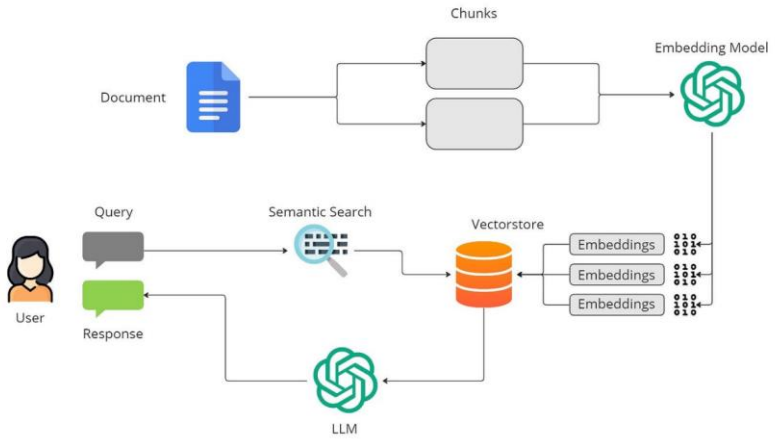
پیاده‌سازی چت بات با استفاده از رویکرد(Retrieval-Augmented Generation (RAG

مقدمه

در این گزارش، ما به بررسی و پیاده‌سازی یک چت‌بات با استفاده از رویکرد RAG می‌پردازیم. RAG یک روش پیشرفته است که با ترکیب قابلیت‌های بازیابی اطلاعات و تولید متن، امکان پاسخگویی به سوالات پیچیده را فراهم می‌کند.

کامپوننت‌های اصلی RAG شامل **بازیابی** (**Retrieval**)، **امبدینگ** (**Embedding**) و **تولید** (**Generation**) می‌باشد. RAG با ترکیب این سه کامپونت اصلی، به عنوان یک رویکرد قوی برای سیستم‌های پرسش و پاسخ پیچیده استفاده می‌شود. این رویکرد بهبود قابل توجهی در دقت، سرعت و کیفیت پاسخگویی به سوالات کاربران را فراهم می‌کند، به خصوص در مواقعی که نیاز به استفاده از اطلاعات دقیق و جمع‌آوری شده از منابع مختلف داریم.

معماری سیستم



مراحل انجام پروژه

1 .دریافت فایل‌های PDF از کاربر.

فایل PDF پس از دریافت بررسی می‌شود تا اطمینان حاصل شود که فایل PDF باشد، خالی نباشد و متن آن به زبان انگلیسی باشد.

2 .استخراج متن از فایل‌هایPDF

پس از دریافت فایل PDF، باید متن موجود در فایل استخراج شود. برای این کار از کتابخانه‌ی Pymupdf استفاده کردیم. این کتابخانه می‌تواند متن و متاداده‌های مربوط به صفحات PDF را استخراج کند.

3 .تقسیم متن به chunk کوچک

مدل‌های زبان طبیعی معمولاً محدودیت‌هایی در تعداد توکن‌ها دارند. بنابراین، متن استخراج شده باید به chunk کوچک‌تر تقسیم شود تا هر چانک به تعداد توکن‌های مجاز مدل بخورد.

4 .ایجاد Embedding برای chunk

برای تبدیل هر چانک به یک بردار متنی، از یک مدل Embedding در کتابخانه Transformers استفاده شده‌است. مدل sentence-transformers/multi-qa-mpnet-base-dot-v1 یک مدل امبدینگ پیشرفته است که برای تولید بردارهای عددی از جملات به کار می‌رود. این مدل با استفاده از معماری MPNet بهینه‌سازی شده است و در وظایف مانند پرسش و پاسخ و بازیابی اطلاعات عملکرد بسیار خوبی دارد. به عنوان یکی از مدل‌های مدرن امبدینگ در زمینه پردازش زبان طبیعی، شناخته می‌شود.

5 .ذخیره‌سازی Embedding در وکتور دیتابیس

Embedding تولید شده برای chunk باید در یک وکتور دیتابیس ذخیره شوند. برای این کار می‌توان از FAISS استفاده کرد که یک ابزار سریع و مقیاس‌پذیر برای جستجوی شباهت بردارها است.

6 .دریافت سوال کاربر

سوال کاربر از طریق واسط کاربری دریافت می‌شود. این سوال سپس به همان روش متن چانک‌ها، به بردار متنی تبدیل می‌شود.

برای اطمینان از معتبر بودن سوال کاربر، بررسی می‌شود که سوال خالی نباشد و یا تنها شامل اعداد نباشد.

7 .جستجوی نزدیک‌ترین چانک‌ها به سوال کاربر

با استفاده از FAISS، می‌توان نزدیک‌ترین چانک‌ها به سوال کاربر را جستجو کرد. این کار با جستجوی بردار سوال در میان بردارهای چانک‌ها انجام می‌شود.

8 .استفاده از مدل generative برای تولید پاسخ

چانک‌های نزدیک به همراه سوال کاربر به مدل generative داده می‌شوند تا پاسخ نهایی تولید شود.

برای وظیفه پرسش و پاسخ بر اساس PDF، انتخاب مدل مناسب بستگی به نیازها و محدودیت‌های خاص پروژه دارد.

مدل T5 (Text-to-Text Transfer Transformer) یک مدل زبان پیشرفته است که توسط Google توسعه یافته است. این مدل به طور خاص برای تبدیل همه وظایف پردازش زبان طبیعی به یک فرمت متن به متن طراحی شده است.

در زمان کنونی، چندین مدل T5 مختلف وجود دارد که هر کدام برای وظایف خاصی تنظیم شده‌اند. این مدل‌ها به تفاوت‌هایی مانند اندازه، زمینه کاربردی، و ترکیب تنظیمات توجه دارند.

در این تحقیق از مدل google/flan-t5-xl استفاده نمودیم که با ملاحظه منابع محاسباتی ما انتخاب شد. هدف ما در این تحقیق، مدیریت بهینه ترید آف بین منابع محاسباتی و دقت مدل با استفاده از این مدل بوده است.

9. در فرآیند دریافت خروجی از مدل generative، کاربر دو گزینه دارد: اول، استفاده از روش پارس ساده که خروجی مدل را به صورت یک رشته متنی به کاربر ارائه می‌دهد. دوم، استفاده از پارسر JSON که این امکان را فراهم می‌آورد که داده‌ها را به فرمت JSON دریافت کند، این فرمت معمولاً استفاده می‌شود برای تبادل داده‌های ساختارمند بین برنامه‌ها و سیستم‌ها. کاربر می‌تواند بر اساس نیاز خود یکی از این دو روش را برای دریافت خروجی مدل generative انتخاب کند.

نکته

برای استفاده از سرویس، لطفاً کد ضمیمه شده را اجرا کنید. پس از اجرای سرویس، ابتدا از شما درخواست می‌شود که نحوه‌ی پارس کردن خروجی را مشخص کنید. سپس، درخواست می‌شود که مسیر فایل‌های خود را مشخص کنید و پس از مشخص کردن مسیر برای همه داکیومنت‌ها، رشته "done" را وارد کنید. در نهایت، از شما خواسته می‌شود تا سوالات خود را مطرح کنید.