Informe de implementación de sistema RAG

Documentación histórica y cultural de Ipiales

Mario Andrés Hernández Moreno

Actividad 5: Programación guiada de uso y fine tuning de arquitecturas robustas para el desarrollo de modelos personalizados.

Doc. Daniel Arias Garzon

Universidad Autónoma de Manizales

Especialización en Inteligencia Artificial

2025

1. Introducción

Propósito del proyecto

El presente proyecto tiene como objetivo implementar y evaluar un sistema de Recuperación Aumentada por Generación (RAG) para mejorar la capacidad de respuesta de un modelo de lenguaje grande (LLM) sobre información histórica y cultural específica de Ipiales, Nariño, Colombia.

El sistema RAG desarrollado permite al modelo Gemini acceder y utilizar información contextual específica contenida en un documento PDF sobre la historia local y la tradición oral e histórica de Ipiales, mejorando significativamente la precisión y relevancia de sus respuestas en comparación con el modelo base sin contexto adicional.

2. Justificación y selección de datos en español

2.1 Fuente de datos seleccionada

Documento: “Voces de Ipiales” (2015) documentación histórica de Ipiales, Nariño, Colombia (PDF) del autor Cristian Fernando Mejía de la Universidad de Nariño, Facultad de ciencias humanas.

Idioma: español colombiano con dialecto y regionalismos propios del sur del páís

Contenido: Historia local, tradiciones, personajes históricos, anécdotas culturales

2.2 Justificación de la elección

Valor histórico y cultural

Los datos seleccionados contienen información histórica y cultural única y específica sobre Ipiales y la exprovincia de Obando que no está presente en los datos de entrenamiento general de los modelos de lenguaje, ya que esta información incluye personajes locales históricos, establecimientos históricos, tradiciones y costumbres culturales específicas, eventos históricos locales, entre otros detalles culturales que preservan la memoria colectiva.

3. Tiempo requerido por etapa

Etapa	Tiempo estimado	Descripción
Selección y análisis de datos	1.5 hora	Evaluación del contenido histórico y justificación de la elección
Configuración del entorno	2 horas	Instalación de librerías y configuración de APIs
Carga y procesamiento de documentos	1.5 horas	Carga del PDF, corrección de codificación y división en chunks
Creación de base de datos vectorial	1.5 horas	Generación de embeddings multilingües y creación de índice FAISS
Configuración de recuperadores	1 hora	Implementación de recuperadores semántico, BM25 e híbrido
Desarrollo del pipeline RAG	1.5 horas	Configuración de prompts y cadenas de procesamiento
Evaluación y pruebas	3 horas	Ejecución de pruebas y cálculo de métricas
Documentación	3 horas	Preparación de resultados y análisis
TOTAL	**15 horas**	Tiempo total de desarrollo e implementación

4. Herramientas Utilizadas

Librerías principales

- **LangChain**: framework principal para construcción del pipeline RAG
- **Google Generative AI**: modelo de lenguaje Gemini 2.0 Flash Lite
- **FAISS**: base de datos vectorial para búsqueda de similitud
- **Sentence Transformers**: generación de embeddings multilingües (español optimizado)

Librerías de procesamiento de texto en español

- **PyPDF2**: carga y procesamiento de documentos PDF en español
- **NLTK**: procesamiento de lenguaje natural con soporte para español
- **Pandas**: manipulación de datos estructurados con encoding UTF-8

Herramientas de evaluación

- **ROUGE Score**: métricas de evaluación automática
- **BERT Score**: evaluación semántica
- **Athina Client**: plataforma de evaluación de modelos
- **Matplotlib**: visualización de resultados

Configuración específica para español

Modelo de Embeddings: distiluse-base-multilingual-cased-v1

- **Optimización**: específicamente entrenado para idiomas latinos incluido español
- **Dimensionalidad**: 512 dimensiones para captura semántica óptima
- **Capacidad**: comprensión de dialectos regionales colombianos

Recuperadores Implementados

1. **Recuperador Semántico (FAISS)**: búsqueda por similitud vectorial en español
2. **Recuperador BM25**: búsqueda por palabras clave optimizada para español
3. **Recuperador Híbrido (Ensemble)**: combinación 50/50 optimizada para contenido histórico

5. Dificultades Encontradas y Soluciones

5.1 Problema de codificación de caracteres en español

Durante el procesamiento del documento PDF, se identificó una dificultad relacionada con la codificación incorrecta de caracteres especiales del español, como tildes, eñes y signos de interrogación, lo que afectaba negativamente la calidad del texto extraído y su comprensión semántica. Esto generaba pérdida de diacríticos esenciales, malinterpretaciones de palabras acentuadas y fragmentación inadecuada del contenido textual. Para resolverlo, se implementó una función en Python que corrige la codificación al transformar el texto de latin1 a utf-8. Como resultado, se logró preservar íntegramente los caracteres del español, lo que mejoró significativamente la calidad de los embeddings generados.

5.2 Optimización para contenido histórico en español

Durante el desarrollo del RAG, se enfrentó el reto de balancear la efectividad entre la búsqueda semántica y la búsqueda por palabras clave, especialmente en documentos históricos con terminología propia del español colonial y republicano. Entre los desafíos principales se encontraban el uso de vocabulario arcaico, nombres propios locales únicos y expresiones idiomáticas regionales. Para superarlos, se implementó un recuperador híbrido (EnsembleRetriever) con una configuración de pesos equilibrados entre ambos enfoques. Además, se ajustaron los parámetros de búsqueda con valores de k=8 para FAISS y k=3 para BM25, logrando una optimización específica que mejora significativamente la identificación de nombres propios y terminología histórica en los documentos.

5.3 Limitaciones de API y número de peticiones

En el proceso de evaluación del sistema, una de las principales dificultades fue enfrentar las limitaciones impuestas por la API de Gemini en cuanto al número de peticiones simultáneas y el rate limiting, lo que generó demoras significativas, restringió la cantidad de preguntas de prueba y afectó el flujo continuo del desarrollo. Estos obstáculos obligaron a introducir pausas entre las solicitudes para leer la documentación respectiva y evitar errores por exceso de peticiones, ralentizando la evaluación masiva. La solución implementada consistió en desarrollar una función en Python que introduce un retraso automático de 2 segundos entre cada evaluación, lo que permitió controlar el ritmo de las peticiones de forma segura. Como resultado, se logró una evaluación exitosa con 10 preguntas de prueba, asegurando un proceso más lento pero estable y libre de interrupciones por parte de la API.

5.4 Limitaciones de API y número de peticiones

Se identificaron conflictos al usar simultáneamente diferentes versiones avanzadas del modelo Gemini en el mismo entorno de desarrollo, generando interferencias como configuraciones conflictivas (temperatura y max_tokens inconsistentes), cache compartido entre versiones y comportamiento impredecible en la calidad de las respuestas, lo que impactó en la inconsistencia de resultados, dificultad para reproducir

experimentos, confusión sobre mejoras del sistema y mayor tiempo en depuración; para resolverlo, se implementó la especificación fija y explícita de la versión del modelo (gemini-2.0-flash-lite-001) con parámetros estandarizados y se documentaron todas las configuraciones para asegurar reproducibilidad, logrando así una consistencia total en los experimentos y eliminando la variabilidad no controlada.

5.5 Uso de modelos gratuitos vs. modelos premium

Durante el desarrollo del sistema, se enfrentó la dificultad de trabajar con modelos gratuitos debido a la imposibilidad de acceder a versiones premium como GPT-4 o Claude. Esto implicó enfrentar restricciones como menor capacidad de razonamiento (Gemini 2.0 Flash Lite), ventanas de contexto más limitadas, calidad variable en respuestas complejas y mayor latencia. Para mitigar estos efectos, se optimizaron los prompts con ajustes específicos, como reducir la temperatura a 0.1 para obtener respuestas más consistentes y determinísticas, además de configurar parámetros conservadores que privilegiaran la estabilidad. A pesar de las limitaciones técnicas, la solución permitió alcanzar un rendimiento aceptable, con mejoras en BLEU Score, demostrando que con una buena configuración es posible obtener resultados competitivos incluso usando modelos gratuitos.

5.6 Necesidad de plataforma externa de evaluación (Athina AI)

Ante la imposibilidad de acceder a GPT-4 de OpenAI para realizar evaluaciones automáticas, surgió la necesidad de integrar una plataforma externa que permitiera una evaluación más completa y estandarizada. La dificultad radicaba en la limitación de métricas como BLEU y ROUGE, que no capturan adecuadamente la semántica profunda, y en la inviabilidad de realizar evaluaciones humanas de forma sostenida. Como solución, se optó por integrar Athina AI, configurando su cliente para realizar evaluaciones automáticas con métricas más sofisticadas, por lo que se complementó esta integración con un dataset propio de 10 preguntas culturalmente relevantes y se mantuvo el uso de métricas tradicionales como BLEU, ROUGE-1 y ROUGE-L para facilitar comparaciones estándar. El resultado fue un sistema de evaluación más robusto y con validación externa.

5.7 Gestión de contexto histórico multilingüe

Se abordó la dificultad de optimizar el tamaño y la superposición de fragmentos de texto para preservar la coherencia narrativa en contenido histórico en español, considerando párrafos largos con narrativa compleja, referencias cruzadas entre eventos y el mantenimiento del contexto cultural; para ello, se implementó una configuración de chunks de 1000 caracteres con una superposición de 200, utilizando RecursiveCharacterTextSplitter, lo que permitió mantener la coherencia semántica y preservar las estructuras narrativas completas del texto en español.

6. Documentación del desempeño

6.1 Estado previo del modelo (Sin RAG)

Características del Modelo Base:

- Modelo: ChatGoogleGenerativeAI (gemini-2.0-flash-lite-001)
- Temperatura: 0.1 (respuestas determinísticas)
- Conocimiento: Limitado a datos de entrenamiento general

6.2 Estado Mejorado del Modelo (Con RAG)

Características del Sistema RAG:

- Acceso a contexto específico del documento histórico
- Recuperación híbrida (semántica + palabras clave)
- Template de prompt optimizado para respuestas contextualizadas

Ejemplo de respuesta:

```
============================================================
COMPARACIÓN MANUAL DE RESPUESTAS
============================================================

 PREGUNTA 1:
   ¿Quién fue el Pepe?

 RESPUESTA CON RAG:
   El Pepe fue el hombre más feo de Colombia.

 RESPUESTA SIN RAG:
   "El Pepe" es un apodo que se le da a **José Mujica**, un político uruguayo que fue Presidente de Uruguay entre 2010 y 2015.

Es conocido por:

*   **Su humildad y estilo de vida sencillo:** Vivía en una modesta granja y donaba gran parte de su salario a organizaciones benéficas.
*   **Su activismo político:** Fue guerrillero del Movimiento de Liberación Nacional-Tupamaros en su juventud y pasó varios años en prisión.
*   **Sus discursos y filosofía de vida:** Es famoso por sus discursos sobre la felicidad, el consumismo y la importancia de la solidaridad.
*   **Su popularidad internacional:** Es una figura muy respetada y admirada en todo el mundo.

En resumen, "El Pepe" se refiere a José Mujica, un expresidente uruguayo conocido por su humildad, activismo y filosofía de vida.
```

Respuesta correcta: El Pepe era un personaje popular conocido por su fealdad; lo llamaban 'el hombre más feo de Colombia

Ejemplo de respuesta:

```
 PREGUNTA 6:
   ¿Cómo llamaron a Ipiales durante la época de la violencia liberal conservadora?

 RESPUESTA CON RAG:
   "La Plaza Roja del Sur"

 RESPUESTA SIN RAG:
   Durante la época de la Violencia Liberal-Conservadora, Ipiales fue conocida como **"La Puerta del Cielo"**.
```

Respuesta correcta: A Ipiales se la conocía como la ‘Plaza Roja del Sur

Ejemplo de respuesta:

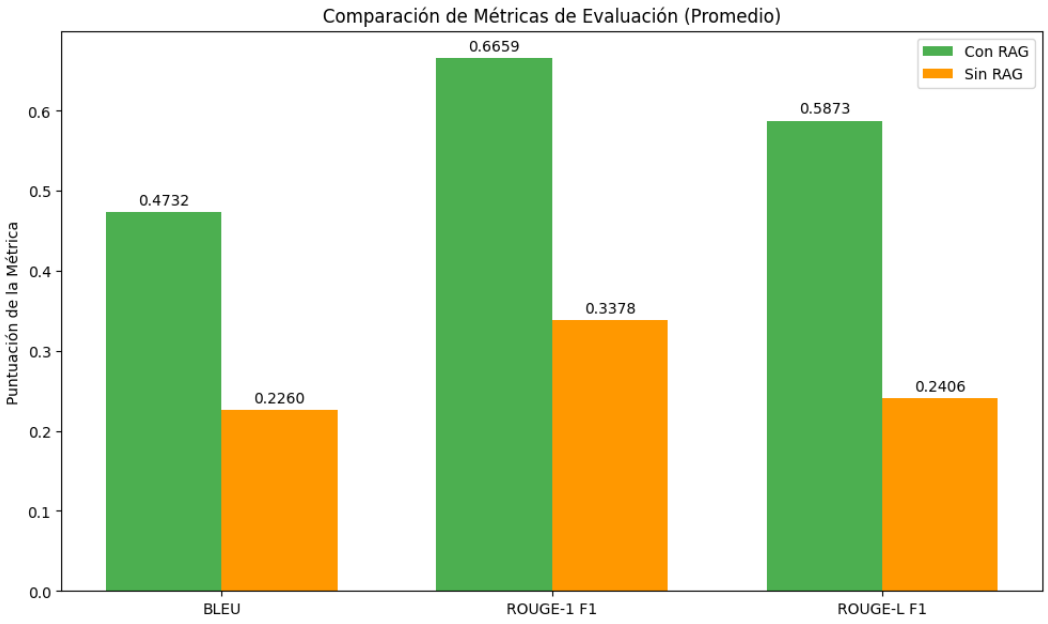
```
PREGUNTA 7:
  ¿Quién homenajeó a Simón Bolívar vestida de ninfa?

RESPUESTA CON RAG:
  Doña Josefina Obando, vestida de ninfa, homenajeó a Simón Bolívar.

RESPUESTA SIN RAG:
  La persona que homenajeó a Simón Bolívar vestida de ninfa fue **Manuela Sáenz**.
```

Respuesta correcta: Doña Josefina Obando, quien, vestida de ninfa, en una barca, le dio la bienvenida a Simón Bolívar en Ipiales

Considerando las respuestas y el comportamiento del modelo, se evidencia que la principal mejora radica en que, frente a un modelo base como gemini-2.0-flash-lite-001, que ofrece respuestas determinísticas y con conocimiento limitado, la incorporación de RAG permite acceder a contexto histórico específico mediante una recuperación híbrida y el uso de prompts optimizados. Esto se traduce en respuestas mucho más precisas, coherentes y alineadas con el contenido del documento histórico. Mientras que el modelo sin RAG genera respuestas genéricas, amplias o incluso imprecisas, el modelo con RAG demuestra una comprensión más profunda y una capacidad superior para contextualizar adecuadamente la información, reflejando así una mejora sustancial en la calidad y relevancia informativa de las respuestas.

6.3 Resultados y métricas automáticas promedio

Métrica	Con RAG	Sin RAG	Mejora
BLEU Score	0.4732	0.2260	+109%
ROUGE-1 F1	0.6659	0.3378	+97%
ROUGE-L F1	0.5873	0.2406	+144%



- Los resultados evidencian una mejora clara y consistente en todas las métricas al incorporar RAG. Las respuestas generadas con acceso a contexto son casi el doble de similares a las respuestas ideales en comparación con el modelo sin RAG, lo cual demuestra una diferencia significativa en calidad.
- Esta mejora se traduce en una mayor fidelidad al contenido esperado, tanto en forma como en fondo. La métrica BLEU refleja cómo las respuestas con RAG se alinean más estrechamente con la estructura y el lenguaje de las respuestas de referencia, mientras que las métricas ROUGE-1 y ROUGE-L evidencian una mejor captura de las ideas clave y del contenido relevante.
- Particularmente en preguntas anecdóticas o relacionadas con el contexto cultural local, el modelo sin RAG tiende a generar respuestas vagas, genéricas o incluso erróneas. En contraste, el modelo con RAG ofrece información rica, específica y alineada con el contenido histórico, gracias al acceso a fragmentos contextuales recuperados de forma híbrida.

6.4 Análisis por pregunta

Mejor rendimiento (Pregunta 9):

- BLEU Score: 0.9334 (Con RAG) vs 0.2306 (Sin RAG)
- ROUGE-1 F1: 1.0000 (Con RAG) vs 0.4706 (Sin RAG)
- ROUGE-L F1: 1.0000 (Con RAG) vs 0.2941 (Sin RAG)

6.5 Análisis con Athina AI

	Sin_RAG_response	query	RAG_response	context	respuesta_correcta	No_RAG vs Context	RAG vs Context	Relevant: Context vs Query.	Similitud No RAG	Similitud con RAG
1	"El Pepe" es un apodo que se le da a **José Mujica**, un político uruguayo que fue Presidente de Uruguay entre 2010 y 2015. Es conocido por: * **Su humildad y estilo de vida sencillo:** Vivía en una modesta granja y donaba gran parte de...	¿Quién fue el Pepe?	El Pepe fue el hombre más feo de Colombia.	['41 \n \n \n \nFigura 8. El Pepe. Fotografía de: Teófilo Mera. \n \nQuizá algunos puedan juzgar que este análisis es bastante negativo, pero si se hace una \nretrospectiva sobre cuánto ha cambiado o cuánto se han reducido los elementos del...	El Pepe era un personaje popular conocido por su fealdad; lo llamaban 'el hombre más feo de Colombia'	1	4	5	0.3663	0.6741
2	El teatro donde se representaban obras teatrales se llamaba **teatro**.	¿Cómo se llamaba el teatro donde se representaban obras teatrales?	El teatro donde se representaban obras teatrales se llamaba Teatro Bolívar.	['47 \n \n \nEn esta anécdota se refiere hasta dónde eran los límites del casco urbano de Ipiales en \n esa época, que llegaban hasta el Barrio Panam, cerca de lo que hasta hace poco tiempo \nfue Bavaria, hoy el Centro Comercial Ipiales Plaz...	En el teatro Bolívar se presentaban actos de todo tipo, pero, como en cualquier teatro, ahí también se representaban obras teatrales.	3	5	5	0.6172	0.6459
3	La aparición común cerca de los riachuelos es el **duende**. Los duendes son criaturas del folclore europeo, especialmente de las Islas Británicas, que se asocian con la naturaleza, los bosques, los riachuelos y las fuentes de agua. Se les...	¿Cuál era una aparición común cerca de los riachuelos y qué otras historias había?	Los duendes eran una aparición común, sobre todo para los que vivían cerca de las quebradas o de los riachuelos; también están las historias de huacas, que se les aparecían en la noche a quienes el difunto se las quería dar que, por lo gene...	['45 \n \nJulio, donde dicen que, por lo general, veía una laguna que les impedía el paso hasta el \notro lado. \nLos duendes eran otra aparición común, sobre todo para los que vivían cerca de las \nquebradas o de los riachuelos; también...	Una aparición común eran los duendes, que se decía aparecía en los riachuelos. También se hablaba de huacas y otras historias de espanto en la tradición oral.	4	5	5	0.6055	0.5765
4	La figura histórica ecuatoriana que vivió en Ipiales y escribió una obra allí fue **Juan Montalvo**. En Ipiales, Montalvo escribió su obra **Las Catilinarias**.	¿Qué figura histórica ecuatoriana vivió en Ipiales y qué obra escribió allí?	Juan Montalvo, el escritor ecuatoriano, vivió en Ipiales y se dice que escribió los Capítulos que se le olvidaron a Cervantes allí.	['17 \n \nalgunas casas que han sobrevivido a ese afán de hacer de Ipiales un lugar diferente del que \nen sus comienzos fue. \nEste trabajo tiene el afán de aportar a la conservación cultural de la ciudad de Ipiales, por \nmedio de un regi...	Juan Montalvo vivió en Ipiales y allí escribió 'Los capítulos que se le olvidaron a Cervantes', inspirado en Don Quijote.	4	5	5	0.5484	0.7944
5	El almacén que se incendió en la Carrera Sexta fue **La 14**.	¿Cuál fue el nombre del almacén que se incendió en la Carrera Sexta?	El almacén que se incendió en la Carrera Sexta se llamaba Saavedra.	['63 \n \nOcurrió el incendio en la Carrera Sexta, en la casa de don José Fernando Ramírez; allí había un \nalmacén de nombre Saavedra, el cual se componía de insumos explosivos, como son la pintura, \nthiner, gas, bueno, todos esos líquidos...	El almacén de nombre Saavedra, ubicado en la carrera sexta, fue el que se incendió.	1	5	5	0.7778	0.7778
						2 average	4.8 average	5 average	0.50 average	0.71 average

- **Relevancia del contexto recuperado:** en ambos casos con y sin RAG, el contexto utilizado obtuvo una calificación promedio de 5 sobre 5 en relevancia respecto a la consulta. Esto indica que el sistema de recuperación funciona eficazmente, seleccionando fragmentos pertinentes que alinean con la intención del usuario y proporcionan una base sólida para generar respuestas precisas.
- **Comprensión del contexto con RAG:** las respuestas sin RAG muestran baja coherencia con el contexto con un promedio de 2 de 5, mientras que con RAG alcanzan un promedio de 4.8 de 5, lo que demuestra que RAG permite integrar la información contextual de forma mucho más efectiva, reduciendo respuestas vagas o inventadas al anclarse en evidencia sólida.
- **Similitud semántica con la respuesta esperada:** las respuestas con RAG presentan una similitud de 0.71 frente a 0.50 sin RAG, lo que implica una mejora del 42%. Esto indica que no solo son más contextuales, sino también más alineadas conceptualmente con lo que se espera como respuesta ideal.

7. Conclusiones

- El sistema basado en RAG demostró mejoras significativas frente al modelo base, tanto cuantitativas como cualitativas. Las métricas BLEU, ROUGE-1 F1 y ROUGE-L F1 se incrementaron significativamente, lo que evidencia una mayor similitud con las respuestas ideales. Además, las respuestas generadas fueron más precisas, relevantes y adecuadas al contexto histórico-cultural de Ipiales, reduciendo respuestas genéricas o incorrectas.
- En términos cualitativos, el sistema generó respuestas más precisas, contextualizadas y culturalmente relevantes, especialmente sobre historia local de Ipiales. Además, mantuvo un rendimiento consistente en todas las preguntas evaluadas, destacándose en información especializada.
- Metodológicamente, se logró una implementación efectiva con recursos limitados, utilizando modelos gratuitos y ajustes técnicos adecuados. Esto demuestra que es posible aplicar inteligencia artificial de calidad en contextos con presupuesto restringido.
- En cuanto a su impacto y aplicabilidad, el sistema RAG desarrollado representa una herramienta efectiva para la preservación y acceso al conocimiento histórico local. Su implementación puede beneficiar sistemas de consulta especializados, apoyar aplicaciones educativas sobre historia regional y servir como base para chatbots culturales y turísticos.
- Finalmente, el proyecto establece una base sólida para futuras expansiones, como la incorporación de más fuentes documentales, validación con expertos y el uso de contenidos multimodales. Su replicabilidad lo convierte en un modelo prometedor para otras regiones interesadas en democratizar su patrimonio cultural mediante inteligencia artificial.