Analysis of city council meeting minutes

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

October 11, 2024

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

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.

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

Almost 20.000 documents with no annotated target variable.

Options to deal with the annotation process:

- Manual annotation.
- Prompting.
- Generating the annotation via clustering with LLM Embeddings [3].

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

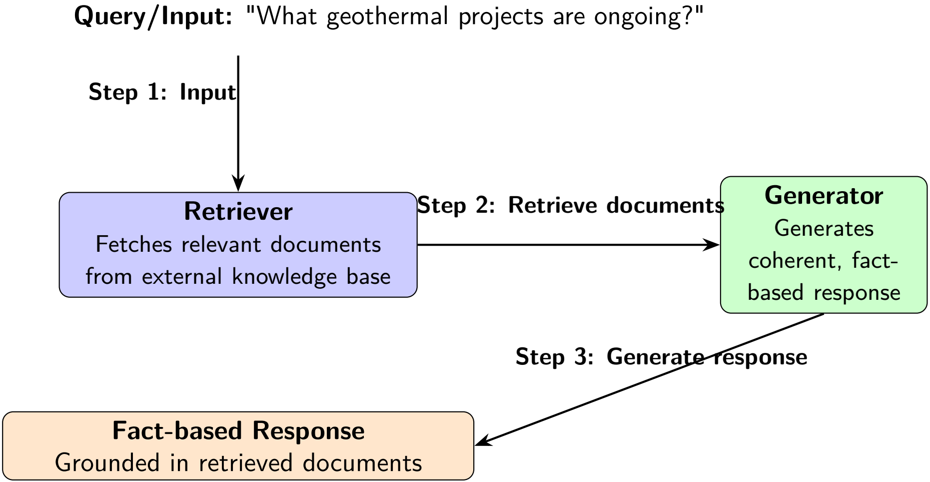
Results from Petukhova et al. (2024)

Next step: train the classifier

Considerations when training the classifier with automatically generated labels:

- Validating the labels generated: manual and automatic approaches.
- Training with label noise [4].

RAG (Retrieval-Augmented Generation) Workflow: A Simplified View



Why and How to use RAG in our project

WHY RAG?

- **Handling Document Complexity**
  The large volume of documents is diverse in format and content.
- **Relevant NLP Tasks**
  Ensures high-quality, contextually accurate results for tasks like classification and extraction.

HOW TO USE RAG

- **Application in Our Project:**
  - Document Filtering: Use the retriever to pull relevant documents from a dataset of 20,000 PDFs.
  - Classification and Stage Identification: Use the generator to classify document stages.
  - Information Extraction: Extract budget and timelines conditioned on relevant document sections.

References: Lewis et al. (2020)[2], Izacard & Grave (2021)[1]

References

G. Izacard & E. Grave – "Leveraging passage retrieval with generative models for open domain question answering", in *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume* (Online) (P. Merlo, J. Tiedemann & R. Tsarfaty, éds.), Association for Computational Linguistics, 2021, p. 874–880.

P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel & D. Kiela – "Retrieval-augmented generation for knowledge-intensive nlp tasks", 2021.

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

H. Song, M. Kim, D. Park, Y. Shin & J.-G. Lee – "Learning from noisy labels with deep neural networks: A survey", 2022.