Part 1 Documentation: Business QA Bot

Overview

This section provides documentation for a system that retrieves relevant documents based on a user query and generates answers using a generative AI model. The process involves creating an index with Pinecone, encoding queries using a sentence transformer model, and generating responses via Cohere's generative language model. Below is an explanation of each component.

Model Architecture

The system utilizes two main components for information retrieval and response generation:

1. **Sentence Transformer for Encoding**:
   - A sentence transformer from the `sentence-transformers` library is used to encode both documents and queries into dense vector representations (embeddings). The transformer converts text into 768-dimensional embeddings that capture semantic meaning.
   - The encoded embeddings allow comparison between the user's query and the stored documents by measuring the cosine similarity between the vectors.
2. **Cohere for Generative Responses**:
   - After relevant documents are retrieved, Cohere's generative model, specifically `command-xlarge-nightly`, is used to generate human-like answers based on the retrieved context.

    This model is capable of processing the input prompt (including the query and the context) and generating concise and accurate responses.
   -

Approach to Retrieval

1. **Pinecone Vector Index**:
   - Pinecone serves as the backbone for document retrieval by storing and searching through high-dimensional vectors. The documents are indexed by first converting their content into 768-dimensional vectors using the sentence transformer model. These vectors are then stored in Pinecone.
   - The cosine similarity metric is employed to find documents that are semantically close to the user's query.
2. **Document Insertion into Pinecone**:
   - When a document or a piece of information is added to the index, it is converted into a vector (using the same sentence transformer). This vector is then upserted into Pinecone under a specific namespace.
   - If the index doesn't exist, it is created with the following configuration:
     - **Dimensions**: 768 (matching the sentence transformer output).
     - **Metric**: Cosine similarity to determine the closest match.
     - **Serverless Specification**: The index runs on AWS infrastructure in the 'us-east-1' region.
3. **Retrieving Documents**:
   - When a user submits a query, it is transformed into a vector using the sentence transformer model. This query vector is then used to search the Pinecone index, retrieving the top 3 most relevant documents based on their cosine similarity to the query.

- These results contain metadata, such as the original text of the document, which is used to construct the context for the generative response.

Generative Response Creation

1. **Context Generation**:
    - After retrieving the relevant documents, the text from these documents is combined to form a "context." This context is essentially a concatenation of the text metadata from the top matching documents retrieved from Pinecone.
2. **Generating the Response**:
    - Using the combined context and the user's query, a prompt is created in the following format:
    -
    - context: {retrieved_context}Question: {user_query}Answer:
    -
    - This prompt is then fed into Cohere's large language model (`command-xlarge-nightly`), which generates a natural language answer to the query. The model processes the prompt and returns a concise response that leverages the provided context.

Example Workflow

1. **User Query**:
    - Suppose the user asks: *"Can you tell me about the 2022 Chardonnay?"*
2. **Query Encoding and Retrieval**:
    - The query is encoded into a vector using the sentence transformer model and passed to Pinecone. Pinecone searches for the top 3 matching documents based on cosine similarity.
3. **Context Creation**:

    - The retrieved documents' text is combined to form the context, such as:
    - The 2022 vintage of our Chardonnay is a true expression of the unique terroir of our vineyard. This wine showcases a brilliant straw color with hints of green, inviting you to take a sip...
    -

4. **Response Generation**:

- The system generates the following answer using Cohere: *"The 2022 Chardonnay has a brilliant straw color, fresh citrus aromas, and a balanced acidity with a creamy texture. The finish lingers with flavors of green apple and oak spice."*

Predefined Queries

The system also supports predefined queries for demonstration purposes. For example:

1. **Query**: "What white wines do you have?"
    - **Answer**: "Some popular white wine varieties include Chardonnay, Sauvignon Blanc, Pinot Grigio, Riesling, and Moscato."
2. **Query**: "Can you tell me about the 2022 Chardonnay?"
    - **Answer**: A detailed description of the 2022 Chardonnay based on the stored document.

Conclusion

This system combines semantic document retrieval with generative AI to deliver accurate, context-aware responses. By leveraging Pinecone for vector-based search and Cohere for language generation, it ensures that users receive relevant and coherent answers based on the knowledge present in the indexed documents.