Brance <Founding AI Applied Researcher> Task

Name: Soura Dutta

Linkedin Profile: <https://www.linkedin.com/in/soura-dutta/>

Date Challenge Received: 6th July, 2023

Date Solution Delivered: 6th July, 2023

1. Problem Statement

Task is to create a chatbot using Retrieval Augmentation Generation (RAG) on a given knowledge document. It receives question from users and provides answers on the knowledge found in the document. It has two phases, fetching relevant contexts from the document and answering using LLM from the contexts.

2. Approach

1. Try QA with knowledge source chain first with custom prompt to prevent hallucination.
2. Fall back to Zero Shot Learning agent with Google search tool.
3. Use persistent DB for knowledge source embeddings and memory for conversation.
4. *Bonus feature* - Added speech recognition feature for input and speak aloud feature for output.

3. Solution

**Procedure**

* Read file -> Tokenize text -> Create Embeddings -> Persist embeddings in DB
* Create prompt to stop hallucination -> Create shared persistent memory -> Create QA with sources chain (1) -> Create zero shot agent chain with Google search tool (2)
* Create infinite loop -> Take speech input -> Transcribe -> Try question as input to (1) -> If fails send question as input to (2) -> Speak out output loud -> Break loop if asked to exit

**Performance**

* Working well on in-context data. No signs of hallucination. Falls back to ZSL agent with google search and finds answer accordingly.

4. Future Scope

1. Google search results can be used to enrich knowledge base.
2. Knowledge base can be stored in unstructured DB like – MongoDB.
3. Add relevant chain to add multi-lingual conversation support.