



INFORMATION RETRIEVAL

Techathon Data Science & AI

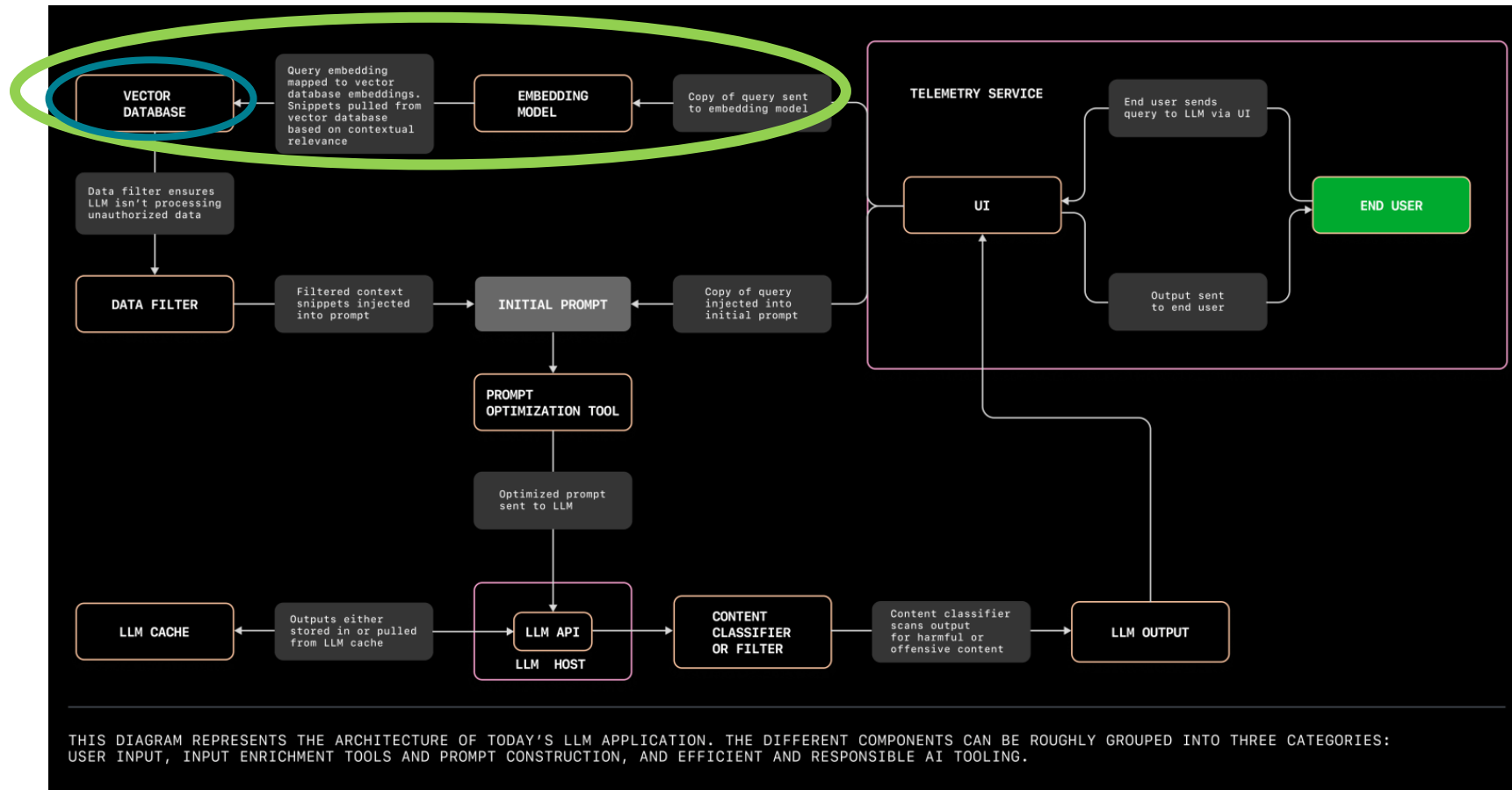
25/01/2024

Ordina-Group/hackathon-vectorsearch-
benchmark: techathon Data Science & AI
about vector search, embedding models
and benchmarks (github.com)

By Pauline van Nies

INFORMATION RETRIEVAL

Finding documents of text that satisfies the information need from within a large collection.



GOAL OF THIS TECHATHON

Have a good understanding of the components and concepts



Embedding models
Data to embed



Information retrieval
benchmarks



Vector database
Search settings


How?

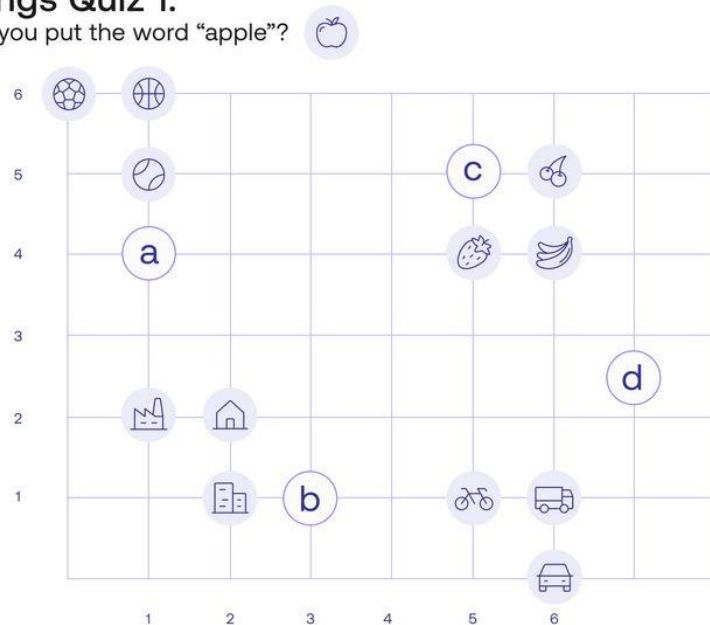
- Working in groups investigating questions and being satisfied with the answers
- Check repo and run evaluation on benchmark dataset with model in weaviate

EMBEDDING MODEL

Text -> vector representation

Embeddings Quiz 1:

Where would you put the word "apple"? 



Questions

- What is an embedding model?
- How is an embedding model trained?
- On what properties do the embedding models differ?
- Which embedding model would you like to test and why?

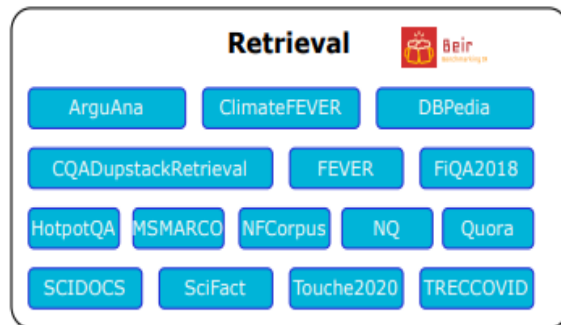
Resources

The internet

[Massive Text Embedding Benchmark Leaderboard](#)

COMPARISON OF EMBEDDING MODELS

Tasks, datasets, evaluation metrics , BEIR and MTEB



Questions

- On which tasks are embedding models compared?
- What is the difference between reranking and retrieval tasks?
- What are the different datasets used for the retrieval tasks, its different properties?
- What is the format of the datasets? Understand corpus, queries, qrels.
- What are the evaluation metrics used for information retrieval in BEIR and in MTEB?
- Is there anything you find in the evaluation approach that stands out?
- Which dataset is interesting for your use case and why?

Resources

[Massive Text Embedding Benchmark Leaderboard](#)

[BEIR colab for exploration and evaluation](#)

[BEIR Paper](#)

[MTEB Paper](#)

VECTOR DATABASE

Weaviate



Questions

- What are alternatives to Weaviate and what are pros and cons?
- What does Weaviate embed in the vector?
- What can be changed in preprocessing and for different languages?
- What does hybrid search do?
- Which parameters are adjustable for hybrid search?
- What is the default distance metric for vector search?
- What are different ranking methods and how do they influence the vector search?
- Which parameters should be chosen in weaviate to reproduce a benchmark evaluation?

Resources

- [Weaviate documentation](#)
 - [Hybrid search](#)
 - [text2vec-transformers](#)
 - [Collection schema](#)
 - [Getting started with Weaviate Python Library](#)