Integrantes: Sofía Villamizar, Juan Sebastián Sierra, Nicolás Rich y Sofía Anzola Ortegón

RAG: Asesor de Crédito Agropecuario para Caficultores en Colombia

1. Introducción y descripción general del proyecto

El proyecto consiste en el diseño y construcción de un sistema de Recuperación Aumentada por Generación (RAG) capaz de brindar asesoría crediticia personalizada a pequeños y medianos caficultores. Los productores rurales, en especial los cafeteros, suelen tener dificultades para acceder a información clara sobre líneas de financiamiento, tasas, plazos, garantías e incentivos gubernamentales. También enfrentan barreras geográficas, educativas y tecnológicas que complican aún más la toma de decisiones financieras.

El sistema busca cubrir esas necesidades ofreciendo respuestas claras y basadas exclusivamente en documentos oficiales y técnicos. Para ello se construyó un RAG que combina documentos de entidades financieras y estudios técnicos, como formularios del Banco Agrario, información sobre el Fondo Agropecuario de Garantías (FAG), documentos sobre drones de aspersión regulados por la Aeronáutica Civil y estudios de costos recientes del cultivo, con modelos de lenguaje avanzados capaces de generar explicaciones accesibles para el usuario final.

2. Dataset y recolección de documentos

El sistema se alimenta de documentos auténticos utilizados por el sector. Entre ellos se encuentran:

- Formularios y guías del Banco Agrario que explican cómo solicitar crédito, qué requisitos se deben cumplir, qué garantías exige el FAG y cuáles son los plazos y esquemas de amortización permitidos.

- Documentos técnicos de Cenicafé y la Aeronáutica Civil sobre drones de aspersión, donde se describen sus características, regulaciones, parámetros de vuelo y ventajas en el manejo de la broca del café.

- Estudios de costos de producción del café que reportan información sobre fertilización, recolección, rendimientos y estructura de costos del cultivo.

Estos archivos fueron cargados manualmente en Google Colab a través del botón “Choose Files”, que aparece al correr el código. Todos los documentos están almacenados en el GitHub, después de cargar los 14 pdf, el sistema los procesó para incorporarlos al RAG.

3. Procesamiento de documentos

Cada PDF fue leído página por página utilizando la librería **pypdf.** El texto extraído se dividió en fragmentos más pequeños para facilitar la búsqueda posterior. Este proceso, llamado “chunking”, consiste en cortar el texto en partes de aproximadamente 500 caracteres con una superposición de 100. Esto garantiza que los fragmentos sean suficientemente

amplios para contener ideas completas, pero lo bastante pequeños para permitir respuestas precisas y relevantes.

Al dividir los documentos de esta manera, se consiguió que piezas importantes de información, como condiciones de crédito, descripciones técnicas o detalles sobre costos agrícolas, estén disponibles para su recuperación exacta cuando el usuario realice una consulta.

4. Embeddings y almacenamiento en FAISS

Para transformar el texto en representaciones numéricas que un modelo pueda comparar, se utilizó el modelo de embeddings **"paraphrase-multilingual-MiniLM-L12-v2"**, que en la arquitectura corresponde al **encoder** del sistema. Este encoder es liviano, rápido y funciona muy bien con textos en español, lo cual es fundamental para procesar documentos rurales colombianos y preguntas de caficultores.

Este modelo toma cada fragmento y lo convierte en un vector que representa su significado. Todos estos vectores se almacenaron en un índice FAISS, una base de datos especializada para buscar información por similitud. De esta forma, cuando un caficultor hace una pregunta, el sistema puede encontrar rápidamente los fragmentos más parecidos en contenido al texto de su consulta. El uso de MiniLM y FAISS permite que el sistema funcione de manera eficiente incluso en computadoras con pocos recursos, como las que se usan en Colab.

5. Reranker: selección fina de los mejores fragmentos

Aunque FAISS encuentra fragmentos relevantes, no siempre el primer resultado es el más adecuado. Para mejorar la precisión, el sistema incorpora un segundo modelo llamado **"mixedbread-ai/mxbai-rerank-base-v1"**, que reordena los fragmentos recuperados de FAISS evaluando cuál responde mejor a la pregunta. Este reranker actúa como un filtro más estricto y garantiza que al modelo generador solo se le entreguen fragmentos realmente útiles.

Esta etapa es clave porque muchos documentos contienen tablas, numerales o explicaciones largas y técnicas. El reranker ayuda a seleccionar únicamente lo esencial para contestar la consulta.

6. Generación de respuestas con el modelo Mistral 7B

Una vez seleccionados los fragmentos más relevantes, se construye un prompt que combina la pregunta del usuario, las instrucciones para el modelo y el contexto recuperado. Para responder se utiliza el modelo **"Mistral-7B-Instruct-v0.3"**, que corresponde al **decoder** dentro de la arquitectura del sistema.

Este decoder fue elegido por varias razones: ofrece un excelente rendimiento en español, sigue instrucciones de manera precisa y puede generar explicaciones claras sin inventar información cuando se le restringe a usar solo el contexto. Además, su versión cuantizada en

4 bits reduce el consumo de memoria, lo cual permite ejecutarlo en Google Colab a pesar de su tamaño. El prompt instruye al modelo para que responda solamente con la información presente en los fragmentos, evitando interpretaciones creativas o datos no respaldados. También se le indica que use español colombiano y que explique de forma clara para un usuario rural.

Este modelo es capaz de generar respuestas completas, coherentes y fáciles de entender, así como de citar cada fragmento utilizado.

7. Flujo general del sistema RAG

El funcionamiento del sistema se puede describir en seis pasos: primero, el usuario realiza una pregunta en lenguaje natural. Luego, esa pregunta se transforma en un vector mediante el encoder MiniLM. FAISS recupera los fragmentos más similares y el reranker selecciona los más pertinentes. Con esta información se construye un prompt que incluye la pregunta, el contexto y las instrucciones. Finalmente, el decoder Mistral 7B genera la respuesta en lenguaje natural, acompañada de las fuentes utilizadas.

Este flujo completo garantiza que el sistema responda de forma transparente, citando la evidencia y sin inventar información.

8. Limitaciones y consideraciones

Aunque la arquitectura es robusta, el proyecto presenta algunas limitaciones. Varios documentos no incluyen tasas vigentes o información financiera actualizada, por lo que el sistema debe responder que esos datos no están en el contexto. Además, algunos archivos poseen estructuras complejas o están escaneados, lo que afecta la extracción de texto. Estas limitaciones no provienen de la arquitectura del RAG, sino de las características del dataset disponible.

9. Conclusión

Este proyecto demuestra cómo un sistema RAG puede convertirse en una herramienta de apoyo para productores rurales, integrando información técnica, financiera y operativa en un solo asistente accesible. La arquitectura construida, basada en un encoder eficiente (MiniLM), un índice FAISS optimizado, un reranker para mejorar la precisión y un decoder moderno como Mistral 7B, permite entregar respuestas claras y contextualizadas para caficultores que buscan orientarse sobre opciones de crédito, tecnologías como drones o costos de producción. Al tener trazabilidad de cada fragmento utilizado y evitar generar información no respaldada, el sistema se convierte en un aliado confiable para apoyar decisiones productivas y financieras en el sector cafetero colombiano.

10. Bibliografía

- Arcila-Moreno, A., & Benavides-Machado, P. (2025, abril). Drones de aspersión en la caficultura para el manejo integrado de la broca.
- Banco Agrario de Colombia. (2024). 2.2 Requisitos para crédito con recursos FINAGRO pequeño productor agropecuario - joven rural - mujer rural bajos ingresos.
- Calle, A. J., & Proescher, M. (2024, septiembre). Los retos financieros y la oferta de productos y servicios financieros para el sector agropecuario en Colombia.
- Estrada, D. (2019, abril). Financiamiento agropecuario y rural.
- FINAGRO. (n.d.). Justificación para asignación del FAG.
- FINAGRO. (2006, marzo). Manual de servicios FINAGRO.
- FINAGRO. (2017). Manual de servicios FINAGRO: Capítulo sexto. Línea de redescuento "A Toda Máquina e Infraestructura".
- FINAGRO. (2020). Manual de servicios FINAGRO: Título cuarto. Líneas especiales de crédito (LEC) con tasa subsidiada.
- FINAGRO. (2025a). Manual de servicios FINAGRO: Título primero. Crédito agropecuario y rural.
- FINAGRO. (2025b). Manual de servicios FINAGRO: Título tercero. Incentivos.
- Ministerio de Agricultura y Desarrollo Rural & Agencia de Desarrollo Rural. (2019, mayo 10). Manual operativo: Clasificación y registro de usuarios del servicio público de extensión agropecuaria.
- Secretaría técnica. (2023). Costos de producción de café. Colombia 2022.
- Secretaría técnica. (2024). Costos de producción de café 2023. Colombia.
- Sistema Nacional de Crédito Agropecuario: Propuesta de reforma. (2014, diciembre).