Assignment: Medical Keyword Extraction

Background:

Keyword extraction involves identifying and extracting essential terms or phrases from text data, in this case, medical transcriptions. Extracting keywords from these transcriptions aids in summarizing the content, facilitating searchability, and enabling better information retrieval in the medical domain. Your task is to develop a keyword extractor that extracts the keywords from the transcription

For this assignment, you will utilize a big collection of transcribed medical reports for various medical specialties.

Assignment Tasks:

- Dataset Understanding: describe the structure of dataset, including the features and labels.
- Data Preprocessing:
    - o Text cleaning (removing special characters, lowercasing, tokenization).
    - o Handling missing values and duplicates.
    - o Splitting the data into training and validation sets.
- Train/Fine-tune on given domain-specific dataset.
- Incorporate a language model-internally and externally in the system architecture if possible.
- Evaluate the effectiveness of fine tuning in handling domain-specific NLP tasks compared to pre-trained (baseline) end-to-end models.
- Perform EDA on the train data and test results.

You are required to present the working code, approaches, explaining the steps followed, challenges faced, and their solutions, Finding and results.

Provide documentation outlining the code, algorithms used, and reasoning behind choices made during the process.

Dataset:

https://www.kaggle.com/datasets/tboyle10/medicaltranscriptions/data

Platform:

https://colab.google/

Time Duration:

3 days