Database Creation and Management: I start by setting up a SQLite database, defining schemas, and inserting a small subset of data. This involves connecting to the database, executing SQL commands to create tables, and inserting data into these tables. The goal is to organize my data in a way that supports efficient storage, retrieval, and querying of textual information.

Text Processing and Embedding Generation: I use transformer-based models, specifically BERT (tried both distilibert and bert-base-uncased), to generate embeddings for text documents. These embeddings are high-dimensional vectors that capture the semantic meaning of the texts. By leveraging pre-trained models and tokenizers from the transformers library, I process the texts, convert them to embeddings, and store these embeddings for later use. This step is crucial for enabling semantic search capabilities over the text corpus.

Efficient Similarity Search: To facilitate efficient similarity search within the high-dimensional embedding space, I employ FAISS (Facebook AI Similarity Search), an open-source library optimized for searching similarities in large-scale datasets. I create a quantized index using FAISS to store the embeddings, which allows for fast retrieval of documents that are semantically similar to a query. This involves training a FAISS index with the generated embeddings, adding the embeddings to the index, and then performing search operations to find the most similar documents based on the query embeddings. I return the indices of the most similar documents. However most of these indices do not correspond to remotely similar documents at all when I check the text of the documents they refer to. I don't understand why the search won't produce relevant documents. Some considerations to note are that each document represents a newspaper article and some are very long and some are just a sentence. I possibly think maybe because of the nature of the documents the encodings might be bad, but also it could very possibly be something else.

Here is a sample document:

BODIES OF VICTIMS REACH WASHINGTON

Those of Lansdowne, Lawrence and Sheppard Placed in Arlington Receiving Vault.
Special to The New York Times. WASHINGTON, Sept. 5. -- Three of the bodies of victims of the wreck of the Shenandoah, those of Lieut. Commander Zachery Lansdowne, who commanded the dirigible; Lieutenant John B. Lawrence and Lieutenant Edgar W. Sheppard, arrived in Washington this morning and were placed in the receiving vault at Arlington National Cemetery. The body of Lieut. Commander Louis H. Hancock, executive officer of the ship, will arrive tomorrow morning and be taken to the same vault to await completion of plans for an official burial. The funeral services will be held at 11 o'clock next Tuesday, when these officers will be buried with full military honors in the Dewey section at Arlington. The bodies will be laid side by side, not far from the grave of Lieutenant Lewis H., who was killed in the work of the ZR-2 on her trial trip. Just before she was to be sent to the United States by the British Government. Captain Evan "W". Scott, chief of the Chaplains' Corps of the Navy, will officiate at the funeral. Drawn up around the grave will be the Navy Band, three companies of bluejackets, and three companies of marines, under command of Commander John B. Rhodes of the Washington Navy Yard. One company of bluejackets will be from the Washington Navy Yard and two

companies from the Naval Air Station at Anacostla. The marines will be drawn both from the Washington Navy Yard and the Marine Barracks here. The naval escort will go directly to the grave. The procession from the vault will have no escort. The pallbearers will be as follows: For Lieut. Commander Lansdowne -- Perry I. Hall of Greenville, Ohio; Hard Knox of Long Island, Commander E. G. Allen, and Lieut. Commander W. A. Edwards of Washington, Lieut. Commander Max B. De Mott of Philadelphia and Lieut. Commander I. R. Pierce of Lakehurst. For Lieut. Commander Hancock -- Lieut. Commanders D. ... Beary, A. M. A. Mitscher, E. K. Long and F. C. Sherman, all of Washington, and Lieutenant J. C. Arnold of Lakehurst. For Lieutenant Lawrence -- Lieut. Commander D. and Lieutenants F. W. Wead, T. T. Patterson of Washington and Lieutenants T. G. W. Settle, J. V. Leahy and J, 13. Carter of Lakehurst. For Lieutenant Sheppard -- Commander E. E. Wilson, Lieut. Commander A. S. Carpenter and R. R. Bi. Parsons and Harry Gardfler of Washington and Lieutenants C. E. and T. W. Spear of Lakehurst. The Navy Department stated today that, in accordance with the expressed wish of members of their families, the bodies of the three of the killed, relatives had not been heard from yes-" terday, -ill le shipped as follows: Charles H. Broom, aviation chief machinist's mate, to Atlantic City, N. J.; James W. Cullinan, aviation pilot, to Binghamton, N. Y.; Bartholomew O'Sullivan, aviation machinist's mate, who was a native of New York City, to Lowell, Mass.