Perspective-Aware Healthcare Summarization

This project focuses on generating high-quality summaries for healthcare Q&A data using multiple perspectives such as CAUSE, INFORMATION, SUGGESTION, QUESTION, and EXPERIENCE. We implement both supervised and weakly-supervised techniques to train models and generate summaries.

Project Structure

- `stacking-ensemble_nlp.ipynb`: Fine-tuning of Flan-T5 with LoRA, BART, and PEGASUS. Also includes ensemble logic to combine outputs from multiple models.
- -`snorkel_nlp.ipynb`: Weakly supervised labeling using Snorkel, classification using Sentence Transformers + Logistic Regression or SVM, and a two-stage extractive-abstractive summarization pipeline.
- `train.json`, `valid.json`, `test.json`: Dataset splits used for training, validation, and evaluation.
- `classified-test-output.json`: Output of Snorkel-based classification on the test set.

Stacking notebook:-

This notebook performs **supervised fine-tuning** for perspective-aware summarization using multiple transformer models. It:

- Fine-tunes **Flan-T5 with LoRA** on the perspective-labeled data
- Fine-tunes **BART** and **PEGASUS** using the same instruction-style prompts
- Uses a **stacked ensemble** to generate summaries from all models
- Compares outputs and selects the better one based on fluency and relevance

The goal is to generate more accurate and perspective-aligned summaries by combining the strengths of multiple summarization models.

Snorkel Notebook:-

This notebook implements a **weakly supervised pipeline** using **Snorkel** and lightweight classifiers to label answer sentences by perspective. It:

- Embeds sentences using Sentence Transformers
- Applies **Snorkel labeling functions** based on keywords

- Uses **Logistic Regression or SVM** to classify perspective labels
- Applies **zero-shot classification** when Snorkel and classifiers are unsure
- Groups sentences by perspective and summarizes them:
    - First using **BART (extractive)**
    - Then refines with **PEGASUS (abstractive)**

The goal is to generate perspective-based summaries **without relying on gold summary labels**, making it suitable for unlabeled or weakly labeled data

Plasma Model Link:-
https://drive.google.com/drive/folders/1fSkgWWQRqOLh9H4O-YfxwzM3rs7baTNH