Informe Técnico - Proyecto final: Implementación de un Sistema RAG (Retrieval-Augmented Generation)

Integrantes: Andrade Dilan

Calahorrano David

Zambrano Andrés

Fecha de entrega: 13 de febrero de 2025

---

1. Resumen del proyecto

El presente informe describe el desarrollo de un Sistema RAG (Retrieval-Augmented Generation), el cual integra técnicas de Recuperación de Información (RI) con Modelos de Generación de Texto para responder consultas basadas en documentos relevantes.

El sistema fue diseñado para recibir consultas en lenguaje natural, recuperar información relevante almacenada en una base de datos y generar respuestas utilizando un modelo de inteligencia artificial. Para su implementación, se utilizaron diversas técnicas avanzadas de Procesamiento de Lenguaje Natural (PLN), incluyendo:

- **Preprocesamiento de datos:** Limpieza, tokenización, lematización y normalización del corpus.
- **ChromaDB como base de datos vectorial:** Permite almacenar documentos en embeddings y realizar consultas eficientes.
- Generación de embeddings con Sentence Transformers (all-MiniLM-L6-v2).
- **Generación de respuestas con T5 (Flan-T5-Large):** Modelo de lenguaje que produce respuestas basadas en los documentos recuperados.
- **Desarrollo de una API con Flask:** Permite la interacción del usuario mediante una interfaz web.

La evaluación del sistema se realizó con métricas de recuperación como Precision@k, Recall y F1-Score, y métricas de generación como BLEU y ROUGE-L, garantizando la efectividad del modelo en ambas tareas.

2. Descripción del corpus seleccionado

El corpus utilizado se compone de documentos textuales relacionados con los planes de trabajo de los candidatos presidenciales en Ecuador y transcripciones de entrevistas.

Para estructurar mejor la información, el corpus se dividió en:

- **Raw corpus**: Contiene los documentos en su estado original.
- **Processed corpus:** Incluye textos preprocesados tras limpieza, tokenización y lematización.

Se aplicaron técnicas de transcripción de audio a texto, segmentación del contenido y normalización de texto para mejorar la calidad de los datos antes de generar embeddings.

3. Metodología utilizada

El desarrollo del sistema se basó en la implementación de tres módulos principales:

- **Preprocesamiento del Corpus**
  El preprocesamiento fue clave para mejorar la calidad del texto antes de ser utilizado en los modelos. Se realizaron los siguientes pasos:
  - **Eliminación de ruido:** Se corrigieron errores tipográficos y se eliminaron caracteres especiales y palabras sin relevancia semántica.
  - **Tokenización y lematización**: Se procesó el texto para segmentarlo y reducir las palabras a su forma raíz.
  - **Conversión de texto a embeddings:** Se utilizó Sentence Transformers (all-MiniLM-L6-v2) para convertir los documentos en representaciones vectoriales.
- **Módulo de Recuperación de Información**
  El módulo de recuperación se encargó de buscar y seleccionar documentos relevantes a partir de la consulta del usuario.
  - **Implementación de ChromaDB:** Se utilizó como base de datos vectorial, almacenando los documentos en forma de embeddings.
  - **Conversión de consultas a embeddings:** Cada consulta se convierte en un vector numérico y se compara con los documentos almacenados en la base de datos.
  - **Recuperación basada en similitud semántica:** Se utiliza una métrica de distancia para identificar los documentos más relevantes.
  - **Evaluación del módulo de recuperación:** Se aplicaron métricas como Precision@k, Recall y F1-Score para evaluar la precisión del sistema.
- **Módulo de Generación de Respuestas**
  El módulo de generación transforma la información recuperada en respuestas coherentes y comprensibles.
  - **Selección de fragmentos relevantes:** Los documentos más relevantes son analizados y se extraen fragmentos clave.
  - **Generación de respuestas con T5 (Flan-T5-Large):** Se utiliza este modelo para estructurar respuestas basadas en la información recuperada.
  - **Optimización del modelo:** Ajuste de hiperparámetros como temperatura y top-k sampling para mejorar la coherencia.
  - **Evaluación de la generación de texto:** Se aplicaron métricas como BLEU y ROUGE-L para medir la calidad de las respuestas generadas.

4. Resultados obtenidos

Evaluación de la Recuperación

Métrica	Valor	Explicación
Precisión	63.40%	Indica qué porcentaje de la información recuperada es relevante. En este caso, el 63.4% de los datos recuperados fueron correctos en relación con la respuesta esperada.
Recall	79.60%	Mide qué porcentaje de la información relevante esperada fue recuperada. Un valor del 79.6% significa que el sistema logró recuperar la mayoría de los datos clave, pero aún falta optimizar la recuperación total.
F1-Scroe	70.60%	Es el equilibrio entre precisión y recall. Con un 70.6%, el sistema demuestra un desempeño aceptable, aunque puede mejorar en la exactitud de los resultados.

Evaluación de la generación

Métrica	Valor	Explicación
BLEU Score	0.6036	La respuesta generada tiene un 60.36% de similitud con la respuesta esperada en términos de palabras y frases coincidentes. Un valor entre 0.60 y 0.75 se considera aceptable para generación de texto.
ROUGE-L Score	0.7581	El 75.81% de la estructura del texto generado coincide con la respuesta esperada. Esto indica que la respuesta generada mantiene la coherencia y fluidez, aunque puede mejorarse.

Ejemplo de consulta y respuesta generada

- **Consulta:** "Dame información acerca de Noboa".
- **Documentos recuperados:** Discursos del candidato, entrevistas en medios de comunicación y análisis de expertos.
- **Respuesta generada:** "El conversatorio abordó temas clave como democracia y participación ciudadana, destacando la importancia del voto informado y la necesidad de fortalecer la institucionalidad. En seguridad, se discutieron la sobrepoblación carcelaria, el uso de cárceles barcazas, la reforma del SNAI y la implementación del plan Fénix con tecnología e inteligencia."

5. Conclusiones y recomendaciones

Conclusión

- El sistema RAG desarrollado ha demostrado ser una herramienta efectiva para recuperar información y generar respuestas basadas en documentos relevantes. La combinación de ChromaDB y el modelo Flan-T5-Large permitió obtener respuestas estructuradas con un nivel aceptable de precisión. Sin embargo, la evaluación muestra que, aunque el sistema logra recuperar la mayoría de los documentos

clave (Recall del 79.60%), la Precisión del 63.40% indica que aún hay margen de mejora en la selección de información relevante.

- A nivel de generación de texto, los valores obtenidos en BLEU Score (0.6036) y ROUGE-L Score (0.7581) reflejan que el modelo genera respuestas coherentes, aunque con oportunidades de optimización en la exactitud de los detalles. Esto sugiere que, si bien el sistema cumple su objetivo principal, se pueden implementar mejoras para aumentar la precisión y la relevancia de las respuestas.

Recomendación

- Para mejorar la precisión del sistema, se recomienda optimizar la selección de documentos recuperados ajustando los filtros de relevancia en ChromaDB y refinando la conversión de consultas en embeddings. Esto permitiría reducir la cantidad de información irrelevante en las respuestas generadas.
- También sería beneficioso aplicar fine-tuning al modelo Flan-T5-Large utilizando datos más específicos para el dominio del problema. Además, se podría ampliar el corpus de entrenamiento y explorar modelos más avanzados como Flan-T5-XL o GPT-4, lo que ayudaría a mejorar la generación de respuestas con mayor precisión y detalle.