Intel Vtune Profiler Assistant-GenAI powered chatbot.

**Team members:**

Jaideep, Shanmukh, Aishwarya CV, Shivani

**Problem Statement:**

Develop a chatbot powered by GenAI to provide targeted assistance to developers seeking information about Intel Vtune Profiler and its components. The chatbot should specialize in addressing queries specifically related to Intel technologies and their application within the oneAPI ecosystem.

**Solution approach:**

Converting the input pdf file to text in document

Pipeline creation

Input encoding

Model load

Data Retrieval(RAG)

Output

**Technologies used:**

1. Python standard library
2. Third party libraries

* Chromadb
* PyPDF2
* Transformers

1. Langchain

* CharacterTextSplitter
* RecursiveCharacterTextSplitter
* HuggingFacePipeline
* HuggingFaceEmbeddings
* RetrievalQA
* Document
* Chroma

1. Transformers

* AutoTokenizer
* AutoModelForQuestionAnswering
* AutoModelForCausalLM
* pipeline

**Implementation:**

The input PDF file is converted into text within a document using the PyPDF2 library. Following this conversion, the text is further processed by employing the RecursiveCharacterTextSplitter to split it into smaller segments or chunks based on specific criteria.

Text undergoes encoding using Hugging Face embeddings, leveraging the capabilities of the " hkunlp/instructor-large " model. This process involves transforming the text into a numerical representation that captures the semantic meaning and context of the words and phrases.

The text is stored into a database utilizing Chroma. This involves saving the processed text data into a structured storage system provided by Chroma, which offers efficient data management and retrieval capabilities.

Loading the Hugging Face model involves initializing the specified model, denoted as "< zephyr-7b-beta or Xwin-LM-13B-V0.1 >," along with an autotokenizer. The autotokenizer automatically selects and configures the appropriate tokenizer based on the model being used, ensuring compatibility and optimal tokenization for text inputs. This combined setup enables efficient processing and utilization of the Hugging Face model for tasks such as text generation.

Data retrieval using RAG (Retrieval Augmented Generation) with the retrievalQA method involv.es a process that integrates retrieval-based and generation-based techniques in natural language processing (NLP). In this context, the retrievalQA method utilizes a retriever to fetch relevant data from a knowledge base or corpus based on a given query or context. The retriever employs keyword matching models to identify and retrieve the most relevant information.

Retrieved data is served as an input augmentation, This augmentation enhances the context and completeness of the input data by incorporating additional information from the retrieved sources. The augmented input is then passed through generative model< zephyr-7b-beta or Xwin-LM-13B-V0.1 >to generate responses, answers or summaries to the input and the specific task requirements.

**Enhancements:**

This Intel Vtune Profiler GenAI chatbot is currently designed to work with a single toolkit. We can enhance its capabilities to support multiple toolkits of Intel oneAPI in the future.

**Source Code:**

[source.ipynb](https://github.com/shanmukha-sajja/vtuneprofiler_genAI_bot/blob/main/source.ipynb)