Embeddings and Llama.cpp

SQLite VSS - Lightweight Vector DB

SQLite VSS is a SQLite extension that adds vector search on top of SQLite. It's based on FAISS[1]

There are some examples of how to use Pure SQLite VSS on the blog post here

LangChain

You can use SQLite VSS with Langchain which makes it easier to use. The documentation is here for sqlite-vss and here for using llama for embedding.

You need to install `sqlite-vss` python package to use it via `pip install sqlite-vss`

Zephyr embeddings

Load the zephyr model with long contet and set gpu layers up.

```
llama = LlamaCppEmbeddings(model_path="/path/to/models/zephyr-7b-alpha.Q5_K_M.gguf",
            n_batch=512,
            verbose=True, # Verbose is required to pass to the callback manager
            n_ctx=16000,
            n_gpu_layers=32)
```

NB: I found that Zephyr isn't actually very good for generating embeddings - I suppose this is likely because it is fine-tuned for chatting rather than for embedding.

It actually turns out that the default MiniLM that comes with sentence-transformers does a pretty reasonable job:

```
embedding_function = SentenceTransformerEmbeddings(model_name="all-MiniLM-L6-v2")
```

Revision #3
Created 2023-10-15 19:00:21 UTC by James
Updated 2024-01-21 14:49:29 UTC by James