Analysis of city council meeting minutes

Institut des Sciences du Digital, Management et Cognition
Université de Lorraine

November 29, 2024

Oyetunji Abioye, Alberto Lorente Galé, Mina Oulhen and Ziyan Xu

1 Project Overview

2 Tasks

3 Classifier

4 Information Retrieval Stage

Project Overview

Datapolitics Project

- **Objective**: Develop an automated detector to identify and categorize projects implemented by local authorities.
- **Scope**: Examine around **20,000** geothermal energy PDF documents from the past five years. Ensure the methodology can be applied to various local projects

Data Overview

- **doc_id**: Unique identifier for each document.
- **url**: Original source URL of the document.
- **cache**: Link to the cached PDF version.
- **fulltext**: Link to the plain text version of the document.
- **nature**: Automatically classified document type (e.g., deliberation, minutes).
- **published**: Publication date of the document.
- **entity_name**: Name of the local authority responsible for the document.
- **entity_type**: Type of the entity (e.g., municipality, intercommunality).
- **geo_path**: Hierarchical administrative path indicating the geographical scope.

1. Project Overview
2. Tasks
3. Classifier
4. Information Retrieval Stage

Tasks

Filtering and Classification Process

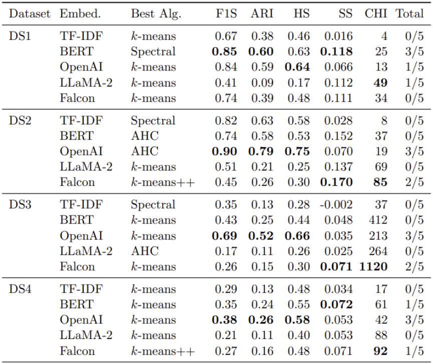
1. **First Level: Binary Filter**
   - **Concerns a Geothermal Project**
   - **Unrelated to a Geothermal Project**
2. **Second Level: Project Stages**
   - Idea/Wish
   - Preliminary Studies
   - Budget Voted for the Definitive Project
   - Implementation in Progress
   - Implementation Completed
3. **Final Level: Data Extraction**
   - Initial Budget
   - Final Cost
   - Estimated Duration
   - Actual Duration

1. Project Overview
2. Tasks
3. Classifier
4. Information Retrieval Stage

WHEN YOU HAVE TO ANNOTATE 20K DOCS BUT YOU'RE JUST A CHILL GROUP

Exhausted options to deal with the annotation process:

- Generating the annotation via clustering with LLM Embeddings [1].



Dataset	Embed.	Best Alg.	F1S	ARI	HS	SS	CHI	Total
DS1	TF-IDF	k-means	0.67	0.38	0.46	0.016	4	0/5
	BERT	Spectral	**0.85**	**0.60**	0.63	**0.118**	25	3/5
	OpenAI	k-means	0.84	0.59	**0.64**	0.066	13	1/5
	LLaMA-2	k-means	0.41	0.09	0.17	0.112	**49**	1/5
	Falcon	k-means	0.74	0.39	0.48	0.111	34	0/5
DS2	TF-IDF	Spectral	0.82	0.63	0.58	0.028	8	0/5
	BERT	AHC	0.74	0.58	0.53	0.152	37	0/5
	OpenAI	AHC	**0.90**	**0.79**	**0.75**	0.070	19	3/5
	LLaMA-2	k-means	0.51	0.21	0.25	0.137	69	0/5
	Falcon	k-means++	0.45	0.26	0.30	**0.170**	**85**	2/5
DS3	TF-IDF	Spectral	0.35	0.13	0.28	-0.002	37	0/5
	BERT	k-means	0.43	0.25	0.44	0.048	412	0/5
	OpenAI	k-means	**0.69**	**0.52**	**0.66**	0.035	213	3/5
	LLaMA-2	AHC	0.17	0.11	0.26	0.025	264	0/5
	Falcon	k-means	0.26	0.15	0.30	**0.071**	**1120**	2/5
DS4	TF-IDF	k-means	0.29	0.13	0.48	0.034	17	0/5
	BERT	k-means	0.35	0.24	0.55	**0.072**	61	1/5
	OpenAI	k-means	**0.38**	**0.26**	**0.58**	0.053	42	3/5
	LLaMA-2	k-means	0.21	0.11	0.40	0.053	88	0/5
	Falcon	k-means++	0.27	0.16	0.48	0.071	**92**	1/5

Results from Petukhova et al. (2024)

Educated (?) wish: the embedding distance captures the project vs non-project relationship.

Process for cluster explainability: 2 clusters for each document type. Train a RF model with TF-IDF transformations of the text. Look at the most important features.

Top 20 most important words for different document types:

- **acte.arrete**: région, direction, subdélégation, recueil, agrément, relatif, mission, transport, arrêter, mois, département, décret, national, devoir, e, être, signer, maire, faire, délai.
- **bdj**: local, avoir, fonds, travail, investissement, compte, service, être, opération, personnel, subvention, public, recette, général, année, entrer, devoir, dépense, budget, charge.

inconclusive results = prompt LLAMA

1. Project Overview
2. Tasks
3. Classifier
4. Information Retrieval Stage

RAG

After setting up a RAG *poc* and doing some trials with Llama 3.2 1B the retrieval was not as accurate as we hoped. Example with one document loaded with *sentence-camembert-base* embeddings:

US: C'est un project?

LLAMA: Il y a des projets dans l'extrait du registre des délibérations du Conseil municipal le 1er juillet 2022.

Node

> Mis en ligne le : 12/07/2022 - vUle t1eV n1ers.-le.-bel Séance du Conseil Municipal du vendredi l juillet 2022 Extrait du Registre des Délibérations du Conseil Municipal Ville de Villiers-le-bel Séance ordinaire du vendredi 1 juillet 2022 N°1/Com pte rendu Compte rendu commenté (procès-verbal) de la séance du Conseil Municipal du 24 mai 2022 Le vendredi.

Chain-of-Thought Prompting

So we are using Chain-of-Thought (CoT) prompting to guide a GPT4-o model towards structured, logical answers. But its not working that well either:

US: You are a helpful AI Question Answerer. You are about to be passed a project description text in French by a User. Your task is to extract the estimated duration of the project. Make sure to generate a series of logical steps and reason about the problem before generating a solution. (...)

GPT: The actual duration is: 3. Chain-of-Thought: (...) in the context here of understanding a duration from a publishing perspective, the publication covered listings of 3 pages regarding a decision from 25th August. Hence, the most visible duration specifiable from **document length is 3 pages**, interpreting via the listings as grasped directly from the text.

Conclusion

Moving forward:

- Prompt LLMs to annotate the target variable for the binary classifier.
- Keep working on the RAG system with CoT generation.

References

A. Petukhova, J. P. Matos-Carvalho & N. Fachada – "Text clustering with llm embeddings", 2024.