Emplay Assignment - Demo Script Localization

Report

Thought Process:

When approaching the Emplay Assignment on Demo Script Localization, the following thought process was followed:

- **Understanding the Problem:** The first step was to carefully read and understand the assignment instructions. This involved comprehending the objectives, requirements, and constraints of the assignment.
- **Data and Preprocessing:** The demo script document provided in DOCX format was to be localized. This required extracting text from the document, segmenting it into chunks, and identifying names that needed localization.
- **Translation and Localization:** To localize the document, an OpenAI language model was used to translate the text into the user-specified language. Additionally, the provided names in the document were to be replaced with the user's input names.
- **User Interface:** A Streamlit UI application was developed to provide an easy and interactive way for users to upload documents, select target language, and provide their names for localization.
- **Deployment:** ngrok was utilized to create a tunnel and expose the Streamlit app to a public URL. This allowed for accessing the app remotely.

Experiments Conducted:

Several experiments were conducted to ensure the correctness and effectiveness of the pipeline:

- **Pipeline Testing:** Different sample demo scripts were used for testing the entire localization pipeline. This included documents of varying lengths and structures.
- **Name Replacement:** The process of replacing names separately and together was tested to verify accurate name localization.
- **Translation Quality:** The quality of translation using OpenAI's language model was tested by translating text to different languages and evaluating the output.

Conclusions:

Based on the experiments and development process, the following conclusions were drawn:

- **Localization Success:** The pipeline successfully localized demo scripts by translating the text and replacing names as required.

- **User Interaction:** The Streamlit UI app provided an intuitive interface for users to upload documents, choose language, and input their names.
- **Translation Quality:** The quality of translations varied based on the target language. Some languages produced better results than others, highlighting the importance of evaluating translation quality.
- **Error Handling:** The pipeline included basic error handling mechanisms to handle cases such as missing input and connectivity issues.

Error Analysis:

During the development and testing of the pipeline, a few errors and challenges were encountered:

- **ngrok Connection Issues:** There were instances when the ngrok tunnel failed to establish a connection, leading to a "Bad Gateway" error. This might be due to connectivity issues or conflicts with the chosen port.
- **Translation Accuracy:** The accuracy of translations depended on the complexity of the text and the target language. Some translations were not perfectly accurate, and fine-tuning the translation parameters might improve results.
- **Token Limit:** The token limit of the OpenAI model limited the length of text that could be processed in a single API call. Longer documents required segmentation and handling multiple API calls.
- **User Experience:** The Streamlit app offered a straightforward user interface, but further enhancements could be made to provide better user guidance and error messages.