مستندات فنی پروژه: سیستم RAG برای پرسش و پاسخ جغرافیایی

تاریخ :۲۶ اردیبهشت ۱۴۰۳ **توسط**: آرمان خلیلی - 4003623016

---

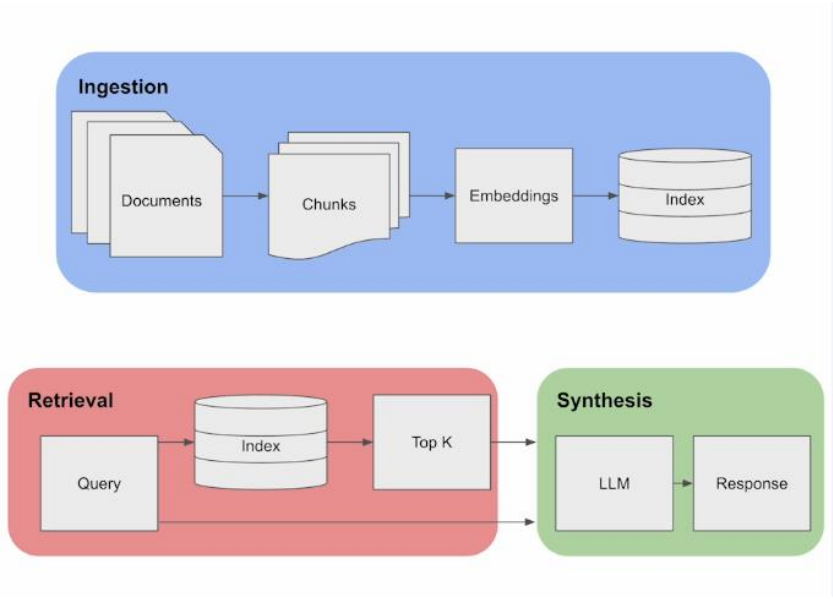
۱ .معرفی کلی پروژه

این سند، مستندات فنی یک سیستم کامل Retrieval-Augmented Generation (RAG)است که از ابتدا با هدف پاسخگویی هوشمند به سوالات مرتبط با جغرافیای فرانسه طراحی شده است. معماری سیستم بر پایه دو فاز اصلی Ingestion(پردازش و آماده‌سازی داده) و Inference(پردازش درخواست و تولید پاسخ) استوار است و از طریق یک API قدرتمند مبتنی بر FastAPI در دسترس قرار می‌گیرد.

این پروژه با به‌کارگیری تکنیک‌های پیشرفته، یک راهکار جامع برای ساخت سیستم‌های پرسش و پاسخ هوشمند ارائه می‌دهد.

۲ .معماری و گردش کار سیستم(Workflow)

گردش کار سیستم شامل دو Pipeline مجزا است:



الف: Ingestion Pipeline :

این فاز داده‌های خام وب را به یک پایگاه دانش ساختاریافته و قابل جستجو تبدیل می‌کند.

Web Content -> Scraper -> Text & Metadata -> Cleaner -> Chunker -> Embedding Model -> Vector Index (FAISS)

ب) Inference Pipeline :

این فاز درخواست‌های زنده کاربران را دریافت کرده، اطلاعات مرتبط را بازیابی و پاسخی دقیق تولید می‌کند.

User Query -> Retriever -> Top-K Chunks -> Re-ranker -> Final Context -> LLM -> Generated Answer

---

۳ .جزئیات فنی و تکنیک‌های استفاده شده

۳.۱ .پردازش و آماده‌سازی داده(Ingestion)

- **استراتژی :Chunking** برای تقسیم‌بندی اسناد، از رویکرد **Sentence-Window Chunking** استفاده شده است. متون به قطعات **۴ جمله‌ای** با **۱ جمله همپوشانی (Overlap)** تقسیم می‌شوند. این روش، ضمن حفظ یکپارچگی معنایی، قطعاتی بهینه برای مدل Embedding ایجاد می‌کند و از دست رفتن اطلاعات در مرز بین قطعات جلوگیری می‌کند.
- **مدل :Embedding** مدل **sentence-transformers/all-MiniLM-L6-v2** برای تبدیل متون به وکتورهای عددی انتخاب شده است. این مدل به دلیل تعادل عالی بین سرعت اجرا بر روی CPU و کیفیت مناسب برای وظایفRetrieval ، گزینه‌ای بهینه محسوب می‌شود.
- **ایندکس‌گذاری** :وکتورها با استفاده از کتابخانه **FAISS** و ایندکس IndexFlatL2 ذخیره می‌شوند که جستجوی دقیق و کامل را در پایگاه دانش تضمین می‌کند.

۳.۲ .بازیابی و تولید پاسخ(Inference)

- **Hybrid Retrieval:**سیستم به کاربران اجازه می‌دهد تا با اعمال فیلتر بر روی **Metadata**مانند category، فضای جستجوی معنایی را محدود کنند. این کار دقت و سرعت بازیابی اطلاعات را به شکل چشمگیری افزایش می‌دهد.
- **مدل LLM و Prompting:** برای تولید پاسخ نهایی، از مدل **mistralai/Mixtral-8x7B-Instruct-v0.1** استفاده می‌شود. یک **Prompt**مهندسی‌شده و دقیق، مدل را ملزم می‌کند تا پاسخ‌ها را **صرفاً بر اساس متن ارائه‌شده** تولید کرده و از دانش داخلی خود استفاده نکند. این تکنیک به طور مؤثری از تولید اطلاعات نادرست (Hallucination) جلوگیری می‌کند.

---

۴ .ویژگی‌های امتیازی و پیشرفته(Bonus Features)

علاوه بر پیاده‌سازی Pipeline اصلی، چندین ویژگی پیشرفته برای افزایش کارایی، کیفیت و قابلیت استفاده سیستم توسعه داده شده است:

- **مکانیسم:Re-ranking**

  پس از مرحله بازیابی اولیه، یک مدل **Cross-Encoder** (cross-encoder/ms-marco-MiniLM-L-6-v2)نتایج را مجدداً امتیازدهی و مرتب‌سازی می‌کند. این فرآیند دقت زمینه (Context) ارسالی به LLMرا به حداکثر می‌رساند و کیفیت پاسخ نهایی را به طور قابل توجهی بهبود می‌بخشد.

- **فراخوانی ناهمگام:(Asynchronous Calls) LLM**

  تمامی ارتباطات با API مدل زبان (LLM) با استفاده از کتابخانه httpxبه صورت **Asynchronous** پیاده‌سازی شده است. این ویژگی باعث می‌شود سرور FastAPI در هنگام انتظار برای دریافت پاسخ از LLMمسدود نشود و بتواند به درخواست‌های همزمان دیگر پاسخ دهد که برای یک سیستم واقعی امری حیاتی است.

- **مجموعه ارزیابی داخلی:(Evaluation Suite)**

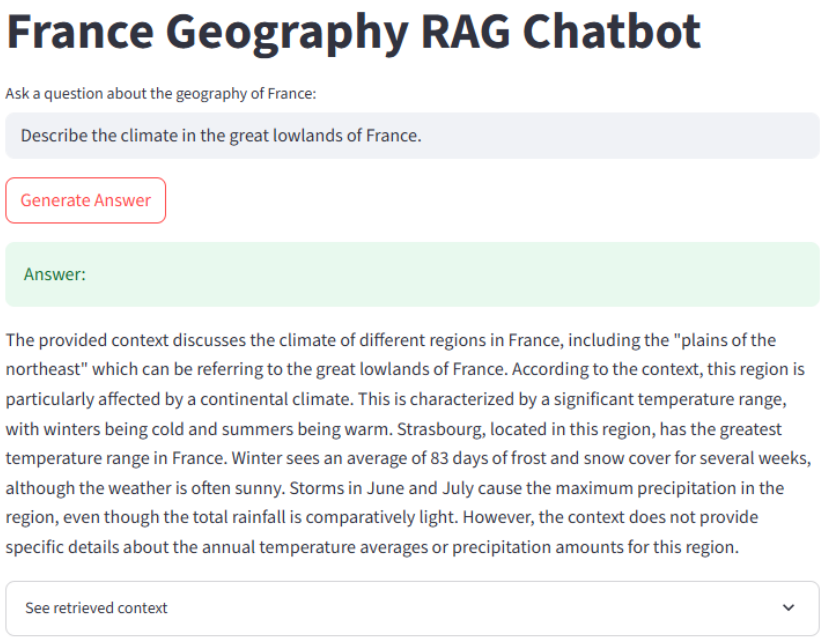
  یک اسکریپت ارزیابی اختصاصی (evaluate.py) برای سنجش کمی عملکرد سیستم طراحی شده است. این ابزار معیارهای کلیدی مانند **Context Recall**، **Faithfulness**و **Answer Similarity**را

محاسبه کرده و به ما امکان می‌دهد تأثیر تغییرات مختلف) مانند تغییر Prompt یا مدل (را به صورت داده‌محور تحلیل کنیم.

- *نتایج کلیدی:* ارزیابی نشان داد که سیستم در بخش **Generation** با Faithfulness ۸۰٪ (بسیار قوی عمل می‌کند، اما در بخش **Retrieval** با Context Recall ۲۰٪ (ضعف دارد که نقطه اصلی برای بهبودهای آینده است.

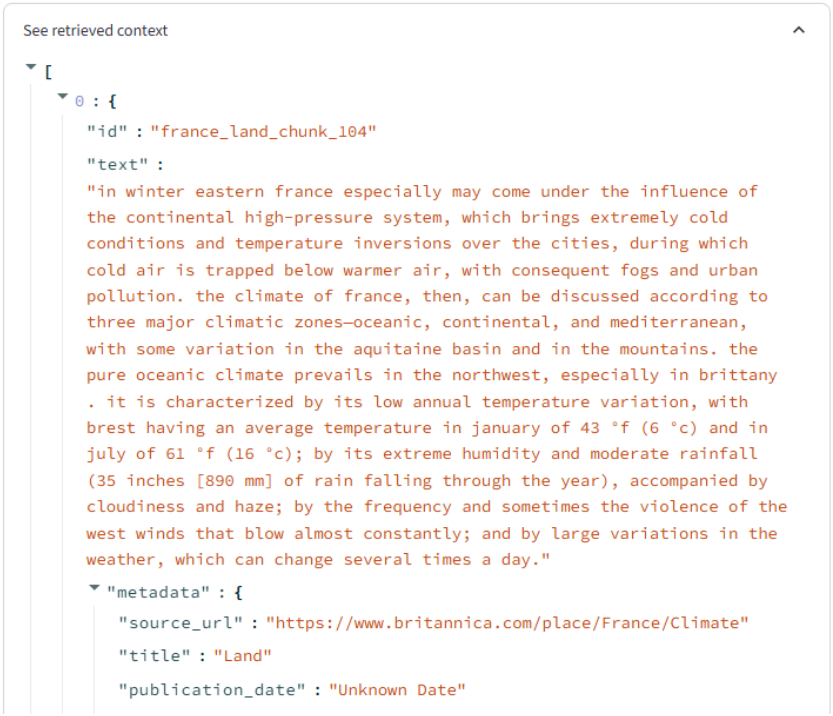
- **رابط کاربری تعاملی:(UI)**

یک رابط کاربری ساده و کاربرپسند با استفاده از **Streamlit** توسعه داده شده است. این UI به کاربران اجازه می‌دهد به راحتی با سیستم تعامل کرده، سوالات خود را بپرسند و پاسخ‌های تولید شده به همراه Context بازیابی‌شده را مشاهده کنند.



تصویر رابط کاربری- پاسخ مدل

specific details about the annual temperature averages or precipitation amounts for this region.



تصویر رابط کاربری - چانک های بازگشت داده شده