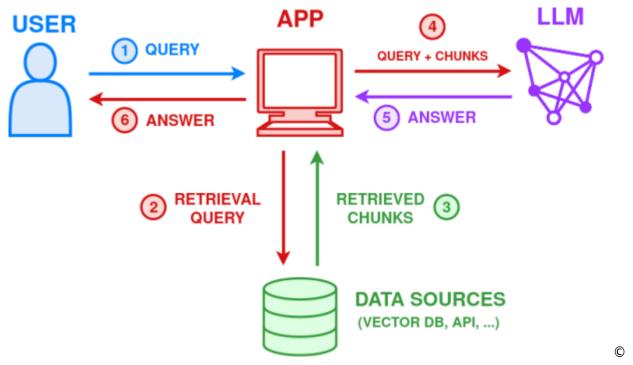
2025/11/18 05:48 1/2 LU11b - RAG Chatbot

LU11b - RAG Chatbot

RAG (Retrieval-Augmented Generation) ist eine gängige Variante, um Fragen zu benutzerspezifischen Dokumenten zu beantworten.



https://www.ridgerun.ai/post/how-to-evaluate-retrieval-augmented-generation-rag-systems

Zu Beginn muss man die gewünschten Daten (z.B. Lernunterlagen) ...

- 1. ... in Chunks unterteilen
- 2. ... "embedden"
- 3. ... in die Vektordatenbank speichern

Der Ablauf funktioniert grob so:

- 1. Der Benutzer gibt eine Frage ein, welche an die Applikation geschickt wird.
- 2. Die Frage wird "Embedded" und an die Vektordatenbank geschickt
- 3. Von der Vektordatenbank werden die n passendsten Chunks im Klartext an die Applikation zurückgegeben.
- 4. Die Applikation schickt die originale Frage im Klartext mitsamt den Chunks an ein LLM-Model
- 5. Das LLM-Model schickt eine Antwort zurück an die Applikation
- 6. Die Applikation kann z.B. Quellen o. Ä. bei Bedarf ergänzen und die Antwort an den Benutzer zurückschicken.

Vektordatenbank

Last update:

2025/11/18 de:modul:ffit:3-jahr:java:learningunits:lu11:b https://wiki.bzz.ch/de/modul/ffit/3-jahr/java/learningunits/lu11/b?rev=1763428332 02:12

From:

https://wiki.bzz.ch/ - BZZ - Modulwiki

Permanent link:

https://wiki.bzz.ch/de/modul/ffit/3-jahr/java/learningunits/lu11/b?rev=1763428332

Last update: 2025/11/18 02:12



https://wiki.bzz.ch/ Printed on 2025/11/18 05:48