Project 3: Online Health Science Knowledge Chatbot

Methods and Technologies for Sub-Tasks B/C, E/F, L/M

Team 3.2

March 9, 2025

Sub-Tasks B/C: Keyword Extraction & Inverted Index

Objective 1.2: Keyword Extraction and Indexing

- **Method 1: Medical Entity Recognition**
  - SpaCy NER Pipeline:
    - Pre-trained biomedical models (e.g., `en_core_sci_md`) for symptom/disease/drug extraction.
    - Custom rules via `Matcher` to handle domain-specific terms (e.g., rare diseases).
- **Method 2: Inverted Index Construction**
  - Lightweight Indexing with Whoosh:
    - Tokenization and stopword removal using `nltk`.
    - Trade-off: Efficiency for small-to-medium datasets (vs. Elasticsearch).
- **Tech Stack**:
  - Python, SpaCy, NLTK, Whoosh.

Sub-Tasks E/F: Query Classification

Objective 2.1/2.2: Topic Classification

- **Method 1: Unsupervised Clustering**
  - K-means with TF-IDF/PCA:
    - Reduce dimensionality to 50-100 features to avoid the "curse of dimensionality."
    - Validate clusters using `pyLDAvis` for human-in-the-loop refinement.
- **Method 2: Supervised Fine-Tuning**
  - DistilBERT Multi-Label Classification:
    - Freeze base layers; train only the classification head for efficiency.
    - Optimize for Macro-F1 to handle class imbalance.
- **Tech Stack**:
  - Scikit-learn, HuggingFace Transformers, PyTorch.

Sub-Tasks L/M: Next-Question Prediction

Objective 5: Context-Aware Prediction

- **Method 1: Session Chain Extraction**
  - Regex-Based Parsing:
    - Extract [Q→A→Q] chains from structured Amazon conversations.
    - Store in Neo4j for contextual graph traversal (e.g., symptom→treatment→side-effect).
- **Method 2: Generative Prediction**
  - GPT-2-small Fine-Tuning:
    - Input: [Q1, A1]; Output: Top-3 candidate Q2.
    - Beam search (beam=5) with length penalty for diverse yet relevant predictions.
- **Tech Stack**:
  - Neo4j, HuggingFace Transformers, Pandas.