Vector Store Options for <u>RAG</u>

1] FAISS
- simple, lightweight

→ similarity search over embeddings

2] Chroma
- open source vector DB

3] Qdrant
- open source rust-based vector DB
- docker deployment

→ persistance & metadata filtering

sentence-transformers

↳ HuggingFace Transformers

↳ embedding models optimized for

i) semantic similarity
ii) clustering
iii) retrieval

Ollama

↳ framework for running full LLMs locally

↳ generally optimized for <u>generation</u>

(FAISS) Similarity i.e. Distance Metrics

- ↳ L2 Distance
- ↳ Inner Product
- ↳ Cosine

Similarity Search Methods (Indexing)

- ↳ Exact Search [uses Euclidean distance]
- ↳ Approx. Nearest Neighbor
  - ↳ Quantization based
  - ↳ Graph-based
- ↳ GPU accelerated Indexes