Data Collection/ Processing/ Storage

Required package: langchain 0.0.319 PyPDF 3.0.1

Workflow

To run the chatbox, please enter the code below:

```
python3 chatbox.py
```

When chatbox is running, it will read the pdf(Ads cookbook.pdf) and convert it into vector database. It include four step:

1. PDF extraction

In this function, PyPDF2 is applied for extract data from pdf and stored them as list of string. Now, the data in pdf are massive and chaos.

2. Text chunks

Each chunk has size about 500. CharacterTextSplitter is mainly used in this part.

3. Vectorizing

Becasue the RateLimitError caused by openai, we used allminiLM as embedding tool. After this part we get vector store

4. Conversation Chain

Then we can input the vector store and create a converstaion chain. This step combine the ConversationBufferMemory and ConversationalRetrievalChain

Now the everything is prepared, you can ask any question you want. Then AI will answer you and print it in the screen. And type exit in question to end the script.

```
vboxuser@dsci-560-back:~/Desktop/test folder$ python3 chatbox.py
PDF Extraction Completed
Text Chunk Completed
Vectorizing Completed
Question: Hello
AI:  Hello! How can I assist you today?
Question: exit
Exiting the program. Goodbye!
vboxuser@dsci-560-back:~/Desktop/test folder$
```