DSCI-560 Assignment No. 6 - Part 1

Team Formation

Team name: HealthEdBot
Names & USC ID
Cici Chang (8706354957)
Daoyang Li (7168949829)
Yifan Yang(8386626967)
https://github.com/DSCI560

Minutes of the offline meeting on March 3

Participants: project team members

Conference topic:

Our team started the project by outlining the DSCI-560 assignment goals, emphasizing building a Q&A chatbot that leverages large language models (LLMs) and embedding models to process textbook content and installation instructions. This allowed us to clarify the goals of this project, and we felt that it was necessary to develop a project schedule, specifying milestones for completing different stages, such as data collection, preprocessing, embedding generation, and dialogue chain development. We also identified tools and libraries for the project, including Python for scripting, OpenAI's Embeddings API for generating embeddings, and various text extraction and processing libraries.

This is a very important step

Cici Chang: Assigned the key role of extracting text from PDF documents, a fundamental step that requires precision to ensure that subsequent processing stages (such as embedding generation) are based on clean and accurate extracted text.

Daoyang Li: Responsible for processing extracted text into smaller, manageable chunks suitable for embedding using technologies such as CharacterTextSplitter. This step is crucial for the efficiency of the embedding process and the relevance of the generated responses.

Yang Yifan: Responsible for developing dialogue chains, which involves creating a model that can interpret user queries, retrieve relevant information from embedded text blocks, and generate coherent responses, forming the interactive core of the chatbot.

After clarifying the tasks

Start assigned tasks immediately to maintain project timelines.

Set up a GitHub repository for collaborative code development, version control, and progress tracking.

Implement regular progress checks to ensure consistency, facilitate troubleshooting, and maintain motivation.

Minutes of the online meeting on March 4

Participants: project team members

Conference topic:

I think our cooperation is smooth. Focusing on the initial hurdles encountered in text extraction accuracy and embedded text chunking optimization.

Solutions to improve text extraction accuracy, such as exploring alternative libraries and refining extraction parameters, are discussed. Highlighting the importance of block size in data processing and embedding generation, Daoyang explores ways to optimize it for better performance. This is a difficult place and we need to spend a lot of time to solve it. We encourage each other to do a good job in this project. Our concept is to adhere to a problem-solving approach to perform tasks and explore optimization strategies for chunking and embedding processes.

Minutes of the offline meeting on March 5

Participants: project team members

Conference topic:

Rather than rushing to the next step, we took some time to review progress in text extraction and chunking, acknowledging the iterative improvements made to address previously identified challenges.

Strategies for embedding blocks of text efficiently and effectively are discussed. The quality of the embedding is critical to the chatbot's ability to generate relevant and coherent responses. Yifan came up with a draft design for the conversation chain, incorporating feedback from the team to ensure the system was responsive and user-friendly. daoyang believes in integrating the various components (text extraction, chunking and embedding) for preliminary testing of dialogue chains. cici develops detailed testing strategies, including unit testing, integration testing and user acceptance testing, to ensure system reliability and availability.

Minutes of the online meeting on March 6

Participants: project team members

Conference topic:

Our project is soon to be completed, and until it is done we feel the focus of the meeting will be to finalize the embedding process and ensure a high quality of embedding to facilitate conversational chains that generate accurate and relevant responses. Yifan’ s update to Dialogue Chain highlights its development progress, demonstrating its ability to generate coherent responses based on embedded blocks of text. The accuracy of the dialogue chain and the user interaction flow are the focus of discussion. Cici and Daoyang discussed the design and development of user interfaces that are intuitive and attractive to end users, emphasizing the importance of a seamless user experience. We agreed that after completing the system integration we would begin full testing to identify any errors or areas for improvement. We also feel at the end of the development of user documentation and extensive project documentation to facilitate understanding and future maintenance of the project. We believe that this lab6 is crucial for the final project.

Minutes of the online meeting on March 7
Participants: project team members
Conference topic:
On the last day, we detected previous possibilities and errors, we carefully reviewed the dialogue chain development and planned the implementation of the driver functionality. We started with the idea of providing a clear UI for user interaction and stress-testing the reliability of the conversation chain. The need for a thorough testing phase before final submission was discussed. At the same time, we have been taking care to maintain our project and ensure that all contributions are pushed to the GitHub repository regularly. Cici concluded by paying close attention to project requirements and making sure all features are tested thoroughly. Daoyang believes that communication is the key. Keep the team updated on progress and any issues encountered. yifanfeel aims to provide clear, maintainable code and proper documentation for easy understanding and future modification.

Finally we recorded a video to show our work.

Team name: HealthEdBot
Cici Chang (8706354957)
Daoyang Li (7168949829)
Yifan Yang(8386626967)