توسعه چت‌بات حوزه دانشگاه مبتنی بر RAG بدون نیاز به اینترنت

سارا اسعدی 403443013

مقدمه

این پروژه با هدف توسعه یک چت‌بات هوشمند برای دانشکده طراحی شده است که بر پایه روش بازیابی-افزوده-تولید (Retrieval-Augmented Generation یا RAG) کار می‌کند. هدف این سیستم پاسخگویی به سوالات کاربران با استفاده از اسناد موجود به صورت آفلاین است.

مراحل اصلی پیاده‌سازی

1. بارگذاری و پردازش اسناد: در ابتدا، اسناد PDF و DOC/DOCX از پوشه‌ی تعیین شده بارگذاری می‌شوند. برای پردازش این اسناد، از کلاس‌های PyPDFLoader و UnstructuredFileLoader استفاده شده است. سپس، متون استخراج شده به بخش‌های کوچکتر تقسیم می‌شوند تا مدل‌های زبانی بتوانند به راحتی آنها را پردازش کنند. تقسیم‌بندی متن بر اساس توکن‌ها و با استفاده از توکنایزر bert-base-multilingual-cased انجام می‌شود.

2. ایجاد Embedding : Embedding یکی از مراحل کلیدی در این پروژه است. در اینجا از مدل intfloat/multilingual-e5-large برای تبدیل متون به بردارهای عددی استفاده شده است.

   در این پروژه، مدل intfloat/multilingual-e5-large انتخاب شده است که توانایی پردازش چندزبانه را دارد و برای زبان فارسی نیز عملکرد مناسبی ارائه می‌دهد.

   این بردارها نمایشی عددی از معانی جملات هستند که امکان مقایسه و شباهت‌یابی بین متون مختلف را فراهم می‌کنند. این بردارها در پایگاه داده Chroma ذخیره می‌شوند که یک پایگاه داده برداری سریع و بهینه است.

3. بازیابی اطلاعات (Retrieval): در این مرحله، از پایگاه داده برداری برای بازیابی بخش‌هایی از متن که بیشترین شباهت را به سوال کاربر دارند، استفاده می‌شود. از روش similarity_score_threshold با مقدار آستانه 0.5 برای اطمینان از مرتبط بودن نتایج بهره برده شده است. سه بخش مرتبط (k=3) برای تولید پاسخ نهایی انتخاب می‌شوند.

4. تولید پاسخ (Generation): پس از بازیابی اطلاعات، مدل LLaMA (نسخه 3) برای تولید پاسخ نهایی استفاده می‌شود. مدل با توجه به متنی که از پایگاه داده بازیابی شده، پاسخ را به زبان فارسی تولید می‌کند. اگر اطلاعات کافی برای پاسخ وجود نداشته باشد، چت‌بات به کاربر اطلاع می‌دهد که اطلاعاتی در دسترس نیست.

   این ترکیب باعث می‌شود که چت‌بات بتواند پاسخ‌هایی دقیق‌تر و مرتبط‌تر ارائه دهد، زیرا مدل به جای تکیه بر حافظه داخلی خود، از داده‌های واقعی و مستندات برای پاسخگویی استفاده می‌کند.

5. رابط کاربری: برای تعامل کاربر با چت‌بات، از فریم‌ورک Streamlit استفاده شده است. این رابط کاربری ساده و کاربرپسند به کاربران اجازه می‌دهد به راحتی سوالات خود را مطرح کنند و پاسخ‌های مرتبط دریافت کنند.

ویژگی‌های کلیدی پروژه

- پشتیبانی از زبان فارسی: چت‌بات به طور کامل توانایی پاسخگویی به سوالات به زبان فارسی را دارد.
- عملکرد آفلاین: این سیستم به گونه‌ای طراحی شده که بدون نیاز به اتصال به اینترنت کار می‌کند، که برای محیط‌های محدود به اینترنت بسیار مناسب است.
- استفاده از مدل‌های منبع باز: تمامی مدل‌ها و ابزارهای استفاده شده در این پروژه منبع باز هستند و امکان توسعه و سفارشی‌سازی آن‌ها وجود دارد.
- ذخیره‌سازی پیشرفته: استفاده از پایگاه داده برداری Chroma باعث افزایش سرعت و کارایی در بازیابی اطلاعات شده است.

نتیجه‌گیری

این پروژه نشان می‌دهد که چگونه می‌توان با استفاده از معماری RAG و ابزارهای منبع باز، یک چت‌بات هوشمند و کارآمد توسعه داد. سیستم طراحی‌شده نه تنها توانایی پاسخگویی دقیق به سوالات کاربران را دارد، بلکه می‌تواند به راحتی گسترش یابد و برای کاربردهای مختلف در محیط‌های دانشگاهی، شرکتی و صنعتی استفاده شود.