Optimizations

After completing the task, further optimizations that can be done are as follows:

1. If we need to process more than one document, we should introduce multi-processing, as the current code can process only one document at a time.

2. Different libraries can be tried for removing stop words, which may yield better embedding creation.

3. For the creation and retrieval of embeddings, the open-source Gemini model is used, but other Hugging Face pre-trained models can also be utilized and further tuned for our task.

4. I currently don't have good hardware, but we can use FAISS-GPU for faster retrieval of data from the database.

5. The index type of the FAISS database can be changed; different types like IndexFlatL2 or IndexIVFFlat can be tried.

6. Vector compression can be done by applying product quantization to reduce the memory footprint of the vectors and speed up the retrieval process.

7. Implementing pre-filtering steps to reduce the search space before querying FAISS can be beneficial; we can use metadata to narrow down the candidate vectors.