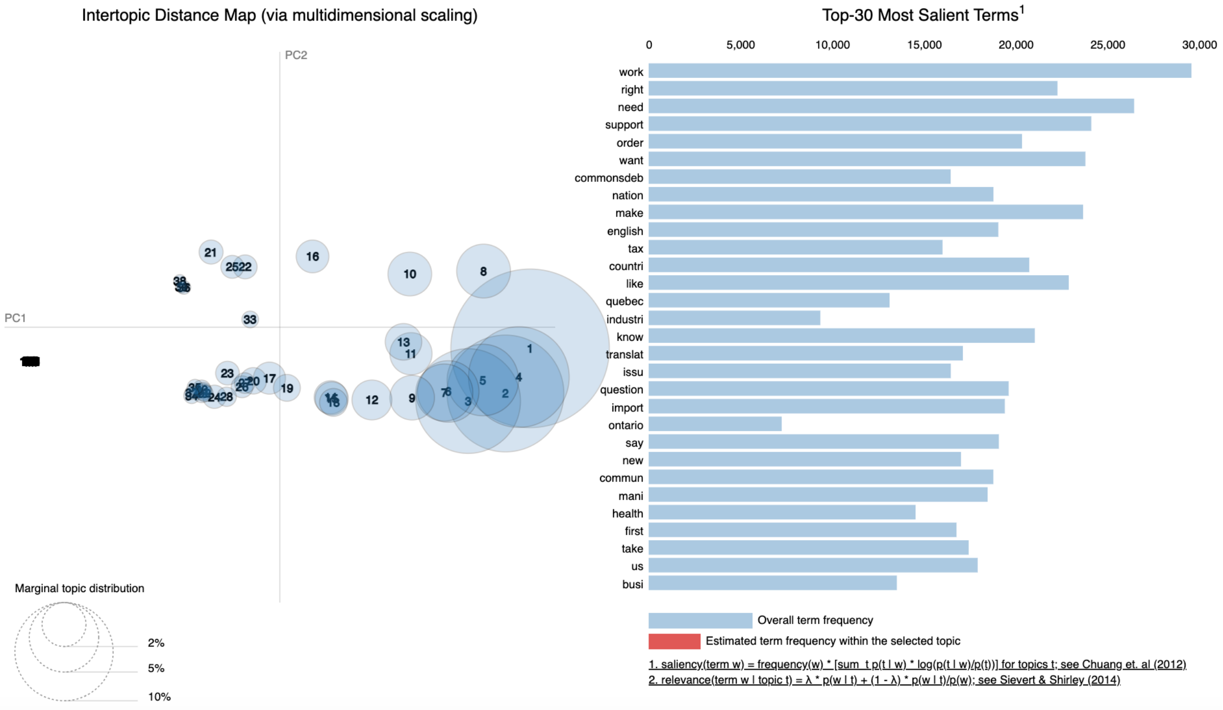


Model	1st Iteration (s=385, default parameters, 35 topics)	2nd Iteration (s=1972, improved training parameters, 38 topics)
HDP	0.31535182848347443	0.3194660065275221

LDA Topics

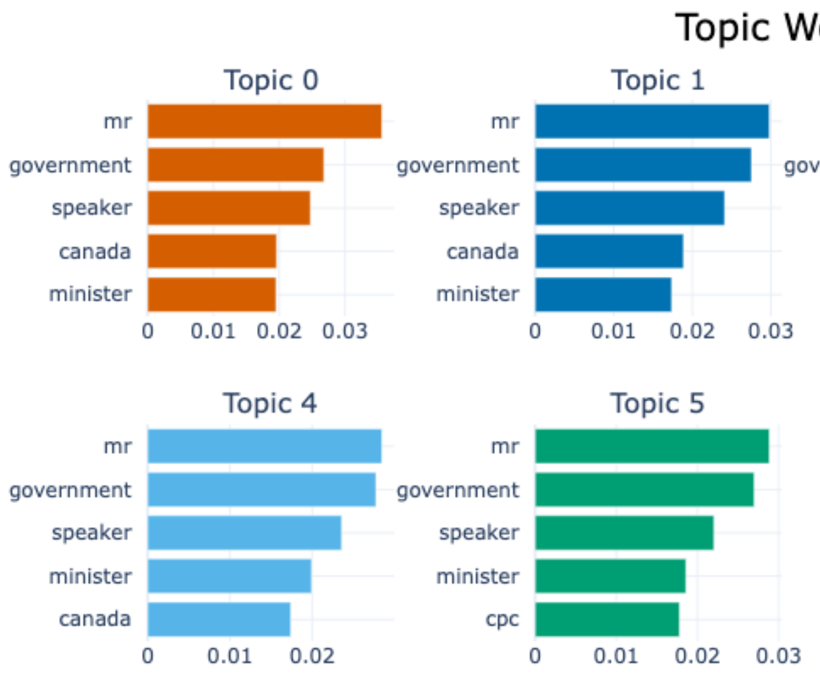


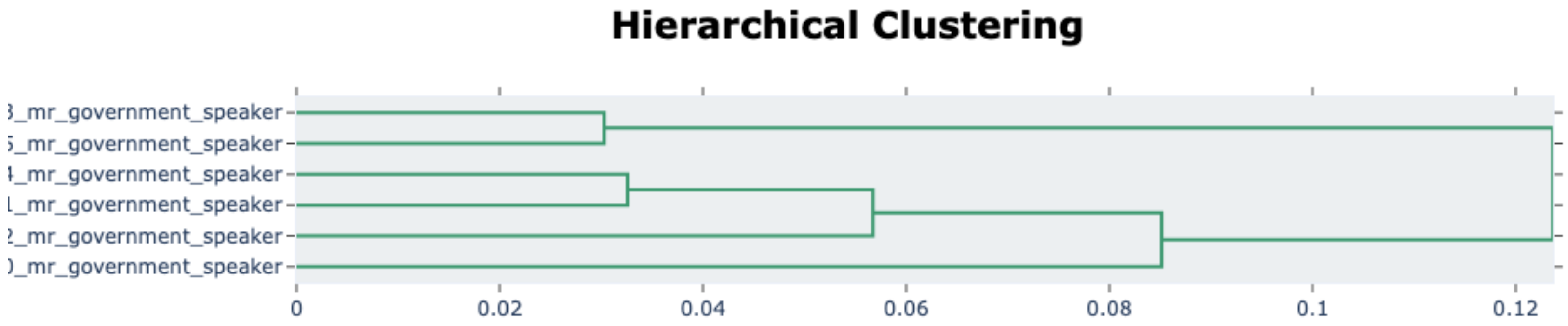
HDP Topics



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 – General Observations





Toronto Metropolitan University

Limitations – Computing Power and Duration

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

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

Toronto Metropolitan University

Future Considerations:
n-grams, stop words, stability and model parameters

	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,...

Toronto Metropolitan University

Revisiting Research Questions

1. Will the application of unsupervised topic modeling produce meaningful insights into patterns and trends in Hansard debate records?

2. How relevant and meaningful are the identified topics from these models?

3. What are the general advantages and limitations of the different topic model algorithms, LDA, HDP and BERTopic?

Toronto Metropolitan University

Toronto Metropolitan University