KIDA
AI for Food and Agriculture

Gefördert durch:
Bundesministerium für Landwirtschaft, Ernährung und Heimat
aufgrund eines Beschlusses des Deutschen Bundestages

Projektträger
Bundesanstalt für Landwirtschaft und Ernährung

Zusammenfassung von Abschlussberichten

Textanalyse und -extraktion mit großen Sprachmodellen (LLMs)

- **Manuelle Auswertung von Dokumenten ist sehr zeitaufwändig**
- **Entwicklung einer RAG Pipeline mit Open-Source-Frameworks (Ollama, Langchain)**
- **Mit RAG können Dokumente wie PDF-Dateien weiterarbeitet werden, sodass die Antworten auf konkrete Fragen aus den Dokumenten extrahiert werden können**

Hintergrund und Fragestellung

Ziel des Projekts war die **Zusammenfassung und Informationsextraktion aus** rund 30 **Abschlussberichten** geförderter Projekte im Bereich Landwirtschaft. Alle Dokumente (PDFs) zeichneten sich dadurch aus, dass sie der gleichen vorgegebenen Struktur folgen. Im Projekt sollten nun mehrere Fragen mit Hilfe von KI zu einem oder mehreren Gliederungspunkten beantwortet werden. Zum Beispiel „wurden die Ziele des Projekts erreicht?“ Die manuelle Beantwortung dieser Frage würde bedeuten, mehrere 100 Seiten Text lesen und zusammenfassen zu müssen. Diesen großen Arbeitsaufwand sollte durch den Einsatz von KI reduziert werden, indem **große Sprachmodelle** (LLMs) die **Texte analysieren**, die relevanten **Informationen extrahieren** und **Antworten** auf die Fragen **geben**.

Vorgehensweise

Der erste Schritt war die **maschinelle Verarbeitung** der Texte, damit diese vom einem LLM benutzt werden können. Dafür wurde eine **RAG** (Retrieval Augmented Generation) **Lösung** erstellt. RAG-Lösungen ermöglichen es dem Nutzenden mit ihren Dokumenten zu interagieren oder zu chatten und Informationen aus verschiedenen Dokumenten zu extrahieren. Folgende Schritte wurden dafür durchgeführt:

1) Die Texte in den Dokumenten wurden in kleineren Abschnitten geteilt (**Text-Chunking**).

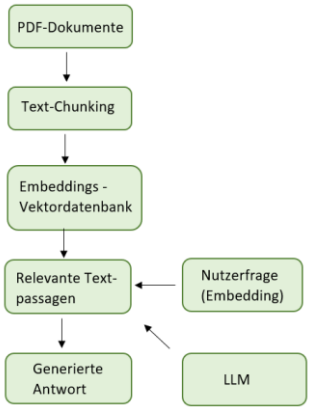
2) Die Textabschnitte wurden **in** sogenannte **Embeddings** (das sind numerische Darstellung der Texte) **umgewandelt**, da Sprachmodelle rohen Text nicht direkt verarbeiten können.

3) Eine **Vektordatenbank** wurde **erstellt** und die Embeddings wurden so in der Datenbank gespeichert, dass semantisch ähnliche Embeddings (Textpassagen) zusammenbleiben.

4) **Jede Frage** wurde **als Embedding** umgewandelt, damit der (Retrieval-)Algorithmus der RAG-Lösung die relevanten Textpassagen in der Vektordatenbank finden kann.

5) Um die Informationen aus allen Dokumenten zu berücksichtigen, werden die **Dateinamen als Filter** verwendet.

6) Die relevanten Informationen aus jedem Bericht, die die Fragen des Anfragenden beantworten, werden anschließend an ein **großes Sprachmodell** (LLM) übergegeben. Dieses **fasst** die Inhalte **zusammen** und **generiert** die finale **Antwort**.



Ergebnisse und Schlussfolgerungen

Mithilfe der Open-Source-Frameworks „Langchain“ und „Ollama“ war es möglich, eine **RAG-Pipeline** zu bauen, die relevante **Textpassagen extrahiert**, **zusammenfasst** und **Antworten** auf Fragen **formuliert**. Damit kann das erstellte Tool dabei unterstützen, die Zusammenfassung von Texten zu automatisieren, und somit Arbeitszeit zu sparen und die Mitarbeitenden zu entlasten. Für den konkreten Bericht lieferte die KI hilfreiche Antworten auf einzelne Fragen zu den Dokumenten und trug so maßgeblich zu einzelnen Berichtskapitel bei.

Kontakt	Informationen
KI-Beratung (KIDA): kida@bmleh.bund.de Petra Raue (Thünen): petra.raue@thuenen.de	**KIDA-Bearbeitende:** Cristina Ortiz Cruz (KIDA, MRI) **Team-Leitung:** Micha Schneider (KIDA KI-Beratung, Thünen) **Anfragende:** Petra Raue (Thünen) **Bericht:** Evaluierung der Förderung von Innovation in Landwirtschaft und ländlicher Entwicklung in Niedersachsen und Bremen. 5-Länder-Evaluierung ##/2025 (in Vorbereitung). www.eler-evaluierung.de