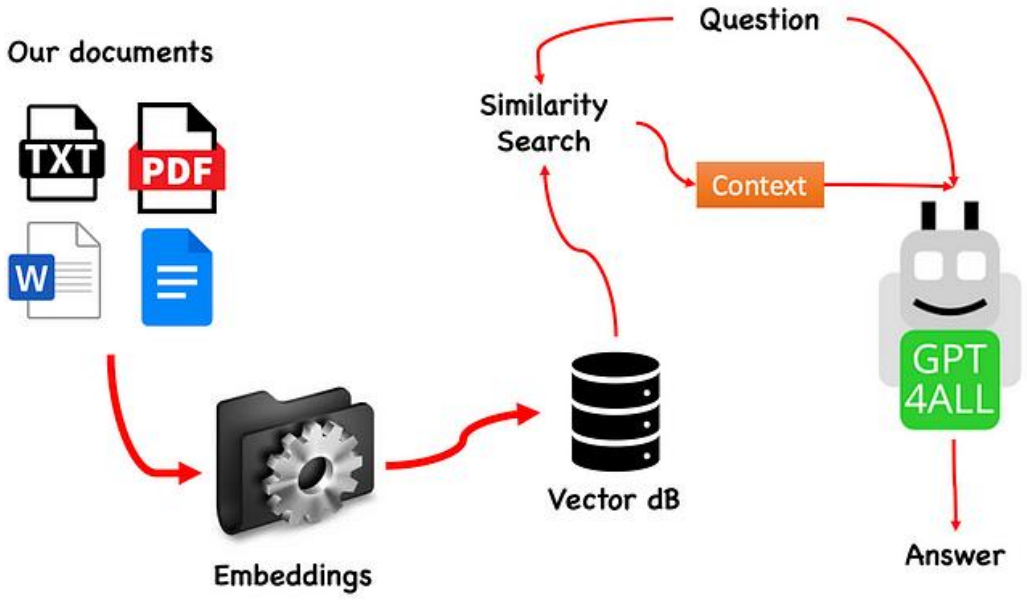
Konto 1260 Entwicklungen

Begründung der veränderungen

Archiv - Abfrage für Öffentliche Verwaltungen. Hier eine kurze Einführung

Wie wird es funktionieren?

Wichtig bei allen Abfragen ist, dass man nicht Openai verwendet, wenn es sich um vertrauliche Dokumente handelt, die werden sonst für das fine-tuning von Openai Verwendet und sin für alle zugänglich. Deshalb diese Entwicklung



Ablauf des QnA mit GPT4All - erstellt vom Autor

Der Vorgang ist wirklich einfach (wenn man ihn kennt) und kann auch mit anderen Modellen wiederholt werden. Die Schritte sind wie folgt:

- das Modell GPT4All laden
- *Langchain* verwenden, um unsere Dokumente abzurufen und sie zu laden

- Aufteilung der Dokumente in kleine, durch Embeddings verdauliche Stücke
- Verwenden Sie FAISS zur Erstellung unserer Vektordatenbank mit den Einbettungen
- Führen Sie eine Ähnlichkeitssuche (semantische Suche) in unserer Vektordatenbank durch, die auf der Frage basiert, die wir an GPT4All weitergeben wollen: dies wird als *Kontext* für unsere Frage verwendet
- Geben Sie die Frage und den Kontext mit *Langchain* an GPT4All weiter und warten Sie auf die Antwort.

Was wir also brauchen, sind Einbettungen. Eine Einbettung ist eine numerische Darstellung einer Information, z. B. von Text, Dokumenten, Bildern, Audio usw. Die Darstellung erfasst die semantische Bedeutung dessen, was eingebettet ist, und das ist genau das, was wir brauchen. Für dieses Projekt können wir uns nicht auf schwere GPU-Modelle verlassen: daher werden wir das native Alpaca-Modell benutzen und von *Langchain* die *LlamaCppEmbeddings* verwenden.