Automated Extraction of Parliamentary Information and Updates Using Artificial Intelligence

All Protocol Observed: Journalist focused assistive dashboard for parliamentary data

INTRO

- Parliamentary data is unstructured documents with high amounts of data and communication.
- This is unfeasible for a single journalist to process and find interesting sources.

METHODS

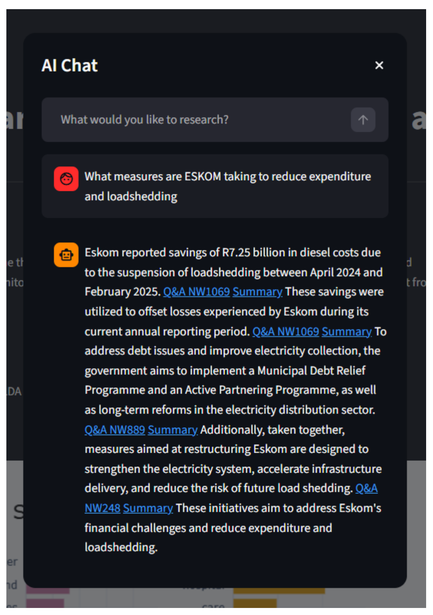
1. Collection of multimodal data. (Text, PDF files)
2. Extraction of text/OCR where needed.
3. Cleansing and topic mapping of overall topics
4. Ingestion into a vector database
5. Generative Summarisation of articles

RESULTS

- Clear topic clusters in the different modalities of data. e.g., Hansards, and questions and replies have overlapping topics.
- LLM-RAG pipeline yielded relevant information for searches
- Pipeline yielded searchable information, enhancing research capability for journalists.

DISCUSSION

Although being successful for English structured and unstructured text documents, the solution can benefit from additional modules to detect and translate non-English sources, as well as the conversion of audio transcripts to text to ensure less lag between parliamentary events and accessibility for journalists



Dreyer JG, **Smith** A

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Department of Computer Science

Faculty of Engineering, Built Environment and Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en Inligtingtegnologie / Lefapha la Boetšenere, Tikologo ya Kago le Theknolotši ya Tshedimošo

Capstone Project - MIT 808

Course Coordinators:
Dr. Vukosi Marivate (vukosi.marivate@cs.up.ac.za)
Abiodun Modupe (abiodun.modupe@cs.up.ac.za)

Scan me