



INFORMATION RETRIEVAL

Techathon Data Science & AI

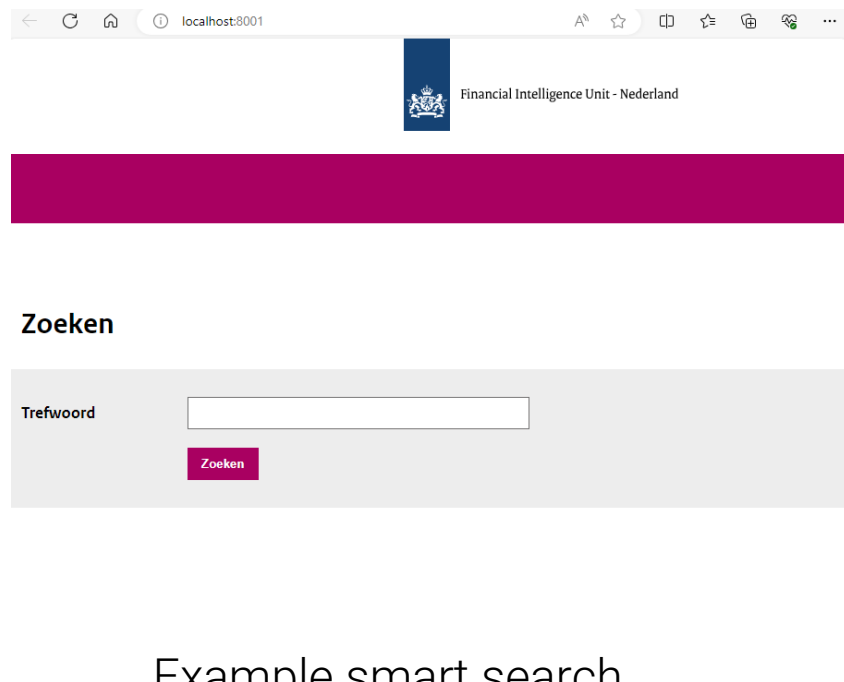
25/01/2024

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.



Example smart search

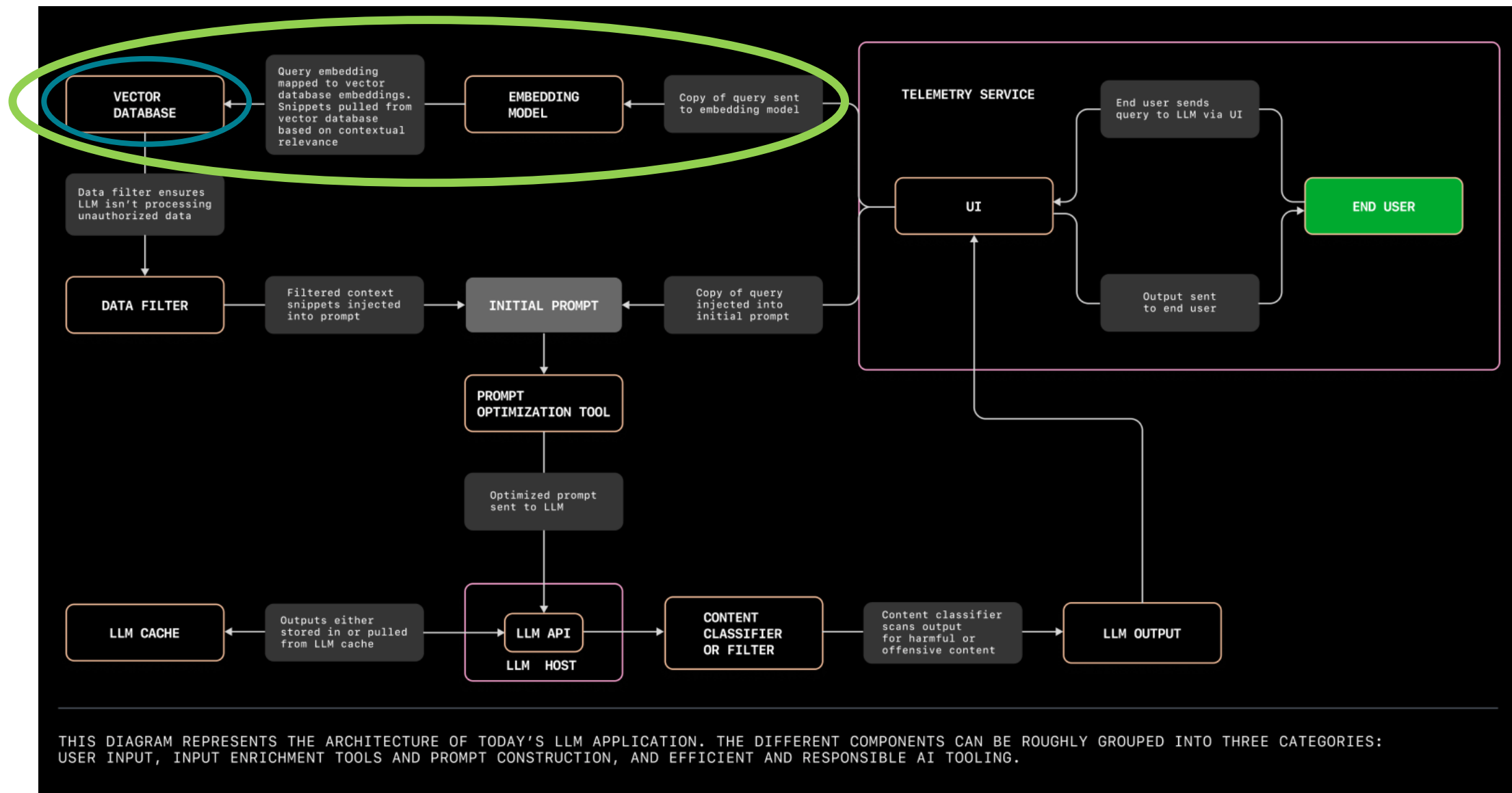


> wanneer maakt een accountant een melding van een ongebruikelijke transactie?

Een accountant maakt melding van een ongebruikelijke transactie zodra hij de transactie als ongebruikelijk kan aangemerken. Het is verplicht om een melding te doen, zelfs wanneer er geen concrete aanwijzingen bestaan dat sprake is van witwassen of het financieren van terrorisme. Een meldingsplichtige instelling moet de transactie melden als er vermoedens bestaan dat deze verband houdt met witwassen of financiering van terrorisme, zelfs indien het vermoeden alleen is gebaseerd op een ongebruikelijke transactie. Na een melding van een ongebruikelijke transactie bepaalt de FIU of de transactie als verdachte transactie moet worden aangemerkt. Hoe en wanneer maakt een accountant een melding van een ongebruikelijke transactie? Een accountant maakt een melding van een ongebruikelijke transactie zodra hij de transactie als on

[Leidraad Wwft Accountants en belastingadviseurs](#)

kwam naar voren (i) dat een bedrag van ruim € 350.000 via rekening -courant werd onttrokken, en (ii) er was sprake van onduidelijke geldleningen en daarmee samenhangende geldstromen waarvoor geen verklaring was. Ook is er een FIOD -inval geweest bij de cliënt waarbij een groot bedrag in contanten is aangetroffen, waarvan de accountant op de hoogte was. Pas na onderzoek van het BFT en ontvangst van het conceptrapport van het BFT gaat de accountant over tot het melden van zestien ongebruikelijke transacties. De Accountantskamer legt een berisping op en de accountant gaat in hoger beroep. Het CB b bevestigt de conclusie van het BFT dat, in onderling verband en samenhang bezien, sprake is van één of meer ongebruikelijke transacties. Het CB b oordeelt verder onder meer: "De professionele oordeelsvorming van de accountant in het kader van artikel 16, eerste lid, van de Wwft is gericht op de vraag of een transactie een ongebruikelijk karakter heeft of niet.



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?

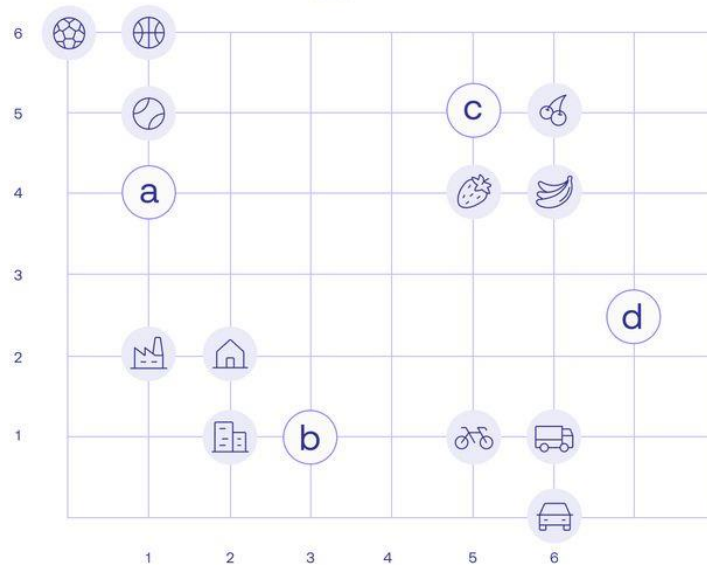
- Working in groups investigating questions and being satisfied with the answers
- Check repo and run evaluation on benchmark dataset with model in weaviate

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

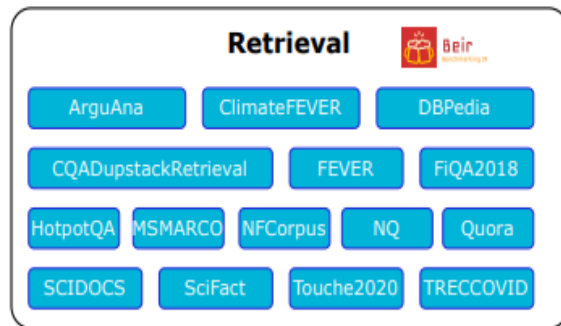
The internet

[Massive Text Embedding Benchmark Leaderboard](#)

Eveline, Gijs, Paul, Thomas

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](#)

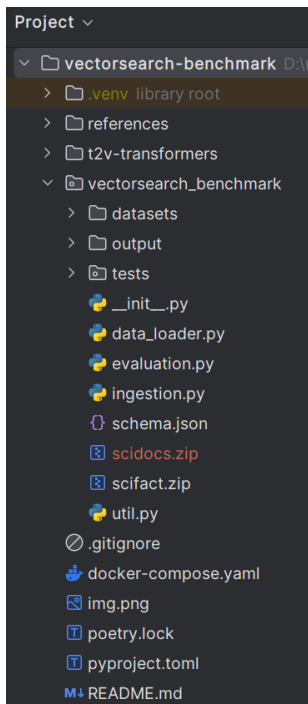
PERFORMING EVALUATION

Techathon repo

hackathon-vectorsearch-benchmark



vectorsearch-benchmark	Running (2/2)
weaviate-1 f5320f0d2f37	semitechnologies/weaviate:1.23.3 Running
t2v-transformers-1 6d1eb4c34129	vectorsearch-benchmark-t2v-transf Running



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?