ORDINA
a Sopra Steria company

# INFORMATION RETRIEVAL

## Techathon Data Science & AI

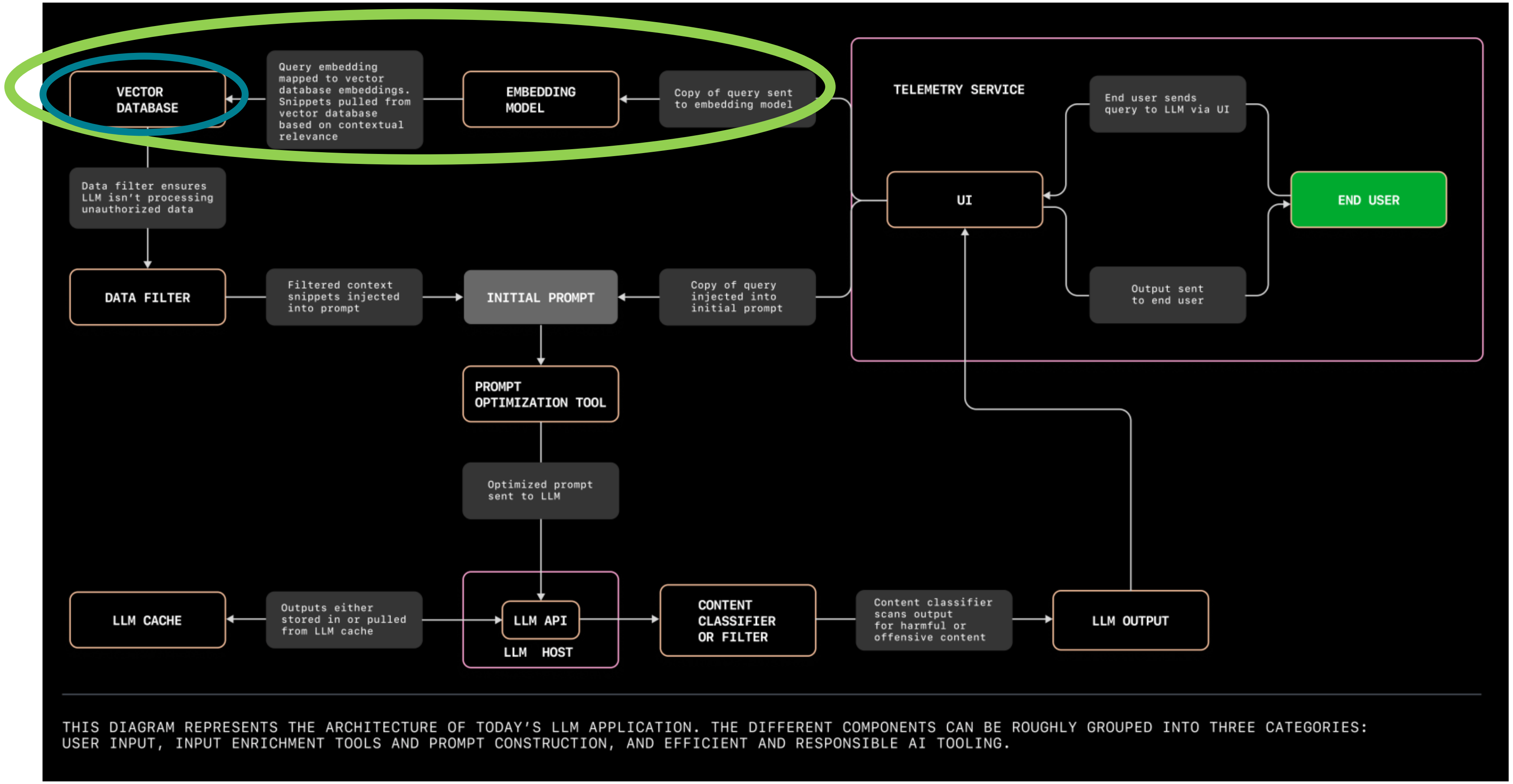
25/01/2024

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.



VECTOR DATABASE

Query embedding mapped to vector database embeddings. Snippets pulled from vector database based on contextual relevance

EMBEDDING MODEL

Copy of query sent to embedding model

TELEMETRY SERVICE

End user sends query to LLM via UI

UI

END USER

Output sent to end user

Data filter ensures LLM isn't processing unauthorized data

DATA FILTER

Filtered context snippets injected into prompt

INITIAL PROMPT

Copy of query injected into initial prompt

PROMPT OPTIMIZATION TOOL

Optimized prompt sent to LLM

LLM CACHE

Outputs either stored in or pulled from LLM cache

LLM API

LLM HOST

CONTENT CLASSIFIER OR FILTER

Content classifier scans output for harmful or offensive content

LLM OUTPUT

THIS DIAGRAM REPRESENTS THE ARCHITECTURE OF TODAY'S LLM APPLICATION. THE DIFFERENT COMPONENTS CAN BE ROUGHLY GROUPED INTO THREE CATEGORIES: USER INPUT, INPUT ENRICHMENT TOOLS AND PROMPT CONSTRUCTION, AND EFFICIENT AND RESPONSIBLE AI TOOLING.

Bigger picture : The architecture of today's LLM applications - The GitHub Blog

# GOAL OF THIS TECHATHON

Have a good understanding of the components and concepts

Embedding models
Data to embed

Information retrieval
benchmarks

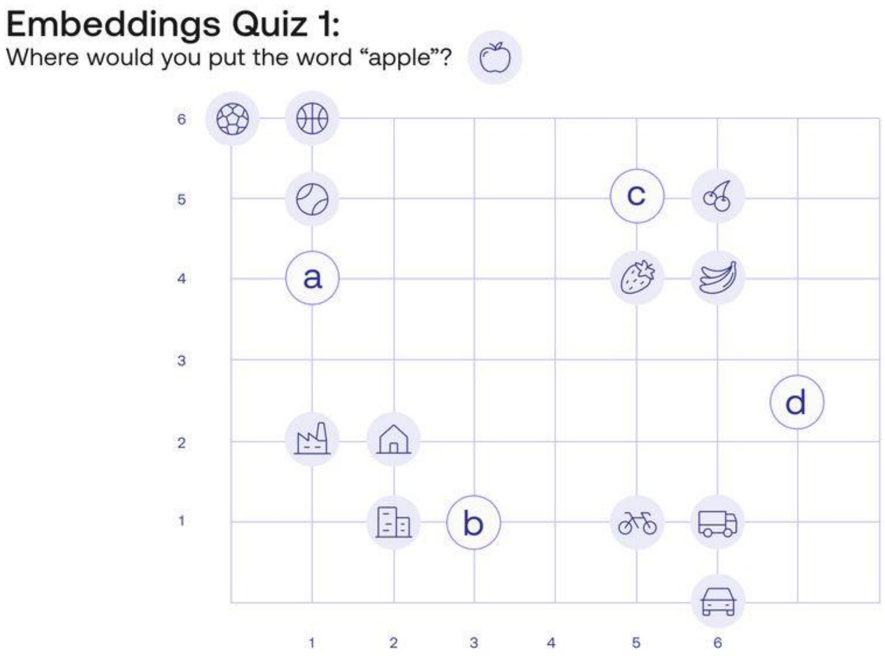
Vector database
Search settings

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?

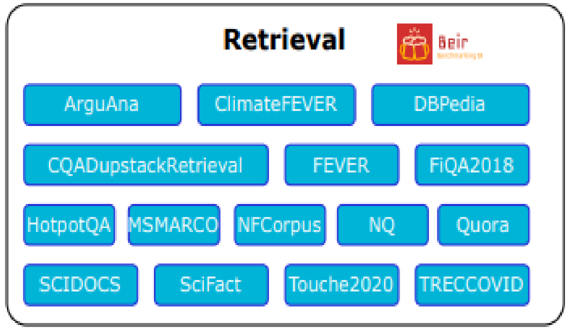
### Resources

The internet

Massive Text Embedding Benchmark Leaderboard

# COMPARISON OF EMBEDDING MODELS

## Tasks, datasets, evaluation metrics , BEIR and MTEB



Retrieval

ArguAna, ClimateFEVER, DBPedia

CQADupstackRetrieval, FEVER, FiQA2018

HotpotQA, MSMARCO, NFCorpus, NQ, Quora

SCIDOCS, SciFact, Touche2020, TRECCOVID

### 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?

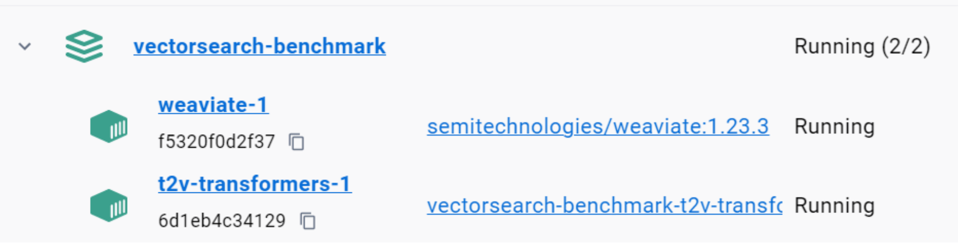
### Resources

- Weaviate documentation
    - Hybrid search
    - text2vec-transformers
    - Collection schema
    - Getting started with Weaviate Python Library

# PERFORMING EVALUATION

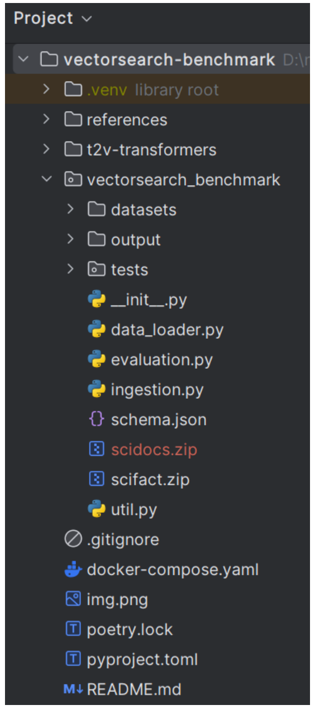
## Techathon repo

hackathon-vectorsearch-benchmark

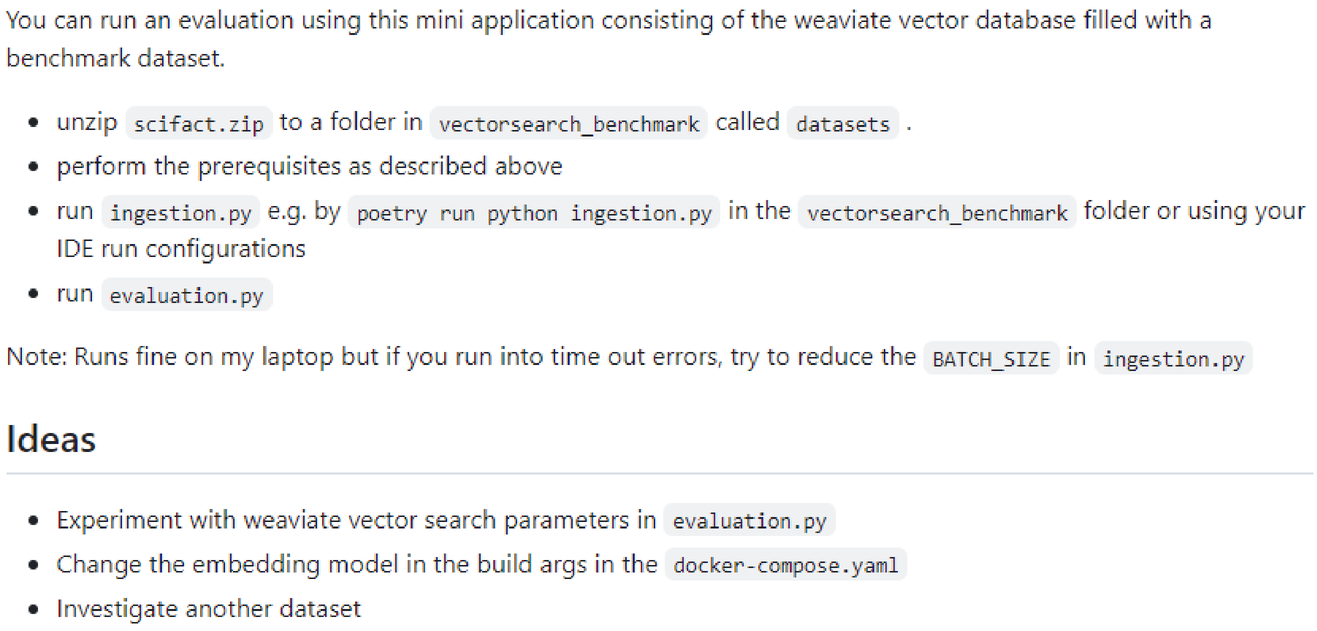


vectorsearch-benchmark Running (2/2)

| | | |
|---|---|---|
| weaviate-1<br>f5320f0d2f37 | semitechnologies/weaviate:1.23.3 | Running |
| t2v-transformers-1<br>6d1eb4c34129 | vectorsearch-benchmark-t2v-transfc | Running |



```
Project
vectorsearch-benchmark D:\
  .venv library root
  references
  t2v-transformers
  vectorsearch_benchmark
    datasets
    output
    tests
    __init__.py
    data_loader.py
    evaluation.py
    ingestion.py
    schema.json
    scidocs.zip
    scifact.zip
    util.py
  .gitignore
  docker-compose.yaml
  img.png
  poetry.lock
  pyproject.toml
  README.md
```



You can run an evaluation using this mini application consisting of the weaviate vector database filled with a benchmark dataset.

- unzip `scifact.zip` to a folder in `vectorsearch_benchmark` called `datasets` .
- perform the prerequisites as described above
- run `ingestion.py` e.g. by `poetry run python ingestion.py` in the `vectorsearch_benchmark` folder or using your IDE run configurations
- run `evaluation.py`

Note: Runs fine on my laptop but if you run into time out errors, try to reduce the `BATCH_SIZE` in `ingestion.py`

## Ideas

- Experiment with weaviate vector search parameters in `evaluation.py`
- Change the embedding model in the build args in the `docker-compose.yaml`
- Investigate another dataset

### Questions

- How does the evaluation metric compare to the benchmarks?
- If there is a discrepancy, do you have an idea where it might come from?