Fortschritte im Projekt

1. Repository-Struktur

Das GitHub-Repository wurde neu strukturiert, um alle wichtigen Elemente wie den Python-Code, organisatorische Dokumente und Analyse-Skripte klar getrennt und übersichtlich zu organisieren.

Struktur:

- src/: Enthält den Python-Code des Chatbots.
- data/: Enthält die organisatorischen Dokumente, die als Grundlage für die Antworten dienen.
- requirements.txt: Listet die notwendigen Python-Abhängigkeiten, einschließlich `spaCy`, zur Installation auf.

2. Datenaufbereitung

Die relevanten organisatorischen Dokumente wurden gesammelt und in digitale Formate wie `.txt` und `.pdf` übertragen.

3. Anwendung von spaCy zur Textanalyse

1. Aufgabe

Im Rahmen der Entwicklung wurde eine separate spaCy-basierte Textanalyse durchgeführt. Ziel war es, die grundlegenden Funktionen von spaCy zur Extraktion von Entitäten und Satzsegmentierung zu testen und zu validieren.

Vorgehen

- **Dateninput**: Ein festgelegter Text wurde als Grundlage verwendet. Dieser wurde bereinigt, um überflüssige Leerzeichen und Zeilenumbrüche zu entfernen.
- **Textverarbeitung**: Der Text wurde mit spaCy analysiert. Dabei kamen zwei zentrale Funktionen zur Anwendung:
  - **Entitäten-Erkennung**: Benannte Entitäten wie Personen, Orte oder Objekte wurden identifiziert und kategorisiert.
  - **Satzsegmentierung**: Der Text wurde in einzelne Sätze unterteilt, um die Struktur des Inhalts zu verdeutlichen.
- **Darstellung der Ergebnisse**: Die identifizierten Entitäten und Sätze wurden separat ausgegeben, um die Resultate übersichtlich darzustellen.

2.Aufgabe

Die Aufgabe bestand darin, PDF-Dokumente einzulesen, ihren Inhalt zu bereinigen mit spaCy analysieren um Entitäten und Sätze zu extrahieren.

Vorgehen

- **Einlesen**: Der Text aller Seiten eines PDFs wurde zusammengeführt.
- **Bereinigung**: Überflüssige Leerzeichen und störende Sonderzeichen wurden entfernt, um einen klaren Text zu erzeugen.
- **Analyse**: Mit spaCy wurden benannte Entitäten identifiziert und der Text in Sätze unterteilt.
- **Verarbeitung mehrerer Dateien**: Alle PDFs in einem Ordner wurden nacheinander analysiert, die Ergebnisse übersichtlich dargestellt.

Ergebnisse

Die Methode ermöglicht eine automatisierte Verarbeitung von PDF-Inhalten. Entitäten und Sätze wurden erfolgreich extrahiert und strukturiert ausgegeben.

Fazit

Alle Aufgaben wurden erfolgreich umgesetzt.

```
Für mehr Informationen besuchen Sie das [GitHub-
Repository](https://github.com/asy0/AI_InfoChatbot/tree/main).
```

Nächste Schritte:

- Erweiterte Bereinigung (z. B. Kopf-/Fußzeilen entfernen).
- Verbesserung der NLP-Funktionalität, insbesondere durch das Trainieren eines maßgeschneiderten Modells.
- Semantische Analyse und Visualisierung.