Assignment NO 12
Programming for Artificial Intelligence
(PAI)

Name: Muhammad Jawad Ahsan

Roll No: BSAIM-F23-043

Section: 4A

Subject: Programming for AI

Project Summary

Disease and Symptoms Embedding with Sentence Transformers

Objective

The goal of this script is to preprocess a dataset of diseases and their related symptoms, generate semantic embeddings using a pre-trained SentenceTransformer model, and prepare the data for further use in an AI-driven application (e.g., similarity search or diagnosis prediction). And integration with a Flask API.

Workflow Summary

1. Data Loading and Cleaning

- File Input: DiseaseAndSymptoms.csv
- Initial Processing:
- The script reads the CSV file and retains the first 4 columns.
- It removes any unnecessary or null entries and saves the cleaned version as refine_data.csv.

2. Symptom Combination

- Creates a new column Combined_Symptoms by joining all symptom-related columns into a single string per disease.
- This string is used as input for the embedding model.

3. Text Embedding

- Uses the SentenceTransformer model: all-MiniLM-L6-v2 to encode the combined symptoms into dense vector embeddings.
- Embeddings are stored as a NumPy array and saved to disk with np.save() for future retrieval.

4. File Output

- Embeddings saved as: disease_sumptoms_embeddings.npy
- Processed CSV saved as: refine_data.csv
- Flask Integration
- The script will be integrated with a Flask web server to:
- Accept user symptom inputs via a web form or API call.

- Encode the input symptoms.
- Compare the input embeddings with precomputed disease embeddings using FAISS for similarity search.
- Return the most likely disease(s) based on symptom similarity.

Libraries Used

pandas: for data manipulation.

re: for text cleaning (planned or placeholder).

sentence-transformers: for semantic sentence embedding.

faiss: for efficient similarity search (setup included but not yet used in this script).

numpy: for array handling and saving embeddings.

Flask: for deployment as an API or interactive web interface.