engAIge®

Aufgabe: RAG Pipeline

Erstelle ein Python-Programm, das den Text aus einer Sammlung von PDF-Dokumenten extrahiert und ihn in einer Vektordatenbank speichert. Verwende diese Datenbank anschließend, um Textanfragen zu beantworten, indem du relevante Passagen aus den gespeicherten Dokumenten abrufst und in die Antwort integrierst.

Beginne mit der Extraktion des Textes aus den PDF-Dokumenten. Lade die PDF-Dateien in das Programm und verwende eine geeignete Bibliothek wie PyMuPDF, um den Text zu extrahieren. Zerlege den extrahierten Text in sinnvolle Absätze oder Abschnitte, die anschließend weiterverarbeitet werden können. Speichere die erzeugten Textvektoren in einer Vektordatenbank. Stelle sicher, dass die Vektoren effizient gespeichert und abgerufen werden können.

Implementiere eine Funktion, die eine Textanfrage entgegennimmt, diese vektorisiert und die relevantesten Passagen aus der Vektordatenbank abruft. Kombiniere die abgerufenen Passagen, um eine kohärente Antwort auf die Anfrage zu erstellen. Achte darauf, dass die Antwort sinnvoll und verständlich ist. Überlege dir auch Möglichkeiten die Ergebnisse zu verbessern!

Folgende Fragen sollten beantwortet werden:

- Wie hoch ist die Grundzulage?
- Wie werden Versorgungsleistungen aus einer Direktzusage oder einer Unterstützungskasse steuerlich behandelt?
- Wie werden Leistungen aus einer Direktversicherung, Pensionskasse oder einem Pensionsfonds in der Auszahlungsphase besteuert?