Προαιρετική Εργασία

Δημιουργία RAG αρχιτεκτονικής για ερωτήσεις φυσικής γλώσσας σε ειδησεογραφικά άρθρα

Καταληκτικές Ημερομηνίες

Παρασκευή 23 Μαΐου 2025	Παράδοση εργασίας
Εβδομάδα 26 Μαΐου 2025	Προφορική εξέταση (οι ακριβείς ημέρες και ώρες θα ανακοινωθούν αργότερα)

Η εργασία μπορεί να γίνει σε ομάδες έως 2 ατόμων.
Η εργασία μετράει σε ποσοστό 50% στο βαθμό σας στο μάθημα και είναι απαλλακτική (αντικαθιστά την τελική εξέταση)

Στην προαιρετική εργασία θα σχεδιάσετε και υλοποιείστε μια RAG αρχιτεκτονική που θα σας επιτρέψει να απαντάτε σε ερωτήσεις σχετικά με τα ειδησεογραφικά άρθρα.

Τα άρθρα είναι τα ίδια με αυτά της υποχρεωτικής εργασίας και μπορείτε να τα βρείτε στο kaggle: **https://www.kaggle.com/datasets/hadasu92/cnn-articles-after-basic-cleaning/data**

Προτείνετε για την υλοποίηση να χρησιμοποιήσετε το εργαλείο LangChain.

Για τη διανυσματική βάση μπορείτε να χρησιμοποιήσετε για παράδειγμα τη MongoDB ή τη FAISS μέσω του LangChain.

Η RAG αρχιτεκτονική θα πρέπει να υποστηρίζει τα παρακάτω:

(α) Διαχωρισμός του κειμένου σε μικρότερα τμήματα (chunking). Δυο συνηθισμένοι chunkers που παρέχει το LangChain είναι οι εξής: RecursiveCharacterTextSplitter και SemanticChunker.

(β) Δημιουργία διανυσμάτων για κάθε τμήμα (chunk). Για την δημιουργία των επιμέρους διανυσμάτων (embeddings) μπορείτε να χρησιμοποιήσετε ένα sentence-transformer model (π.χ. το all-MiniLM-L6-v2), τα οποία θα αποθηκεύσετε στην διανυσματική βάση δεδομένων.

(γ) Ανάκτηση relevant chunks. Με το ίδιο μοντέλο που χρησιμοποιήθηκε για την δημιουργία διανυσμάτων, θα πρέπει να γίνει embed και το query του χρήστη και στην συνέχεια με βάση

την ομοιότητα μεταξύ του embedded query και των embedded chunks να επιστραφούν τα πιο relevant chunks. (Στην περίπτωση που χρησιμοποιήσετε MongoDB υποστηρίζεται η BM25($meta) με επιστροφή των K-top documents, οπότε μετά μπορείτε να χρησιμοποιήσετε cosine similarity μεταξύ του query και των filtered documents και να κρατήσετε το τελικό αποτέλεσμα).

(δ) Επαύξηση του LLM prompt. Προσθέστε τα relevant chunks από το προηγούμενο βήμα μαζί με το user query ως είσοδο στο LLM.

Παράδειγμα ενός prompt:

"You are an AI assistant. Answer the following question: {query}, based on the following context: {retrieved_documents}".

(ε) Evaluation του συστήματος. Δώστε παραδείγματα 5 queries και των αποτελεσμάτων τους με και χωρίς χρήση της επέκτασης (δηλαδή, τα αποτελέσματα με και χωρίς context). Επίσης για τα queries αυτά δώστε και τα έγγραφα (chunks) που χρησιμοποιήθηκαν.

Εγκατάσταση ενός LLM με την χρήση του OLLAMA

1. Για την εγκατάσταση ενός LLM θα χρειαστεί να τρέξετε την παρακάτω εντολή στο terminal: ollama pull {model_name}

2. Για να χρησιμοποιήσετε το LLM που εγκαταστήσατε θα πρέπει το OLLAMA να βρίσκεται σε λειτουργία.

   ollama run {model_name}

3. Τα προτεινόμενα μοντέλα είναι: llama3.2:1b (1.3G) ή llama3.2:3b (2G)

4. Για την χρήση του embedding model (all-MiniLM-L6-v2) μπορείτε να χρησιμοποιήσετε επίσης το OLLAMA, διαφορετικά την βιβλιοθήκη SentenceTransformer από το HuggingFace.

Προτεινόμενα εργαλεία:

LANGCHAIN: https://www.langchain.com/

OLLAMA: https://ollama.com/download

MONGO DB: https://www.mongodb.com/

MONGODB($meta):https://www.mongodb.com/docs/manual/reference/operator/aggregation/meta/