

Analysis of city council meeting minutes

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

February 18, 2025

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

- 1 Introduction
 - Project Overview
 - Approaches
 - Limitations and Scope
- 2 Chain-of-Thought
- 3 RAPTOR Implementation
- 4 Conclusion

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

Tasks: Filtering and Classification Process

- 1 **First Level: Binary Filter**
 - **Concerns a Geothermal Project**
 - **Unrelated to a Geothermal Project**
- 2 **Second Level: Project Stages**
- 3 **Final Level: Data Extraction**

① Chain-of-Thought

② RAPTOR, Sarthi et al. (2024): RAG system in 2 parts

- ① Bottom-up construction of a tree through clustering and generation of summaries
- ② Information retrieval from this tree, allowing different levels of abstraction

- Limited scope: small models
- Working with French data: the issue of finding appropriate monolingual sentence embeddings
- **The data itself:**
 - 1 Presented as: Council minutes
 - 2 Then actually: Data related to (geothermal) projects

Limitations and Scope - The data, in reality



Limitations and Scope - The data, in reality

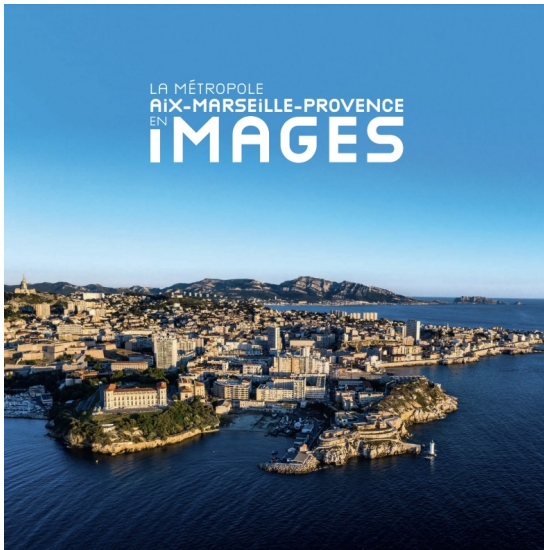
Source OFFICE DES ENTREPRISES BAGNOLS SUR CEZE AGGLOMERATION GARD RHODANEN SEPT 2023
04.66.79.38.00
officesentreprises.fr

LAUDUN L'ARDOISE
04.66.50.55.50
contact@laudun-lardoise.fr

SERVICE COMMUNICATION 25/09/23
04.30.09.03.26
communications@leeds.gov.uk

[illegible]

Limitations and Scope - The data, in reality



Limitations and Scope - The data, in reality



Limitations and Scope - The data, in reality



• NOURRIR

Une ville qui investit dans l'éducation pour nourrir les capacités humaines.

• PROTÉGER

Des services publics municipaux qui prennent soin de chacune et chacun.

• ANTICIPER

Comment Montreuil agit-elle contre le réchauffement climatique?

 • BILAN DE MI-MANDAT

- 1 Introduction
- 2 Chain-of-Thought
- 3 RAPTOR Implementation
- 4 Conclusion

Structured Output

Context:

- LLMs are excellent at generating free-form text, but many applications require structured data.
- Parsing raw text into a reliable format (e.g., JSON) is essential for further processing.

Solution:

- Use **Pydantic**, a Python library based on type hints, to validate structured data.
- Define specific models to validate outputs like project budgets and durations.

Structured Output

```
from pydantic import BaseModel, Field

class GeothermalProjectDetection(BaseModel):
    is_geothermal: bool = Field(description="Whether  
the document concerns a geothermal project")
    chain_of_thought: str = Field(
        description="Logical Steps that were taken  
to derive  
the final concluding statement"
    )
```

Usage:

- The LLM is prompted to generate JSON that fits these models.
- Pydantic then validates and parses the JSON output into structured Python objects.

Binary Filter

```
geothermal_detection_message = {  
    "role": "system",  
    "content": """  
        Your task is to determine if the text  
        concerns a geothermal project.  
        Consider mentions of:  
        - Geothermal energy installations  
        - Geothermal studies or surveys  
        - Heat pump systems  
        // ...additional guidance...  
    """  
}
```

Project Stage Classification

```
stage_classifier_message = {  
    "role": "system",  
    "content": f"""  
        Your task is to determine the stage  
        of the geothermal project  
        based on the text. Classify the  
        project into one of the following:  
        - idea/wish  
        - preliminary studies  
        - budget voted for the definitive  
        project  
        - implementation in progress  
        - installation completed  
        // ...additional guidance...  
    """,  
}
```

- Followed similar custom prompt engineering methods as the earlier steps.
- Returned Initial Budget and Final Cost as Float point values
- Returned Estimated Duration and Actual Duration as Int values measured in days

Chain-of-Thought Reasoning

- Each classification and extraction task is supported by explicit chain-of-thought reasoning. These reasoning steps are essential to ensure that decisions are transparent and can be verified.
- **Temperature Settings:** A temperature value of approximately **0.1** is used to enforce consistency and reduce randomness in the model's logical reasoning.
- **Benefits:**
 - Helps in debugging the decision-making process.
 - Allows technical teams to understand and trust the automated decisions.

Results

```
{  
  "8109": {  
    "original_text_path": "path/to/original_text",  
    "english_path": "path/to/english_translation",  
    "initial_budget": 100000000.0,  
    "final_cost": 45000000.0,  
    "estimated_duration": 365,  
    "actual_duration": 365  
  }  
}
```

COT:

- **initial_budget_chain_of_thought:** Identified the Plan Local de Prévention des Déchets Ménagers et Assimilés (PLPDMA) as a key project. Extracted the initial budget of 100 million euros mentioned in the report as the budget for this plan.
- **final_cost_chain_of_thought:** To extract the final cost of the project, I analyzed the text for any mention of financial figures related to the project ... Therefore, the final cost of the project is 45,000,000 euros.
- **estimated_duration_chain_of_thought:** The project description text mentions several actions and plans implemented throughout the year 2023 ... which is equivalent to 365 days.
- **actual_duration_chain_of_thought:** The project description spans the entire year 2023, as evidenced by multiple actions and initiatives implemented throughout the year ... the actual duration is 365 days.

- 1 Introduction
- 2 Chain-of-Thought
- 3 RAPTOR Implementation**
- 4 Conclusion

Data Processing

- Partition of markdown text in chunks of n sentences.
- Embed those partitions with Gemma-2b.
- Cluster the embeddings into an optimized number of clusters.
- Generate summaries for each cluster with Gemma2-9b.
- Look for tables in the documents and generate descriptions with Llama-3.2-11B-Vision.

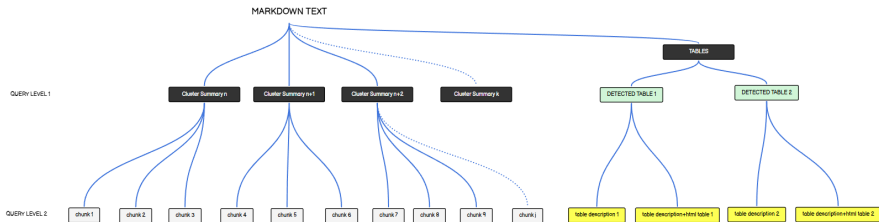
Querying and Retrieval

Querying

- Query the clusters summaries and retrieve the cluster id.
- Query the chunks belonging to said cluster and retrieve relevant facts.
- Query the table transformations.
- Generate summaries for each cluster with Gemma2-9b.
- Look for tables in the documents and generate descriptions with Llama-3.2-11B-Vision.

What is fed into the Augmented Generation Model?

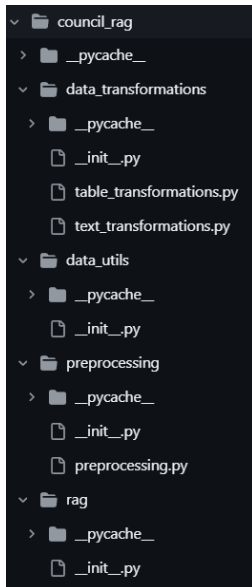
- Relevant chunks coming from the most similar cluster.
- Relevant tables descriptions and relevant table descriptions + html.



Technical Considerations

- As documents get larger, so will the clusters: since the LLM we are using to summarize can only take so many tokens, the minimum number of clusters permissible has to grow along the length of tokens in the document.
- The Groq models are queried in their standard temperature and max completion configuration.

Modules!



Final Results

Sur la partie de la page, il y a mention de la date du 22 septembre et de la légende de samedi 8 h.", "final_response": "D'après le texte fourni, le projet géothermique se situe dans la phase des ****études préliminaires****.

Voici les arguments qui mènent à cette conclusion :

- * ****Recherche de financements:**** La phrase "le but étant de rechercher des financements (CAN, DDT, CAF) sans quoi, le projet ne pourra se faire" indique clairement qu'il s'agit d'une phase pré-financier, où le projet est encore en développement et nécessite des validations financières.
- * ****Définition du coût approximatif:**** Le fait que le coût du projet soit estimé à 2 550 000€ HT, mais qu'il puisse être affiné, suggère que le projet n'a pas encore été finalisé ni chiffré de façon définitive.
- * ****Étude de l'option géothermie:**** La mention de la nécessité de faire appel à un bureau d'études spécialisé pour étudier l'option de géothermie comme moyen de chauffage montre que cette partie du projet n'est pas encore intégrée formellement.
- * ****Adaptation aux commentaires:**** Le bureau d'étude a présenté "le projet amélioré en tenant compte des modifications/remarques demandées" suggère que le projet est encore en cours d'évolution et d'adaptation.

Ces éléments combinés pointent vers une étape préliminaire du projet géothermique, où une étude approfondie est réalisée avant le lancement officiel de la mise en œuvre.

Figure: Conseil Municipal de Coulon, 12 July 2023

After carefully analyzing both the evaluation output and the original document:

- The conclusion is well-structured, precisely extracting the geothermal-related section and budget details from the PDF, while maintaining clarity on its current status.
- It states that further studies are needed before launching the project, avoiding any misinterpretation.

- 1 Introduction
- 2 Chain-of-Thought
- 3 RAPTOR Implementation
- 4 Conclusion

Conclusion

- We have developed a **library** that provides a framework for experimentation and improvements.
- The challenges encountered emphasize the need for **better data quality and preprocessing**.
- The results remain **inconclusive** as we are not utilizing state-of-the-art models for RAPTOR.
- We recommend adopting **better data engineering practices** to enhance performance.
- We hope **monolingual French embedding models** with **long-context capabilities** that can be run locally are developed .

Using Databases for Vector Search

- Traditional DBs support vector search but are **not optimized** for high-dimensional similarity queries.
- Could avoid computing all embeddings upfront, saving **time/cost** for unused texts.

Automatic Annotation via Embeddings

- Continue explore clustering for **automatic annotations**.
- Test **continual pretraining** with noisy data as new data arrives.

Sarhi, P., S. Abdullah, A. Tuli, S. Khanna, A. Goldie, and C. D. Manning (2024). Raptor: Recursive abstractive processing for tree-organized retrieval.