D-RAG

Densifying Retriever for Appropriate Generation

The problem

- Gieni is an LLM agent answering deep reasoning questions over the highly specialized knowledge sector of supply chain management
- Gieni needs to make best use of the sparse information found on company websites

...by analysing our user persona we found that:

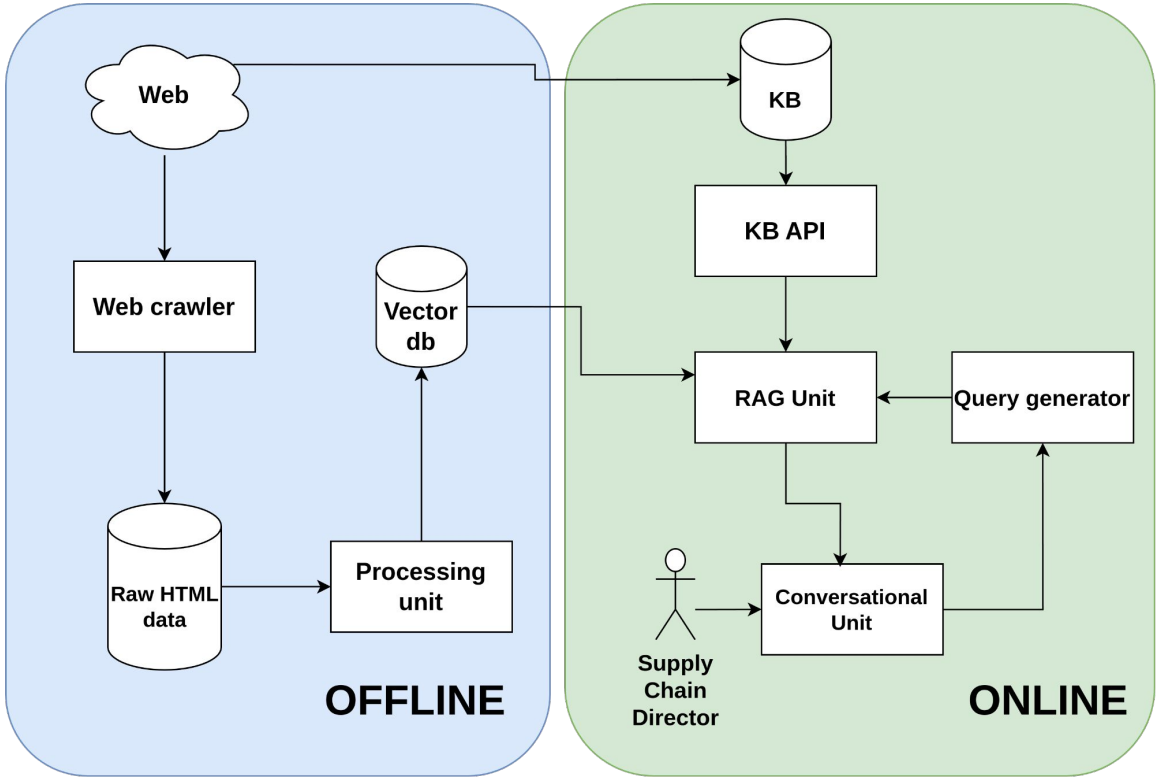
- Spatial awareness is key to provide accurate localized reports
- Feature engineering and query augmentation are required to deepen question answering in the supply chain analysis

The solution

A RAG system designed to:

- Increase the information density, with a lightweight large-scale dataset preprocessing
- Leverage the topological properties of the set
- Incorporate additionally engineered features on geolocation (and potentially much more)
- Rely on KB and anthologies to narrow the user's request by context augmentation
- Run multiquery beams during the conversation (parallel historical data)

The Big picture



Pre-processing

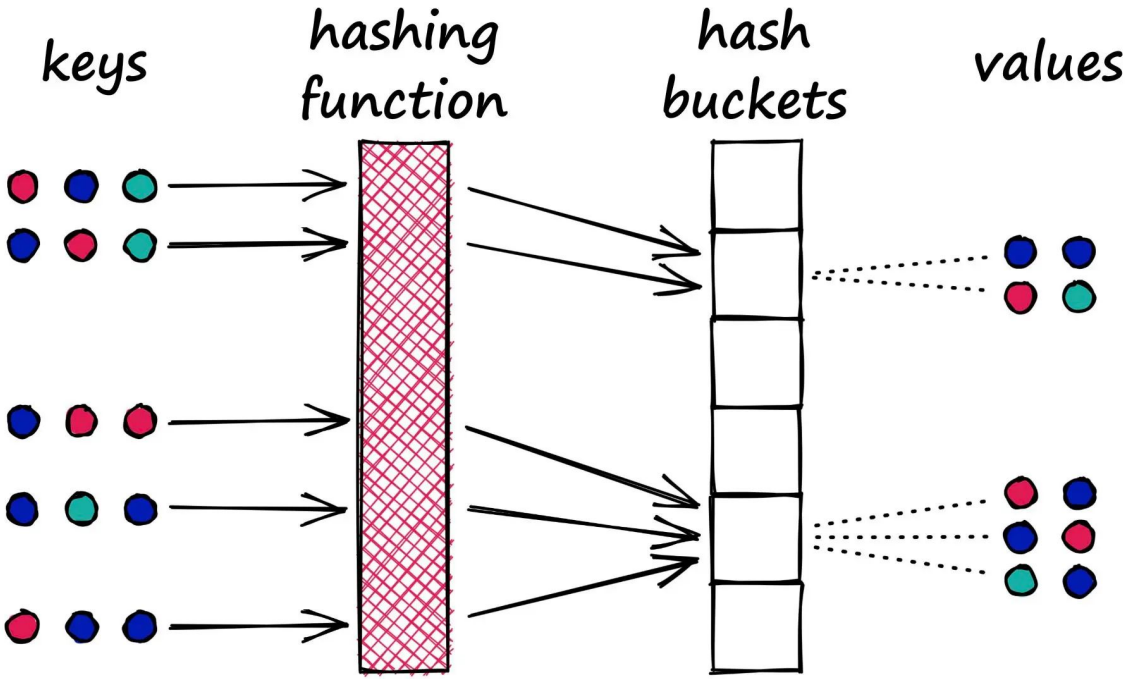
Pre-pro Goals

1. Make the data cleaner
2. Enrich the topological capabilities of the data
3. Increase the density of the dataset (or smallify it)
4. Add useful metadata to the webpages

Making dataset smaller

1. LSH duplicate removal
2. URL-based ranking cutoff

LSH duplicate removal



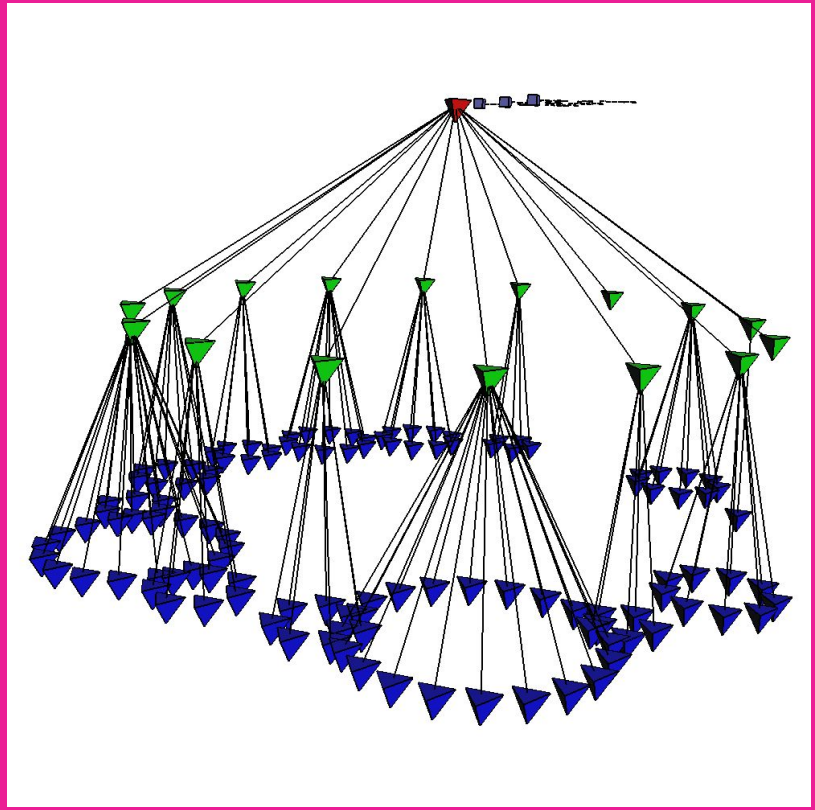
Credits: https://www.pinecone.io/learn/series/faiss/locality-sensitive-hashing

URL based filtering

An intuition: some urls provide a useful information about their importance. We filter them based on this information.

+ A big random sample of the rest (50%)
+ An heuristic based system to detect most information-rich URLs

Hierarchical feature assignment



Geospatial matching

- Every webpage contains a lot of features
  - Some are page-specific
  - Some are company wide
- Our retrieval uses geospatial information inferenced from any page all the related documents

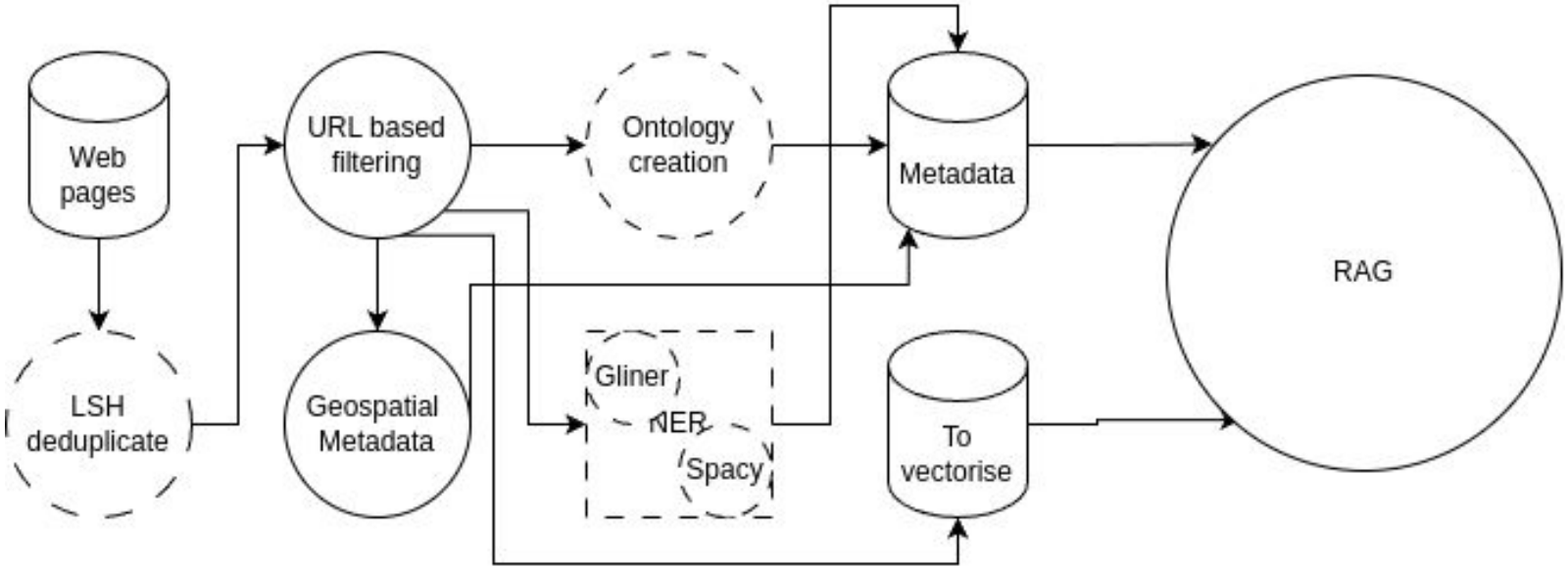
- we extract features
  - FAST : large scale updating dataset
  - ACCURATE : reliable answers
  - SPECIFIC : fine grained details
  - INHERITING: exploiting info from neighboring documents
  - EXTENDING: adding geographical and population features from external datasets

NER extraction

- … geospatial is not everything, the same concept can be extended to a wide amount of entities
- GLiner — open set of entities mined from text (even from the given ontology!)
- Spacy – reliable and well-curated set of entities

spaCy

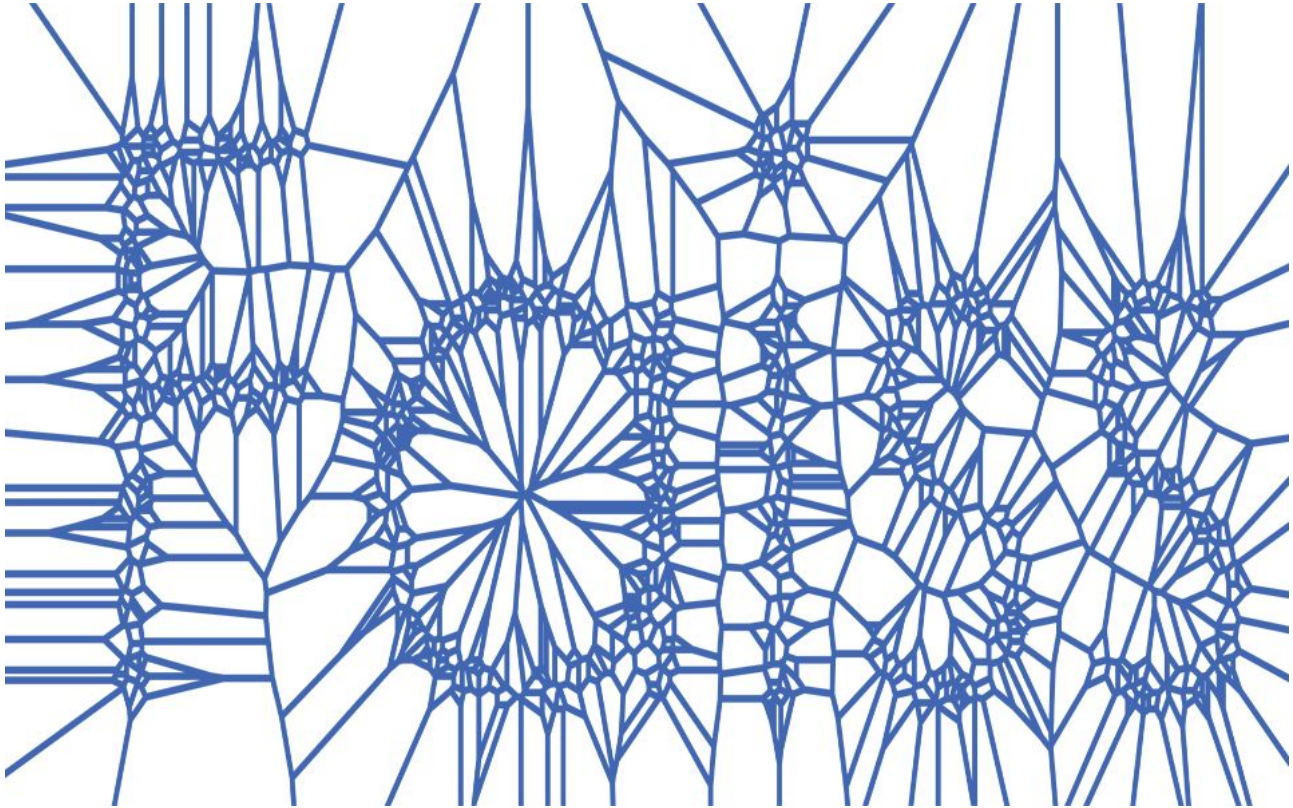
Preprocessing pipeline



RAG

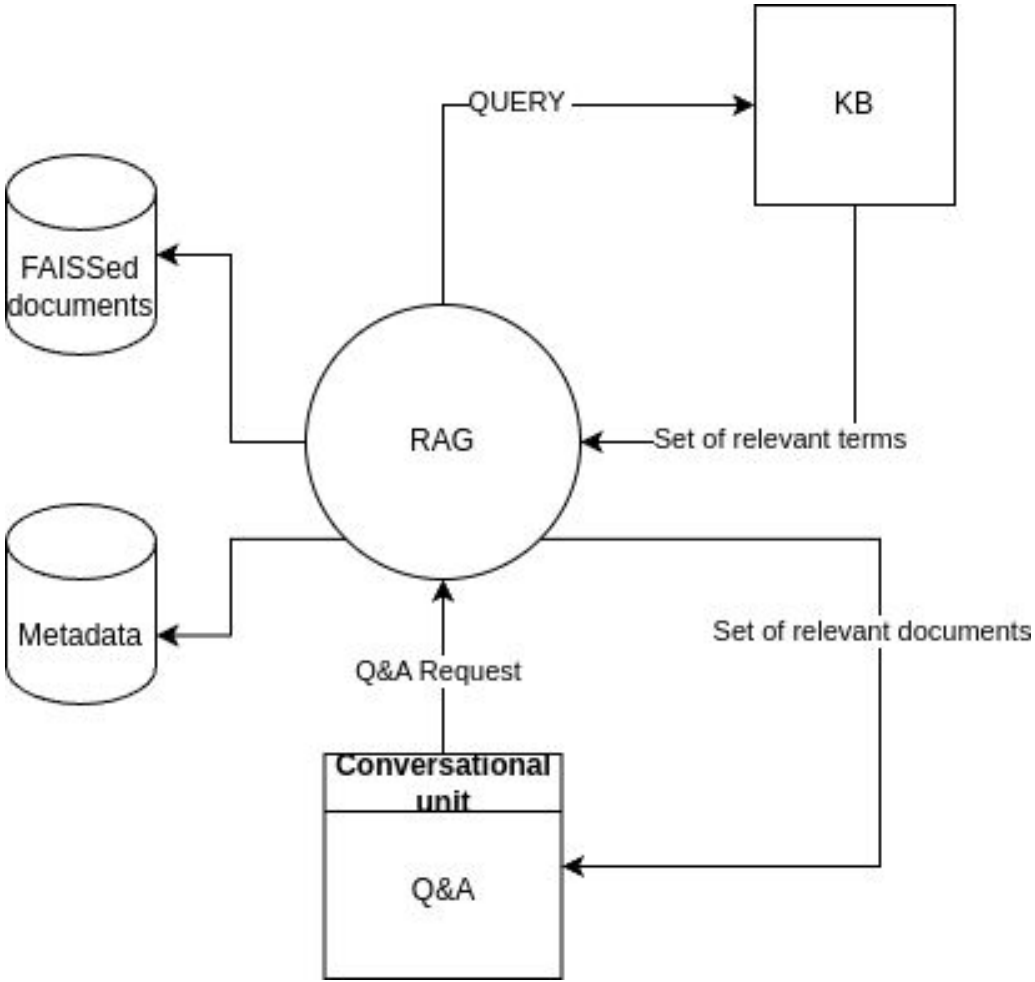
Vectorisation

- FAISS based vectorisation of the sentence-chunked documents
- Metadata information as a separate datas unit



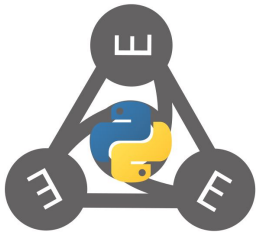
RAG

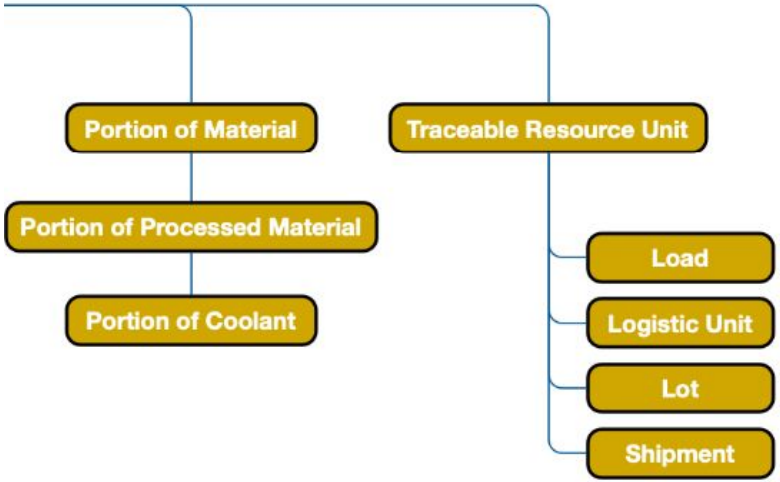
1. KB specifies the request of user
2. RAG queries both databases
3. A set of relevant documents with a combined scoring is retrieved



Knowledge base source

IOF Supply Chain ontology is loaded with an owlready2 library





In-house ontology

Creating a KB by the extracted features with the YARRRML and Morph-KGC

morph

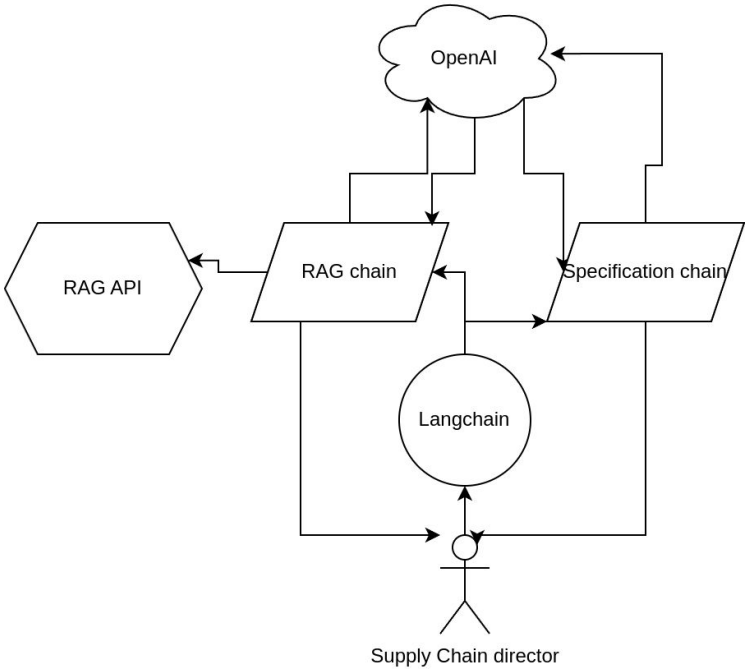
Conversational system

Conversational unit

LangChain

Multi-chain Q&A is used:

- Specification chain is responsible for making additional user questions
- RAG chain is responsible for the retrieval process



The end