

# مستندات فنی پروژه: سیستم RAG برای پرسش و پاسخ جغرافیایی

تاریخ: ۲۶ اردیبهشت ۱۴۰۳

توسط: آرمان خلیلی - 4003623016

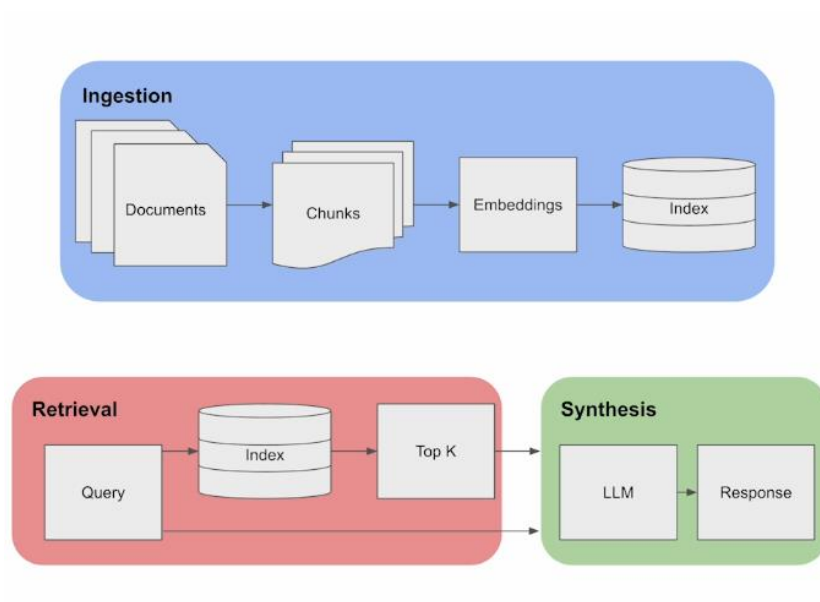
## ۱. معرفی کلی پروژه

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 وظایف Retrieval، 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. بازبایی 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 بازبایی 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 بازبایی اولیه، یک مدل `cross-encoder/ms-marco-MiniLM-L-6-v2` (Cross-Encoder) ارسالی به (Context) را مجدداً امتیازدهی و مرتب‌سازی می‌کند. این فرآیند دقت زمینه (Context) ارسالی به LLM را به حداکثر می‌رساند و کیفیت پاسخ نهایی را به طور قابل توجهی بهبود می‌بخشد.
- **فراخوانی ناهمگام: LLM (Asynchronous Calls)**  
تمامی ارتباطات با API مدل زبان (LLM) با استفاده از کتابخانه `httpx` به صورت **Asynchronous** پیاده‌سازی شده است. این ویژگی باعث می‌شود سرور `FastAPI` در هنگام انتظار برای دریافت پاسخ از LLM مسدود 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 ۸۰٪ Faithfulness (بسیار قوی عمل می‌کند، اما در بخش **Retrieval** با ۲۰٪ Context Recall (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 بازبینی شده را مشاهده کنند.

## France Geography RAG Chatbot

Ask a question about the geography of France:

Describe the climate in the great lowlands of France.

Generate Answer

Answer:

The provided context discusses the climate of different regions in France, including the "plains of the northeast" which can be referring to the great lowlands of France. According to the context, this region is particularly affected by a continental climate. This is characterized by a significant temperature range, with winters being cold and summers being warm. Strasbourg, located in this region, has the greatest temperature range in France. Winter sees an average of 83 days of frost and snow cover for several weeks, although the weather is often sunny. Storms in June and July cause the maximum precipitation in the region, even though the total rainfall is comparatively light. However, the context does not provide specific details about the annual temperature averages or precipitation amounts for this region.

See retrieved context



تصویر رابط کاربری - پاسخ مدل

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

See retrieved context

```
[
  {
    "id": "france_land_chunk_104"
    "text":
      "in winter eastern france especially may come under the influence of the continental high-pressure system, which brings extremely cold conditions and temperature inversions over the cities, during which cold air is trapped below warmer air, with consequent fogs and urban pollution. the climate of france, then, can be discussed according to three major climatic zones—oceanic, continental, and mediterranean, with some variation in the aquitaine basin and in the mountains. the pure oceanic climate prevails in the northwest, especially in brittany . it is characterized by its low annual temperature variation, with brest having an average temperature in january of 43 °f (6 °c) and in july of 61 °f (16 °c); by its extreme humidity and moderate rainfall (35 inches [890 mm] of rain falling through the year), accompanied by cloudiness and haze; by the frequency and sometimes the violence of the west winds that blow almost constantly; and by large variations in the weather, which can change several times a day."
    "metadata": {
      "source_url": "https://www.britannica.com/place/France/Climate"
      "title": "Land"
      "publication_date": "Unknown Date"
    }
  }
]
```

تصویر رابط کاربری - چانک های بازگشت داده شده