MAESTRÍA CIENCIA DE DATOS

TÓPICOS ESPECIALES II

Sistema de Recuperación Aumentada por Generación (RAG) para el Apoyo a Padres en la Educación Temprana

Andrea Aguilera

Mayo 2025

Resumen

Este trabajo presenta el desarrollo de un sistema de Recuperación Aumentada por Generación (RAG) especializado en responder preguntas relacionadas con la educación en la primera infancia. El sistema combina técnicas de recuperación semántica de documentos con un modelo de lenguaje grande (LLM), y está diseñado para asistir a padres y cuidadores con información clara, precisa y científicamente fundamentada. El corpus está compuesto por documentos oficiales en español e inglés, y las respuestas generadas son siempre en español. Se utilizó FAISS para la indexación, HuggingFace para los embeddings multilingües, y LLaMA 3 (70B) vía Groq como modelo generador. El sistema fue evaluado cuali-cuantitativamente y se discuten sus resultados y recomendaciones.

Planteamiento del problema

Los primeros 1000 días de vida son determinantes en el desarrollo físico, cognitivo y emocional de los niños. Durante este período, padres y cuidadores buscan orientación en diversos temas vinculados a la salud, estimulación y crianza. Sin embargo, la información accesible en internet suele ser confusa, contradictoria o carente de respaldo científico. Este proyecto propone un sistema automatizado de recuperación y generación de respuestas que ayude a resolver dudas comunes utilizando fuentes confiables, con un enfoque accesible y en idioma español.

Descripción del corpus

El corpus fue construido manualmente a partir de documentos públicos descargados desde sitios oficiales. Se incluyeron archivos PDF en español e inglés provenientes de:

- Guías y recomendaciones de la OMS y UNICEF
- Artículos científicos de bases como SciELO
- Sitios oficiales de salud como CDC, HealthyChildren.org y OPS
- Materiales educativos del Ministerio de Salud en Paraguay

Para evitar la pérdida de sentido en los textos y mejorar la coherencia de las respuestas generadas, los documentos del corpus fueron fragmentados en bloques de 500 caracteres con un solapamiento de 50 caracteres. Esta técnica permite preservar el contexto semántico, ya que evita cortar frases a la mitad o separar ideas relacionadas, asegurando que cada fragmento mantenga continuidad con el anterior. Además, el solapamiento facilita una recuperación más precisa, ya que el motor semántico (como FAISS) puede identificar fragmentos más coherentes al momento de responder preguntas. Finalmente, esta fragmentación controlada también permite optimizar el uso de modelos de lenguaje, que tienen un límite de tokens, asegurando que cada fragmento sea lo suficientemente corto para procesarse rápidamente, sin perder su significado individual.

Metodología

El sistema implementado se basa en una arquitectura de Recuperación Aumentada por Generación (RAG**)**, que combina técnicas de búsqueda semántica con modelos de lenguaje grandes (LLMs) para responder preguntas sobre la primera infancia utilizando un corpus documental fiable y multilingüe. El proceso puede dividirse en tres grandes componentes: recuperación semántica, generación de respuestas y herramientas de implementación.

Recuperación semántica de información

El sistema utiliza un motor semántico construido con **FAISS** (Facebook AI Similarity Search) y embeddings multilingües generados con el modelo "paraphrase-multilingual-MiniLM-L12-v2" de Hugging Face. Esta combinación permite encontrar fragmentos de texto relevantes no por coincidencia literal, sino por cercanía semántica.

¿Por qué FAISS?

- FAISS es una biblioteca especializada en búsqueda por similitud de vectores.
- Es ideal para sistemas RAG, ya que permite recuperar eficientemente los fragmentos más relevantes de grandes volúmenes de texto.
- Es de código abierto, ampliamente adoptada en la comunidad, y tiene excelente rendimiento.

¿Por qué embeddings multilingües?

- El corpus está compuesto por documentos en español e inglés, por lo que se seleccionó una representación vectorial que funcionara bien en ambos idiomas.
- El modelo "paraphrase-multilingual-MiniLM-L12-v2" fue diseñado para mapear frases en distintos idiomas con el mismo significado a vectores similares.
- Es liviano, compatible con FAISS y adecuado para ejecución en Google Colab.

Generación de respuestas

Una vez recuperados los fragmentos más relevantes, se utiliza el modelo de lenguaje LLaMA 3 (70B) haciendo uso del contexto proporcionado por el motor semántico.

¿Por qué LLaMA 3?

- LLaMA 3 es uno de los modelos de lenguaje más recientes y avanzados desarrollados por Meta.
- Ofrece alta capacidad de comprensión contextual y generación fluida de texto.

¿Por qué a través de Groq?

- Groq permite acceder gratuitamente a modelos avanzados como LLaMA 3, sin necesidad de consumir tokens como en otras plataformas comerciales.
- Ofrece velocidad de inferencia extremadamente alta (respuestas en milisegundos).
- Se integra fácilmente con LangChain, lo que facilita su implementación en sistemas RAG.

El modelo fue configurado para comportarse como un experto en educación en la primera infancia y para responder únicamente en español, independientemente del idioma del texto fuente.

Herramientas y configuración

Para el desarrollo del sistema se utilizaron, además, las siguientes herramientas del ecosistema de procesamiento de lenguaje natural:

- **LangChain**: framework que facilita la construcción modular de sistemas RAG, permitiendo combinar recuperación, contexto y generación de forma estructurada.
- **PyMuPDF**: utilizado para la lectura precisa de documentos PDF del corpus. Se eligió por su capacidad de conservar la estructura del texto mejor que otras alternativas.
- **HuggingFace Transformers**: para la gestión de modelos de embeddings.
- **Google Colab + Google Drive**: como entorno de desarrollo y almacenamiento.

Evaluación de resultados

La evaluación del sistema se realizó en dos niveles: **cualitativo** y **cuantitativo**.

Evaluación cualitativa

Se formularon consultas reales relacionadas con temas centrales de la primera infancia, como:

- ¿Qué cuidados básicos se deben tener en los primeros 1000 días de vida?
- ¿Cómo estimular el lenguaje en un bebé de 1 año?
- ¿Qué vacunas son obligatorias en los primeros dos años?
- ¿Cuál es la importancia de la lactancia?
- ¿Cuáles son las etapas por las que pasa un bebé antes de caminar?

Las respuestas generadas fueron revisadas manualmente y se observó que:

- Eran correctas y coherentes con el contenido del corpus.
- La integración de documentos en inglés no afectó negativamente la calidad de las respuestas en español, gracias al uso de embeddings multilingües.

Evaluación cuantitativa

Para medir objetivamente la calidad de las respuestas, se utilizaron las siguientes métricas:

- **BERTScore F1**: mide la similitud semántica entre la respuesta generada por el modelo y una respuesta de referencia (elaborada por un experto humano).
- **ROUGE-L F1**: evalúa la coincidencia literal y estructural, midiendo cuánto se parece la redacción a la respuesta de referencia en términos de frases o secuencias.

Se evaluaron cinco preguntas representativas, con los siguientes resultados:

	Pregunta	BERTScore F1	ROUGE-L F1
0	¿Qué cuidados básicos se deben tener en los pr...	0.746	0.309
1	¿Cómo estimular el lenguaje en un bebé de 1 año?	0.721	0.316
2	¿Qué vacunas son obligatorias en los primeros ...	0.802	0.419
3	¿Cuál es la importancia de la lactancia?	0.774	0.329
4	¿Cuáles son las etapas por la que pasa un bebé...	0.720	0.271

Como puede observarse, todas las respuestas obtuvieron valores de BERTScore superiores a 0.72, lo que indica una fuerte coherencia semántica con las respuestas esperadas.
Por otro lado, los valores de ROUGE-L F1, si bien más bajos, reflejan una redacción estructuralmente aceptable, considerando que el modelo no fue ajustado para imitar literalmente el estilo humano. Las mejores respuestas se observaron en las preguntas sobre lactancia y vacunas, tanto en contenido como en forma.

Conclusiones y recomendaciones

Este sistema demuestra que es posible construir un asistente especializado en una temática sensible como la educación temprana utilizando herramientas modernas y recursos de código abierto.

Se recomienda:

- Expandir el corpus con más documentos nacionales que respondan a contextos específicos de Paraguay.
- Desplegar el sistema como un chatbot o interfaz web amigable.