EINFÜHRUNG IN DAS MASCHINELLE LERNEN – L6

MINI-RAG-SYSTEM

Bekommen in L6

Abgabe in L7

Implementiere ein minimalistisches RAG (Retrieval Augmented Generation) - System, d.h. ein Sprachmodell welches mit einem vorgegebenen Dataset fine-tuned ist und Fragen aus dem Bereich beantworten kann.

Installiere mit pip die transformers und sentence-transformers Bibliotheken. Siehe hier, hier. Als Grundmodell kannst du mistralai/Mistral-7B-v0.3 benutzen und als Basisembeddings sentence-transformers/all-MiniLM-L6-v2. Diese wurden auf ein Laptop mit intel Core i7 CPU und 16 GB RAM getestet, es funktioniert in einigen Minuten, und es hat positive Bewertungen. Je nach verfügbarer Hardware kannst du ein größeres oder kleineres Modell benutzen.

Benutze zum fine-tuning die Quellen welche als Literaturreferenz für dieses Kurs dienen, diese befinden sich hier, hier, hier. Vergewisse dich, dass du die neuen Gewichte lokal speicherst, damit das Modell wiederverwendbar ist. Diese Embeddings werden auch für die Repräsentierungen der Fragen benutzt. Mit diesen macht man zuerst fine-tuning des Embedding-modells. Du kannst gerne auch zusätzliche Dateien einlesen. Diese werden z.B. mit Hilfe der pypdf2-Bibliothek eingelesen. Verwende danach chunking auf die eingelesenen Daten. Das stellt sicher, dass die eigelesenen Texte in praktikablen Mengen dem Modell gegeben werden (erinnere dich an die token limit von LLMs!!). Der gesamte Kontext ist wichtig, deshalb soll es zwischen den Chunks (eigentlich, n-grams) einige Wörter Überlappung geben. Siehe, z.B. hier.

Implementiere LLM-as-a-judge. Dieser ist ein zweites LLM und es soll die Antwort auf Richtigkeit, Vollständigkeit, Halluzinationen, usw. überprüfen. Siehe hier, hier, hier.

Implementiere Guardrails. Das bedeutet: Inhalt, welches nach allgemeinen gesetzlichen, ethischen oder jugendschutzrechtlichen Standards als nicht angemessen anerkannt ist (Gewalt, Obszönität, Sexualität o.ä.), soll absolut verweigert werden, sowie die Ausgabe von solchem Inhalt soll verboten sein; und im Allgemeinen soll die Annahme von off-topic Fragen verhindert werden. In einer tatsächlichen Anwendung kann es sein, dass problematischer Inhalt geloggt wird und dass man Maßnahmen gegen den Benutzer trifft.

Passe auf die Richtigkeit der Prompts auf. Experimentiere und iteriere. Notiere welche Anweisungen besser oder schlechter funktionieren.

Drucke aus Statusnachrichten bei jedem Schritt (verbose). Benutze logging.

Bei der ersten Benutzung des Programms werden die vortrainierten Gewichttensoren und Embeddings von Huggingface heruntergeladen. Diese haben mehrere GB. Vergewisse dich, dass du genügend Freiraum und eine stabile Internetanbindung hast. Danach läuft alles lokal. Man kann diese auch im Voraus herunterladen, aber man muss dann sicherstellen, dass sie sich an richtiger Stelle befinden.

Implementiere auch ein einfaches GUI, so dass man mit dem System in einem Chatfenster interagieren kann. Für dieses Teil der Aufgabe kannst du gerne den Code mit genAI erstellen.