Praktikum 04: KI-Lern-Tutor

In der heutigen Praxis nutzen Unternehmen große Sprachmodelle (LLMs wie ChatGPT) nicht nur als Chatbot, sondern integrieren sie in ihre eigenen Daten. Das Problem: LLMs kennen die internen Unternehmensrichtlinien und neigen zum "Halluzinieren" (sie erfinden Fakten).

Die Lösung heißt Retrieval-Augmented Generation (RAG). In diesem Praktikum bauen Sie in n8n einen eigenen "KI-Lern-Tutor". Sie nutzen dafür ein vorgefertigtes Template. Ihre Aufgabe ist es, dieses Template zu verstehen, zu konfigurieren und mit einem datenschutzkonformen KI-Modell zu verknüpfen.

Das Ergebnis ist ein n8n-Workflow mit folgenden Funktionen:

- Vorlesungsunterlagen einlesen, z.B. die Folien der Vorlesung "Neuronale Netze"
- Den Text in Vektoren umwandeln und in einer Vektor-DB speichern
- Auf Basis dieser Daten Ihre Fragen beantworten.

Damit kombinieren Sie in diesem Lab die Konzepte Large Language Models, Prompt-Engineering und Retrieval-Augmented Generation um ein nützliches Werkzeug zu erstellen.

Vorbereitung: Zugang zu KI Connect NRW

Für dieses Praktikum nutzen Sie die datenschutzkonformen und kostenlose verfügbaren KI-Modelle der Landesinitiative **KI Connect NRW**. Da deren Schnittstelle "OpenAI-kompatibel" ist, können wir sie nahtlos in n8n einbinden.

1. API-Key generieren:
   - Loggen Sie sich auf https://chat.kiconnect.nrw/app ein.
   - Klicken Sie auf Ihren Benutzernamen und wählen Sie "API-Schlüsselverwaltung".
   - Erstellen Sie sich einen API-Schlüssel und kopieren Sie diesen. Sie benötigen den Schlüssel später.
2. n8n vorbereiten:
   - Starten Sie Ihre n8n Docker-Umgebung aus Praktikum 1 und öffnen Sie http://localhost:5678.
   - Klicken Sie links im Menü auf "Overview" " und dann auf den Reiter "Credentials". Wählen Sie "Add Credential" " und suchen Sie nach *OpenAI*. Klichen Sie auf "Continue".
   - Fügen Sie Ihren KI Connect API-Schlüssel im Feld "API-Key" ein.
   - Fügen Sie die URL https://chat.kiconnect.nrw/api/v1 im Feld "Base URL ein und speichern Sie die Zugangsdaten ab.

Schritt 1: Das RAG-Template importieren

Um das System nicht komplett von null aufzubauen, nutzen wir eine Vorlage, die bereits die wichtigsten RAG-Komponenten enthält.

1. Klicken Sie im linken Menü von n8n auf Templates.
2. Suchen Sie nach dem Template mit dem Namen: RAG Starter Template using Simple Vector Stores, Form trigger and OpenAI
3. Klicken Sie auf das Template und wählen Sie "Use for free". Und importieren Sie den Workflow. Sollte die Option zum Import nicht verfügbar sein öffenen Sie bitte die URL http://localhost:5678/templates/5010/setup. Dies sollte den Import des Templates starten.
4. Benennen Sie den erzeugen Workflow in *KI-Lern-Tutor* um.

Der Workflow wird nun in Ihre Arbeitsfläche geladen. Sie sehen ein Netzwerk aus verschiedenen Knotenpunkten. Machen Sie sich kurz mit der Struktur vertraut (Chat, Agent, Vector Store, Embeddings).

Schritt 2: Den KI-Agenten mittels Prompt Engineering konfigurieren

Der Knoten AI Agent ist die zentrale Komponente des KI-Lern-Tutors. Sie müssen diese so konfigurieren, dass er ein Tutor ist und sich streng an die Vorlesungsfolien hält. Dafür verwenden Sie einfaches Prompt Engineering.

1. Doppelklicken Sie auf den Knoten AI Agent in Ihrem neuen Workflow.
2. Im Bereich "Options" fügen Sie durch einen Klick auf das + das Feld "System Message".
3. Ersetzen Sie den englischen Standard-Text durch folgenden System-Prompt:

   Du bist der offizielle Lern-Tutor für das Modul Informationssysteme. Deine Aufgabe ist es, Fragen der Studierenden zu beantworten.

   REGELN:

   Nutze AUSSCHLIESSLICH das Werkzeug (Query Data Tool) mit den Vorlesungsfolien, um Antworten zu finden.

   Wenn eine Information nicht in den Folien steht, antworte exakt mit: "Das steht leider nicht in den Vorlesungsunterlagen." Erfinde niemals eigene Antworten!

Schritt 3: Die Modelle auf KI Connect NRW umstellen

Das Template ist standardmäßig für die OpenAI-Server konfiguriert. Sie leiten die Anfragen nun auf unsere Server in NRW um. Dies erfodert die Anpassung des Chat-Modells und des Embedding-Modells.

1. Das Chat-Modell anpassen:
   - Öffnen Sie den Knoten OpenAI Chat Model (links neben dem Agenten).
   - Wählen Sie oben Ihre erstellten Credentials aus.
   - Wählen Sie als Modellnamen eines der verfügbaren Modelle aus (z.B. `openai-gpt-oss-120b`).

2. Das Embedding-Modell anpassen:
   - Öffnen Sie den Knoten Embeddings OpenAI.
   - Wählen Sie wieder Ihre Credentials aus.
   - Wählen Sie als Model ein verfügbares Embedding-Modell, z.B. `qwen-qwen3-embedding-8b`, aus.

Damit ist die Konfiguration des Workflows abgeschlossen. Die Anfragen werden nun durch die KI Connect APIs beantwortet.

Schritt 4: Das Vorlesungs-PDF einlesen

Als nächstes müssen Sie Vorlesungsunterlagen hochladen um dem KI-Lern-Tutor die notwendigen Informationen bereitzustellen. Gehen Sie hierzu wie folgt vor:

1. Klicken Sie auf "Execute Workflow".
2. Es öffnet sich ein Formular in dem Sie eine Datei auswählen können. Wählen Sie z.B. die PDF-Datei der Vorlesungsfolien zu künstlichen Neuronalen Netzen aus.
3. Klicken Sie auf "Submit".

Durch das Klicken auf Submit wird das PDF-Dokument hochgeladen und an die API zum Berechnen der Embeddings gesendet. Die Embeddings werden anschließend in einer einfachen Datenbank im Workflow gespeichert. Diese ist nicht persistent. D.h. nach einen Neustart des n8n-Servers gehen die Daten verloren.

Analysieren Sie Daten in der Ausführung des Embedding Modells. Sie erkennen, dass das PDF automatisch in kleiner Teile zerlegt wurde und für diese die Embeddings berechnet wurden.

Sie können den Workflow auch mehrfach ausführen um weitere Dokumente in der Datenbank zu speichern.

Schritt 5: Testen und "Halluzinationen" prüfen

Klicken Sie ganz unten in der Workflow auf "Chat". Dies öffnet ein Chat-Fenster in dem Sie Fragen an den KI-Lern-Tutor stellen können.

1. Test 1 (RAG funktioniert): Stellen Sie eine Frage, die aus den Vorlesungsunterlagen beantworte werden kann. Falls Sie die Vorlesungsunterlagen zur Vorlesung über neuronale Netze hochgeladen haben, können Sie z.B. folgende Fragen verwenden:

   - Was ist die Grundlegende Architektur von neuronalen Netzen.
   - Erstelle mir 5 Prüfungsfragen zum Thema neuronale Netze.

   Der KI-Lern-Tutor sollte kurz nachdenken und die Antwort aus den Folien geben.

Im Log des Workflows können Sie nun nachvollziehe, dass die Daten aus der Datenbank zur Beantwortung der Frage verwendet wurden. Im Knoten "Query Data Tool" sehen Sie z.B. welche Teile der Folien genau verwendet wurden.

2. Test 2 (System Prompt funktioniert): Fragen Sie: "Wie wird das Wetter morgen in Aachen?"

   Ergebnis: Der KI-Lern-Tutor MUSS antworten: "Das steht leider nicht in den Vorlesungsunterlagen." Er wird Ihnen keine Information zum Wetter geben. Stellen Sie die gleiche Frage im Chat-Fenster von KI Connect mit dem gleichen Modell und vergleiche Sie die Antwort.

Schritt 6: Abgabe des Workflows

Um das Praktikum erfolgreich abzuschließen, reichen Sie Ihren funktionsfähigen Workflow ein. Klicken Sie hierzu im n8n-Menü (oben links oder über die drei Punkte) auf Download. Es wird eine .json-Datei heruntergeladen. Laden Sie diese Datei in ILIAS hoch.