Context Refining for Retrieval Augmented Generation

Weekly Report Feb 20

Juan Yu

I. Last week progress

I mostly focused on the data for the project during the last week.

- ✓ **Data Selection:** based on the description of the triviaQA dataset and my personal interest in business-wide application, I decided to use a subset with Wikipedia articles as evidence documents from the RC version (filtered).
- ✓ **Data Collection:** downloaded from Hugging Face
- ✓ **Data Preprocessing:** dropped unnecessary columns; tokenized and cleaned the text within the dataset
- ✓ **EDA and Data Visualization**: looked into the data volumes of questions and evidence documents.
- ✓ **Challenges**: long running time due to the large data volume and limited computing resources

II. Next week to-do list

- ➢ Implementation of TF-IDF for context refining