SAMPLE SET

PART 1

Hursh Karnik

Introduction

# Retrieval-Augmented Generation (RAG) Model for QA Bot

In this notebook, we demonstrate how to build a Question Answering (QA) bot using a Retrieval-Augmented Generation (RAG) model. The model integrates a vector database (Pinecone) for efficient document retrieval and a generative model (Cohere API) for generating answers based on the retrieved documents. We will walk through the process of indexing documents, querying them, and generating responses.

Model Architecture

Model Architecture

The RAG model integrates two main components:

1. Retrieval:

- Pinecone is used to index and retrieve document embeddings efficiently.

- Documents are encoded into embeddings using the `SentenceTransformer` model.

- These embeddings are stored in Pinecone and retrieved based on query embeddings.

2. Generation:

- Cohere API is used to generate answers based on retrieved document excerpts.

- The retrieved document excerpts are combined with the user query to form a prompt.

- The prompt is sent to Cohere API to generate a coherent response.

The architecture ensures that the QA bot retrieves relevant information from a dataset and generates accurate answers.

Approach to Retrieval

1. Document Embeddings:

- Documents are encoded into embeddings using a pre-trained `SentenceTransformer` model.

- Embeddings represent the semantic content of the documents.

2. Indexing:

- Embeddings are stored in Pinecone with associated metadata (e.g., document text).

- Pinecone efficiently manages the retrieval of embeddings based on similarity to the query embedding.

3. Query Processing:

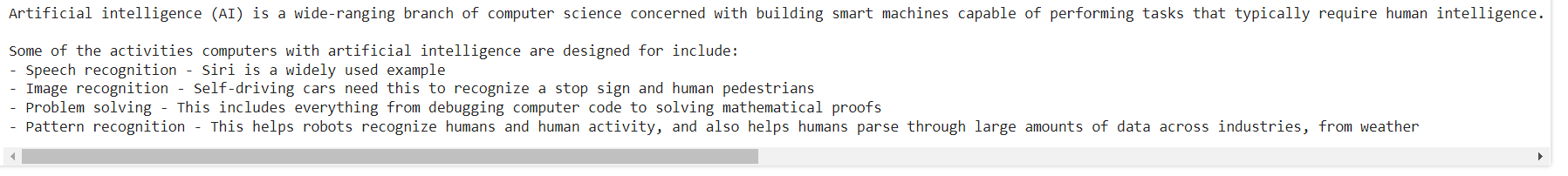
- User queries are encoded into embeddings.

- Pinecone retrieves the most relevant documents by comparing query embeddings to stored embeddings.

**IMPORTANT**

**Replace pinecone and Cohere api key with your api key**

output 1



Output 2

Input

What is html?

A screen shot of a computer

Description automatically generated

Output 3  
What is generative ai?

A screenshot of a computer program

Description automatically generated