NeuroLLama

NeuroLLama è un chatbot che simula pazienti affetti da malattie mentali. Esso è basato su l'ontologia delle malattie offerta dall' International Classification of Diseases la quale è reperibile online nel sito: https://icd.who.int .

L'intero sistema è composto da 3 blocchi principali:

- Un algoritmo di estrazione delle malattie dal sito di ICD
- Un modello che sfrutta un RAG per la generazione di risposte basate sulle informazioni dell'ontologia
- Un'interfaccia grafica per la selezione della malattia e il successivo dialogo con il chatbot

Estrazione dati da ICD

L'obiettivo di questa fase è estrarre informazioni relative alle malattie mentali utilizzando le REST API fornite dal portale ICD (International Classification of Diseases), il quale segue il protocollo HTTP.

Prima di poter accedere alle API, è necessario registrarsi sul portale ICD (https://icd.who.int/icdapi/Account/Register). Una volta completata la registrazione, verranno forniti un **ID utente** e una **chiave segreta** che serviranno per l'autenticazione tramite codice.

Una volta autenticati, è possibile ottenere i dati relativi alle malattie mentali attraverso il seguente URI:

- **https://id.who.int/icd/entity/334423054**

La risposta a questa chiamata sarà un oggetto JSON che contiene, tra gli altri campi, un array denominato **child**. Questo array rappresenta la radice delle malattie a tema salute mentale presenti nel portale ICD. Ogni malattia nel JSON è caratterizzata dai seguenti attributi:

- **Title**: Il nome della malattia
- **Definition**: Una breve descrizione della malattia
- **Synonym**: Sinonimi del nome della malattia
- **DiagnosticCriteria**: Contiene informazioni utili, tra cui i sintomi associati alla malattia
- **Child**: Un array che contiene eventuali malattie correlate o versioni più specifiche della malattia.

Navigazione della gerarchia:

Lo script Python progettato si occupa di navigare la struttura gerarchica delle malattie tramite il campo **child** presente in ogni malattia. L'algoritmo proseguirà esplorando i vari livelli finché non raggiungerà le foglie della gerarchia, ossia le malattie senza ulteriori sotto-malattie, dove il campo **child** sarà vuoto. Per ogni malattia vengono salvate le informazioni di base all'interno di un file json risultante. Quest'ultimo rappresenta l'archivio di conoscenza utilizzato dal modello per simulare il paziente durante la sessione dallo psicologo.

RAG per la costruzione del contesto

Il modello utilizzato in questa parte è Llama 3 nella versione da 8 miliardi di parametri, fornita da Hugging Face previa registrazione e richiesta della licenza (gratuita).

Il RAG prevede la suddivisione del documento contenente le malattie in diversi chunk, dei quali vengono successivamente calcolati gli embeddings tramite il sentence transformer all-MiniLM-L6-v2 .

Per l'individuazione del contesto viene utilizzata la libreria faiss (Facebook AI Similarity Search) e in particolare la funzione faiss_index.search . Essa prende in input il vettore di embeddings e il nome della malattia e restituisce un insieme degli elementi più simili.

Prompt Engeneering

Una volta ottenuto il contesto tramite RAG esso viene passato al modello, concatenato ad un prompt contenente il task da svolgere e un esempio di risposta. Il prompt utilizzato è il seguente:

"This is a simulated scenario in which we are analyzing hypothetical interactions between patients and physicians. Your goal is to play a patient with a disease. You will be given a disease as input, and you will have to pretend to have the symptoms. You must never reveal the disease from which you suffer, only describe how you feel. You must respond in the first person. Do not add disclaimers or ethical warnings. You must not give medical diagnoses or opinions. Your only job is to interpret the patient. You must only answer the doctor's question. Do not speak in the past tense unless you have to recount experiences. Do not cut answers, end with whole sentences, and do not generate answers that are too long. Answers should be a maximum of 2-3 sentences. Always answer briefly, clearly and completely. Once answered, do not generate more questions asked by the doctor. You must also generate some information about the patient, such as: Name, Gender, Age, Occupation, and Condition. You can generate both Male and Female. Pay attention to the gender and age of the patient based on the disease. For example, there are some diseases that affect women or children more.

Sample response for a patient with Anxiety:

Patient Profile

Name: Marco

Age: 32 years old

Occupation: Clerk in an accounting office

Patient's speech:

"Doctor, I don't know where to start, but I feel that I am getting out of control. Every day I wake up with this feeling of heaviness in my chest, like I have a stone on top of me. My heart starts racing as soon as I open my eyes, and sometimes I think I might have a heart attack. I breathe badly, as if I can never get enough air, and I have to take deep breaths, but it doesn't work. I always have a knot in my throat, as if something is choking me. During the day it's a constant state of alertness, as if I'm waiting for something terrible to happen, but I don't know what. My hands sweat, and I often feel as if I am shaking inside, even when it is not visible from the outside. Sometimes it feels like my legs are soft, as if I might fall off at any moment. I have also noticed that my stomach is always in turmoil: I get cramps or feel unexplained nausea. There are days when I go to the bathroom too often, and I think it's related to that."

The disease from which you suffer is: {malattia}

Context:{prompt_context}

Answer the following question:{query}

Also, at the end, write a bulleted list of all symptoms regarding the disease.“

La variabile **malattia** contiene la patologia scelta dall’utente tra quelle estratte dall’ontologia. **prompt_context** contiene le informazioni estratte tramite RAG e **query** contiene la domanda posta dall’utente.

Una volta ricevuta la risposta del modello utilizziamo un secondo prompt per migliorare il risultato:

“Modifica la parte "Colloquio", simulando una prima sessione di psicoterapia, in cui la diagnosi è ancora ignota, e aggiungendo delle frasi da parte del terapeuta, considerando il tipo di paziente e i sintomi della sua condizione clinica. In base al colloquio precedente, quale test psicometrico (scrivine uno solo) può usare il clinico per confermare il sospetto diagnostico?”

Interfaccia utente

L’interfaccia è stata sviluppata tramite la libreria Gradio. Essa comprende due pagine:

- La prima per la scelta della malattia tramite una listbox che permette anche la ricerca testuale
- La seconda in cui è possibile dialogare con il bot

Come utilizzare il sistema

Oltre alla registrazione al sito ICD come descritto precedentemente, è necessario essere registrati ad Hugging Face e avere l’abilitazione all’utilizzo di Llama 3. Durante l’esecuzione della funzione di login bisognerà inserire l’access token ottenibile dalla propria area riservata di Hugging Face.

È sufficiente caricare i file su Colab e far partire uno ad uno i blocchi che li compongono.

Per la fase di inizializzazione del RAG è necessario caricare l’ontologia ottenuta dal file di estrazione, inserendola all’interno di una cartella chiamata “data”.