**Topic Modelling / Clustering of COS debates (PQ questions)**

Data

Official Reports, Parliamentary Debates (Hansard) of the (last 5 parliaments) 14th Parliament (24-08-2020 to Present)

[https://sprs.parl.gov.sg/search/#/home](https://sprs.parl.gov.sg/search/" \l "/home)

Task

To explore common topics from COS debates that has taken place within the 14th Parliament.

1. Write a web scraper to scrape the COS official reports from the Parliament website. Clean and preprocess your data.
2. Unsupervised topic modelling: Find the most suitable number of topics of discussion for each Parliament (state the metric used and why). Do not use LDA method. Visualise the topics and show the impact of different words on the topic (see pyLDAviz for an example of topic visualisation)
3. Every year, what is the most important topic of each parliament and how has it changed across each parliament?
4. Identify the main themes within the President's address of each year and identify the top 10 related PQs to each of the themes for the year. Which President's address themes align closely to the PQs of the year and which do not?
5. An example of the presidents address for 2023. You are encouraged to find other resources related to this address to aid in you task. <https://www.istana.gov.sg/Newsroom/Speeches/2023/04/10/Speech-by-President-Halimah-Yacob-at-the-Opening-of-the-Second-Session-of-the-14th-Parliament>
6. Generate a title of each topic cluster each year + 300 word summary of each topic of discussion
7. Identify which MPs frequently asked questions about each topic

**Points**

b) 1. BERTopic

Define a range of topics or minimum topic sizes

per parliament how many times the minister spoke for all the mps

Google COS dates every year

Part a 10th to 14th parliament COS Data per parliament how many times each mp spoke for COS

1 barchart per parliament 1 axis is the ministers name and the other axis is the no. of times they spoke. Write the barchart using matplotlib python

**Current Task Run BERTopic 3 week exploration**

**-> Next Sprint next Wed**

BERTopic main purpose to understand Topic modelling

- Develop

- Final goal: Through discussion, to build new feature in the current Vertext

\*\*Pre-Task\*\*

Explore BERT Topic in dataset - play around in textual data (Question a)

- Not crawled... need to crawl (last year COS Mar/Apr 2023)

Do we need to use open-source or fine-tuned?

1) If use BERT Topic, what library to use?

2) Try BERT Topic on last year COS (How it aligns to the questions?) -> Do on Jupyter notebook; say what data -> Find Topics discussed from COS?

3) Propose to front-end or backend (standard can like follow the Vertext backend)

e)

President address

- Based on COS discussion

- Attribute what the president say (maybe can use BERTopic?)

f)

Suggest a title - use any language model BERT or BARD

g)

- Propose how to design a front-end

- Propose how to design backend

HANDSAD https://sprs.parl.gov.sg/search/#/home

Parlimentary Seatings - 4 seatings in 4 years

Commitee of Supply - Every public agency need money - Public Servant ask questions to get enough money....

HANDSAD for topic modelling -> a lot of text