SOAP Notes Generation - Project Documentation

Project: SOAP Notes Task
Author: Zeeshitha
Date: 2025

Introduction:

This document provides a detailed explanation of the process used to generate structured SOAP (Subjective, Objective, Assessment, Plan) notes from doctor-patient conversation transcripts.

The goal of this project is to automate medical documentation efficiently using gpt-4 & gpt-4o-mini and NLP techniques.

The solution integrates OpenAI's gpt-4 & gpt-4o-mini API for natural language processing and spaCy for medical entity extraction.

Additionally, chunking was implemented to optimize processing for longer transcripts.

Methodology:

This section outlines the approach taken to process the transcripts and generate SOAP notes.

Step 1: Data Preprocessing & Cleaning

- Removed unnecessary characters, numbers, and special symbols.
- Segmented the transcript by speaker (DOCTOR/PATIENT).
- Converted multiline text into a structured format for better processing.

Step 2: Chunking Implementation

- To improve model efficiency and prevent exceeding token limits, the transcript was divided into chunks of 10 speaker exchanges.
- Each chunk was processed separately, and results were merged into a final SOAP note.

Step 3: Text Classification into SOAP Sections

- Utilized spaCy NLP for medical entity recognition.
- Applied keyword-based classification to assign sentences to:
    - **Subjective (S):** Patient's symptoms, concerns, and history.

  - **Objective (O):** Observations, test results, and vitals.
  - **Assessment (A):** Doctor's diagnosis and condition analysis.
  - **Plan (P):** Treatment plans, medications, and follow-ups.

Step 4: SOAP Note Generation Using gpt-4 & gpt-4o-mini

- Constructed a structured prompt for gpt-4 & gpt-4o-mini, incorporating classified SOAP sections.
- Ensured the generated SOAP notes were concise, structured, and medically relevant.

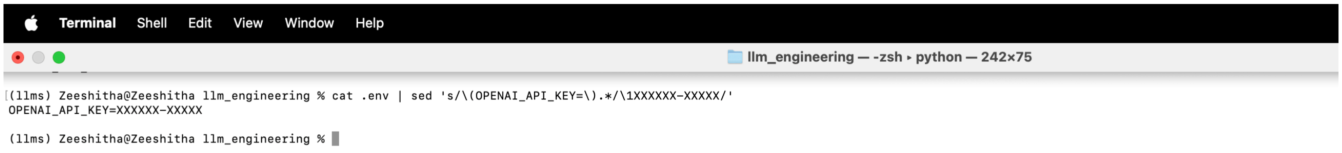
Step 5: Output Storage & Organization

- The generated SOAP notes were saved as (**1_A074_SOAP_gpt4.txt** & **1_A227_SOAP_gpt4.txt**), (**2_A074_SOAP-gpt4.txt** & **2_A227_SOAP-gpt4.txt**) and (**A074_SOAP-gpt-4o-mini.txt** & **A227_SOAP-gpt-4o-mini.txt**)
- The system was designed to handle multiple transcript files dynamically.

API Choice:

API Key Verification:

The OpenAI API key was successfully loaded from the .env file, ensuring secure authentication. Below is a masked version of the key for security purposes:



```
(llms) Zeeshitha@Zeeshitha llm_engineering % cat .env | sed 's/\(OPENAI_API_KEY=\).*/\1XXXXXX-XXXXX/'
OPENAI_API_KEY=XXXXXX-XXXXX

(llms) Zeeshitha@Zeeshitha llm_engineering % 
```

Reasons for Choosing GPT-4:

- **High Accuracy:** Effectively processes complex medical text.
- **Robust Performance:** Reliable for structured documentation.
- **Proven Results:** Frequently used in clinical note generation tasks.

Enhancements & Future Improvements To further refine the system, the following improvements could be implemented:

- **Enhanced Chunking Strategy:** Dynamically detect topic shifts instead of using fixed-size chunks.
- **Fine-Tuned Model for SOAP Classification:** Train a custom NLP model specifically for SOAP note structuring.

- **Integration with Medical Ontologies:** Leverage SNOMED CT or UMLS for more accurate entity recognition.
- **Validation with Healthcare Professionals:** Collaborate with doctors to fine-tune the SOAP note generation process.

Submission Package Contents Final Deliverables:

- **Jupyter Notebook (SOAP_Generation-gpt4.ipynb & SOAP_Generation-gpt-4o-mini.ipynb):** Contains the complete implementation.

- **Generated SOAP Notes** (**1_A074_SOAP_gpt4.txt** & **1_A227_SOAP_gpt4.txt**), (**2_A074_SOAP-gpt4.txt** & **2_A227_SOAP-gpt4.txt**) and (**A074_SOAP-gpt-4o-mini.txt** & **A227_SOAP-gpt-4o-mini.txt**)

  Contains final structured SOAP notes.

- **Project Documentation (SOAP_Notes_Explanation.pdf)** – This document.

Conclusion:

This project successfully automates SOAP note generation by processing doctor-patient conversations using GPT-4 and NLP techniques. The chunking implementation ensures scalability for longer transcripts, while the structured methodology results in clear, clinically useful SOAP notes.