Proyecto integrador (Gpo 10)

Actividad:

Avance 3. Baseline

Objetivos de la actividad:

Proporcionar un marco de referencia para evaluar y mejorar modelos más avanzados.

Integrantes del equipo:

Fernando Benítez Estrada - A01687578
Javier Muñoz Barrios - A01794423
Miguel Ángel Mauriola Medina – A01794830

Fecha de entrega:

09 de febrero del 2025

Contenido

Introducción

La Generación con Recuperación (Retrieval-Augmented Generation, RAG) es un enfoque que combina técnicas de recuperación de información con modelos de generación de lenguaje para que se pueda mejorar la precisión, actualidad y contextualidad de las respuestas generadas. Para que podamos evaluar el rendimiento de un sistema RAG, es fundamental definir métricas adecuadas que aborden tanto la calidad de la recuperación como la efectividad de la generación.

Aquí dividimos las métricas para RAG en dos categorías:

1. Métricas de Recuperación: Evalúan la calidad de los documentos recuperados antes de la generación. Algunas de las métricas más comunes incluyen:
    - Recall@k: Mide la proporción de documentos relevantes dentro de los k elementos recuperados.
    - Precision@k: Evalúa la precisión de los documentos recuperados en el top-k.
    - Mean Reciprocal Rank (MRR): Calcula la posición del primer documento relevante en la lista de recuperación.
    - Normalized Discounted Cumulative Gain (NDCG): Evalúa la relevancia de los documentos recuperados considerando su posición en la lista.
2. Métricas de Generación: Se centran en la calidad del texto generado por el modelo basado en los documentos recuperados. Algunas métricas clave incluyen:
    - BLEU (Bilingual Evaluation Understudy): Mide la similitud entre el texto generado y una referencia, basado en coincidencias de n-grams.
    - ROUGE (Recall-Oriented Understudy for Gisting Evaluation): Evalúa el solapamiento de palabras o frases entre el texto generado y las referencias.
    - METEOR (Metric for Evaluation of Translation with Explicit ORdering): Considera sinónimos y flexiones morfológicas para medir la similitud semántica.
    - BERTScore: Utiliza embeddings de modelos preentrenados para evaluar la similitud semántica entre el texto generado y la referencia.
    - Faithfulness y Groundedness: Métricas cualitativas que miden si el texto generado se basa en la evidencia proporcionada por los documentos recuperados y evita la alucinación (hallucination).

Para que evaluemos un sistema RAG de manera efectiva, es crucial equilibrar estas métricas y optimizarlas en función a nuestro caso de uso específico, garantizando que tanto la recuperación como la generación sean precisas, relevantes y fundamentadas en información verificable.

¿Qué algoritmo se puede utilizar como baseline para predecir las variables objetivo?

En nuestro pipeline se ha implementado un sistema RAG que combina la recuperación de información (usando embeddings generados con OpenAI y almacenados en Pinecone) y la generación de respuestas (mediante la API de Chat de OpenAI). Aunque en nuestro código no se implementa un algoritmo "tradicional" de machine learning para predecir variables de forma supervisada, podemos definir como baseline los siguientes componentes, que actúan de forma sencilla y robusta:

1.**Recuperación de Documentos Baseline:**

Búsqueda por Vecindad en el Espacio de Embeddings (Nearest Neighbor Search):

Utilizamos la generación de embeddings (mediante OpenAI con el modelo text-embedding-3-small) y el almacenamiento en Pinecone para realizar una búsqueda por similitud (cosine similarity).

Este método es directo y reproducible: se transforma cada registro en un vector y, al consultar, se recuperan los documentos más cercanos en términos de similitud. Esto constituye un baseline simple y efectivo para medir la relevancia de la información recuperada.

2.Generación de Respuestas Baseline:

Generación Basada en Plantilla (Rule-Based Generation):

En nuestro código se incluye la función generate_technical_response, que construye una respuesta estructurada a partir de los metadatos recuperados (por ejemplo, mostrando información técnica del dispositivo).

Esta solución rule-based, que mapea directamente los campos del documento recuperado a una respuesta predefinida, sirve como punto de partida para comparar con modelos generativos más complejos (por ejemplo, utilizando el Chat API de OpenAI en modo conversacional).

3.**Pipeline Completo Baseline:**

Combinación del Nearest Neighbor (recuperación) + Plantilla Simple (generación):

Al combinar la búsqueda por similitud en el espacio de embeddings con una generación basada en reglas, se establece un baseline para el sistema RAG completo. Esto permite

predecir o generar las variables objetivo (como el reporte técnico del dispositivo) sin necesidad de ajustes complejos ni entrenamiento adicional.

¿Se puede determinar la importancia de las características para el modelo generado? Recuerden que incluir características irrelevantes puede afectar negativamente el rendimiento del modelo y aumentar la complejidad sin beneficios sustanciales.

Sí, es posible determinar la importancia de las características, y resulta fundamental para evitar incluir información redundante o irrelevante que incremente la complejidad del modelo sin aportar beneficios sustanciales. En nuestro pipeline:

En la Fase de Recuperación:

Los resultados son sobresalientes (Precision, Recall, MRR y NDCG de 1.000), lo que sugiere que las características utilizadas para la indexación (como device_id, user_id, latitude, longitude, battery_level, signal_strength, etc.) están capturando la información esencial de manera muy efectiva.

En la Fase de Generación:

Las métricas muestran resultados más modestos (por ejemplo, BLEU: 0.133, METEOR: 0.287 y Groundedness: 0.267). Esto indica que, aunque la representación de los datos (por ejemplo, a través de la columna ENHANCED_TEXT) es adecuada para la recuperación, podría mejorarse para la generación de respuestas.

¿El modelo está sub/sobreajustado los datos de entrenamiento?

Los resultados actuales muestran una recuperación perfecta, lo cual es positivo, pero también podría sugerir un riesgo de **sobreajuste** en esa fase si el modelo "memoriza" patrones específicos del conjunto de entrenamiento. Sin embargo, en la fase de generación se observan métricas más bajas, lo que indica que esta parte podría no estar capturando la complejidad necesaria o, por el contrario, que aún existe margen para mejorar su capacidad de generalización.

Indicadores:

- **Recuperación:** Con métricas de 1.000, el sistema logra identificar exactamente los documentos relevantes.

- **Generación:** Las métricas moderadas (por ejemplo, BLEU 0.133 y Groundedness 0.267) sugieren que el modelo generativo no está replicando fielmente las referencias o el contexto completo.

¿Cuál es la métrica adecuada para este problema de negocio?

Dado que el sistema RAG combina dos componentes críticos—la recuperación de documentos y la generación de respuestas—se debe evaluar el desempeño utilizando un conjunto híbrido de métricas:

Para la Recuperación:

- **Precision@k, Recall@k, MRR y NDCG:** Estas métricas son esenciales para medir la capacidad del sistema de extraer los documentos relevantes. En nuestro caso, todas estas métricas alcanzan el valor perfecto (1.000), lo que indica una recuperación óptima.

Para la Generación:

- BLEU, ROUGE (rouge1, rouge2, rougeL) y METEOR: Evalúan la similitud y la calidad textual en comparación con las respuestas de referencia.
- Similitud Semántica, Faithfulness y Groundedness: Son críticas para verificar que la respuesta generada sea semánticamente coherente y basada en la evidencia recuperada.

Para el Rendimiento Operativo:

- **Latencia, Uso de Memoria, Tasa de Error y Tasa de hit del caché:** Garantizan que el sistema opere de manera eficiente en un entorno en tiempo real.

¿Cuál debería ser el desempeño mínimo para obtener?

El desempeño mínimo aceptable dependerá de los requerimientos específicos del negocio, pero se pueden establecer algunos umbrales de referencia basados en la literatura y la experiencia práctica:

Para la Recuperación:

- **Precision@k y Recall@k:** Deben ser superiores al 80–90% (idealmente, cerca de 1.000 en escenarios controlados).
- **MRR y NDCG:** Se espera que el primer documento relevante se encuentre entre los primeros resultados (valores cercanos a 1.000).

Para la Generación:

- **BLEU:** Un umbral mínimo ideal podría ser > 0.3 en producción, aunque el valor actual (0.133) indica la necesidad de mejoras.
- **ROUGE:** Valores en torno a 0.5 (como los obtenidos para rouge1 y rougeL) son aceptables, pero se debería apuntar a incrementarlos con ajustes en la generación.
- **METEOR:** Debería ser, al menos, de 0.3 o superior para garantizar una buena calidad textual.
- **Similitud Semántica:** Idealmente superior a 0.70, y se deben mejorar los índices de **Faithfulness** (mejor si supera 0.55) y **Groundedness** (idealmente > 0.30) para asegurar que la respuesta esté bien fundamentada.

Para el Rendimiento Operativo:

- **Latencia:** La latencia promedio debería ser inferior a 2–3 segundos, con un percentil 95 (P95) también en torno a 2–3 segundos para aplicaciones en tiempo real. Actualmente, la latencia promedio es de 3.45 s y P95 de 6.46 s, lo cual sugiere la necesidad de optimizaciones.
- **Uso de Memoria:** Se debería procurar mantenerlo en niveles más bajos (por ejemplo, < 100 MB) si es posible.
- **Tasa de Error y Tasa de hit del Caché:** Idealmente, la tasa de error debe ser 0% y se debe implementar un mecanismo de caché efectivo para mejorar el rendimiento (la tasa actual es 0% de error, pero la tasa de hit del caché es 0%, lo que podría optimizarse).