LAIRHUB – Lab Meeting

March 14, 2025

Participants: Sunny, Paris, Aravind, Divya, Nibras

Action Items

Project	Project Associate	Updates / Action Items	Resource/ Notes
ETTA News	Ruochen/ Nibras	Workflow for Document Representation and Clustering 1. Preprocessing the news Corpus (after ETL) - Collect documents (title + abstracts). - Remove stop words from the text. - One challenge might occur in deciding the vector length, we have specific data-type to manage the scenario. 2. Develop a document representation table 3. Feature Extraction: TF-IDF Weighting - Compute Term Frequency-Inverse Document Frequency (TF-IDF) for each term in the corpus. - Store TF, IDF, and TF * IDF values in the document representation table. 4. Defining a Threshold for Term Selection - Set up a threshold to differentiate between low-weighted, mid-weighted and high-weighted terms. - We will conduct an EDA to decide the threshold. 5. Query Handling and Alternative Workflow Decision - If a query term exists in high-weighted words, proceed with clustering. - If a query term does not exist in high-weighted words, consider an alternative workflow. 6. Our final goal is to use document representation table for retrieving results, clustering and using GA.	

		ETL & Document Distribution Analysis - Compute and demonstrate ● Our current ETL process ● ER diagram ● Average document length ● Maximum document length ● Minimum document length ● Distribution of document length ● top 10 (varies) high frequency words ● top 10 (varies) low frequency words ● In addition, we will find out whether our corpus follows Zipf's law. ● Generate following analysis - **(Later)** ○ Average number of tokens per clusters ○ Get the best tokens and then send them to llama (text preprocessing) ○ Per cluster - tariff & cluster ■ How long does it take to generate ■ How many tokens did I send ● Show the tokens we are sending to understand quality the labels that are coming back ● RQ: These tokens we are sending, many of the labels are not coming back as cluster labels. Later we will find out why it's happening. ■ Use stop word list ■ We can perform higher level of filtering ● TF-IDF (next steps) ■ Remove the query terms (tariffs & cancer) from the llama response ● We can drop it and ask for six labels from llama ■ Perform sanity check? ● Remove the other outputs from llama	

		■ We will focus on time and number of tokens (after using stop word list or TFIDF) ● Make short videos on ○ Incorporated features. ○ How do you perform a basic search with connections? ○ Zoom-in and zoom-out features. ○ Cluster size definitions and color. ○ How to get help. ○ How we constructed the database. ● MISC: ○ Attach askit4.info with CRM.	
ezee.chat	Sunny/Nibras	● User workflow ● Paper outline ● Identify the procedures for publishing the bot (as a research and private entity) on Meta's WhatsApp AI platform. ● Develop a detailed workplan ○ Use cases for newly-admitted students ○ MDL based WhatsApp bot ○ Technology & Research project ○ Publish & present ○ Translation - investment & funding ● Business number for ezee.chat ● Index for dialogue	**Existing Educational Bots** https://www.comm100.com/engage/chatbots-higher-ed/
Grant Submission	Sunny	● Grant Proposal Document Society - OVPI Horizon Europe Keep tracking-NRFK (National Research Foundation of Korea)	
showme.news	Aravind	● Integrate Tobii with showme.news	

Information Addiction **Wellbeing & Thriving**	Paris	• Wait for the feedback for ARIST	

LookupMed	Paris/ Divya	• **Check the recordings for details (7 March, 2025)**
		• Followings are the broader concepts
		  ○ Gerontology
		  ○ Medication compliance
		  ○ Interventional methods | implementation
			Documentation URL:
			https://support.google.com/googleplay/android-developer/answer/13628312?hl=en
			
			https://support.google.com/googleplay/android-developer/answer/10841920#zippy=%2Corganization-account