Report: Development of a Book Recommendation Chatbot

INFO7375. Prompt Engineering & AI- Summer semester

By

Shima Bolboli

Abstract

This report outlines the implementation of a Book Recommendation Chatbot using Pinecone for vector indexing, LangChain for embedding and interaction management, and Streamlit for the web interface. The chatbot aims to provide users with personalized book suggestions based on their input queries, leveraging Natural Language Processing (NLP) models from Hugging Face to embed text data and generate responses.

Introduction

The Book Recommendation Chatbot is designed to assist users in finding books tailored to their preferences and search queries. By integrating advanced technologies such as Pinecone, LangChain, and Hugging Face models, the chatbot can handle large datasets, ensure accurate recommendations, and manage vector embeddings efficiently. Streamlit is used to create an interactive and user-friendly web interface for the chatbot.

System Architecture

Pinecone Initialization

Pinecone is a scalable vector database designed for high-performance indexing and querying of vector data. It plays a crucial role in the chatbot by storing book data and facilitating efficient search operations. Pinecone is initialized using an API key and environment settings, and an index is created if it does not already exist. This index is configured to use cosine similarity as the metric for measuring the similarity between vectors.

Data Preprocessing

The dataset used for the chatbot is preprocessed to ensure it is clean and suitable for embedding. This involves:

- Loading the dataset and inspecting its structure.
- Cleaning the data by removing rows with null values in essential columns such as titles and authors.
- Converting text fields to strings to facilitate embedding.

This preprocessing step ensures that the data fed into the embedding model is consistent and free of errors, which is crucial for accurate recommendations.

Embedding and Upserting Data

The chatbot uses a pre-trained model from Hugging Face for embedding book titles and authors into high-dimensional vectors. These embeddings capture the semantic meaning of the text, allowing for effective similarity searches. The embeddings are upserted into the Pinecone index

in batches to handle the large dataset efficiently. This batch processing approach ensures that the system can scale and manage resources effectively.

Fetching and Displaying Recommendations

When a user enters a query, the chatbot embeds the query text using the same Hugging Face model and retrieves the top-k matches from the Pinecone index. The system checks both exact matches and partial matches to ensure comprehensive search results. This involves comparing the query text with both the title and authors of the books in the database.

Streamlit Interface

Streamlit is used to create a web interface for the chatbot, providing an interactive and user-friendly platform for users to input their queries and receive recommendations. The interface maintains a history of user interactions, displaying previous queries and their corresponding results. This feature enhances the user experience by allowing users to review past recommendations and refine their search queries.

Implementation Details

1. **Pinecone Initialization**:
    - Pinecone is initialized with an API key and environment settings.
    - An index is created with specific dimensions and similarity metrics if it does not already exist.
    - The index is connected for subsequent operations.
2. **Data Preprocessing**:
    - The dataset is loaded and inspected for shape and content.
    - Rows with null values in essential columns are dropped.
    - Text fields are converted to strings to ensure compatibility with the embedding model.
3. **Embedding and Upserting Data**:
    - A Hugging Face model is used to embed book titles and authors into vectors.
    - Embeddings are processed in batches to handle large datasets efficiently.
    - The embeddings are upserted into the Pinecone index, associating each vector with metadata.
4. **Fetching and Displaying Recommendations**:
    - User queries are embedded using the same Hugging Face model.
    - The Pinecone index is queried for the top-k matches based on vector similarity.
    - Results are filtered for exact and partial matches to ensure relevant recommendations.
5. **Streamlit Interface**:
    - A web interface is created using Streamlit, allowing users to input queries and view recommendations.
    - The interface maintains a history of user interactions, displaying past queries and results.

- Recommendations are displayed with titles and authors, along with a summarization of the book details.

Conclusion

The Book Recommendation Chatbot effectively integrates multiple advanced technologies to provide personalized book recommendations. Pinecone efficiently handles vector indexing and querying, while LangChain and Hugging Face models offer powerful text embedding and generation capabilities. Streamlit provides a simple yet effective interface for user interaction, making the chatbot accessible and user-friendly.