**Project Description**

**Current Goal**

Build a system that automatically answer students’ questions posted on Piazza.

**How?**

We will use GPT to answer students’ questions. However, to make GPT’s answers more accurate, we will send along “context” that can help GPT answer the questions. This context is a chunk of transcription of Prof. Kaiser’s lecture.

**Inner Working of the Current System**

**A picture containing text, screenshot, diagram, font

Description automatically generated**

Data preparation

- Transcription

1. Transcribe Prof. Kaiser’s lectures using Whisper from OpenAI.

Result = a lot of text files

- Prepare a database of vectors (vectors = embeddings of text chunks)

1. Split all of the text into chunks
2. Embed all those chunks using SBERT (in the code, you will see HuggingFaceEmbeddings, but it is just a wrapper of SBERT) into vectors
3. Save those vectors in the Chroma database

Real-time response

1. Fetch a student’s question from Piazza
2. Embed that question text using SBERT into a vector (called vector Q)
3. Perform similarity search -> find k nearest neighbors to vector Q (by default k = 4). Neighbors = Vectors saved in Chroma
4. Send that question, along with the text associated with the closest vector (a context), to GPT
5. Post GPT’s answer back to Piazza

**Current Problems**

1. Sometimes, the context vector that has the answer isn't nearest to vector Q, e.g. answer is in the second nearest vector. -> we can try fixing this by sending more than one context vector to GPT.
2. It is hard to know how we should split the transcriptions (what chunk size/overlap to use). -> we can split texts with different parameters, embed and store all of them in the database. (Certain sentences might appear in multiple vectors.)

**Recommended Readings**

Transcription - https://github.com/openai/whisper/blob/main/whisper/transcribe.py

Question Answering - https://python.langchain.com/docs/use\_cases/question\_answering/how\_to/vector\_db\_qa

**Comments**

Current code is using VectorDBQAWithSourcesChain, but it is deprecated. Please change it to RetrievalQAWithSourcesChain.