Escuela Politécnica Nacional

Facultad de Ingeniería de Sistemas

Prof Iván Carrera - Recuperación de Información

Nombres: Oscar Albán, Henry Ramirez

Fecha: 04 de febrero de 2026

Tema: Proyecto Segundo Bimestre "Sistema de recuperación multimodal de información"

Sistema de Recuperación de Información Multimodal con Re-ranking y RAG aplicado a e-commerce

1. Descripción del corpus utilizado

El corpus utilizado en este proyecto proviene del dataset público "Consumer Reviews of Amazon Products" disponible en Kaggle y mantenido por Datafiniti. Este conjunto de datos contiene miles de registros relacionados con productos comercializados en Amazon, incluyendo reseñas, nombres de productos, categorías y enlaces a imágenes.

A diferencia de un catálogo de productos tradicional, este dataset está centrado en reseñas, por lo que múltiples filas pueden corresponder al mismo producto. Por esta razón, el primer desafío técnico consistió en transformar un dataset orientado a opiniones en un corpus estructurado orientado a productos únicos con representación visual.

Para lograrlo, se seleccionaron únicamente las siguientes columnas:

- name: nombre del producto
- primaryCategories: categoría del producto
- imageURLs: enlaces a imágenes del producto

Posteriormente, se aplicaron los siguientes pasos de limpieza y transformación:

1. Eliminación de registros sin imágenes.
2. Agrupación de registros por la firma completa del campo imageURLs para identificar productos distintos.
3. Descarga local de las imágenes.
4. Construcción de un nuevo metadata limpio que contiene:
    - Título del producto
    - Categoría
    - Ruta local de la imagen descargada

Este proceso permitió pasar de un dataset ruidoso y redundante a un corpus multimodal estructurado, donde cada elemento está representado tanto textual como visualmente. El corpus final está compuesto por cientos de productos reales con sus

respectivas imágenes, lo que permite realizar búsquedas tanto por texto como por imagen dentro de un mismo espacio vectorial.

2. Arquitectura general del sistema

El sistema desarrollado sigue una arquitectura moderna de Recuperación de Información Multimodal, compuesta por tres módulos principales:

1. Retrieval vectorial multimodal (CLIP + FAISS)
2. Re-ranking visual fino
3. Generación Aumentada por Recuperación (RAG) con Gemini

Esta arquitectura permite que el sistema no solo recupere productos relevantes, sino que también explique por qué son relevantes, integrando capacidades de búsqueda y generación de lenguaje natural.

3. Retrieval: búsqueda inicial con CLIP y FAISS

Para la etapa de recuperación inicial se utilizó el modelo CLIP ViT-B-32 de sentence-transformers. Este modelo tiene una característica fundamental: puede convertir texto e imágenes en vectores dentro del mismo espacio semántico.

Esto significa que:

- Una imagen de un objeto
- Una descripción textual de ese objeto

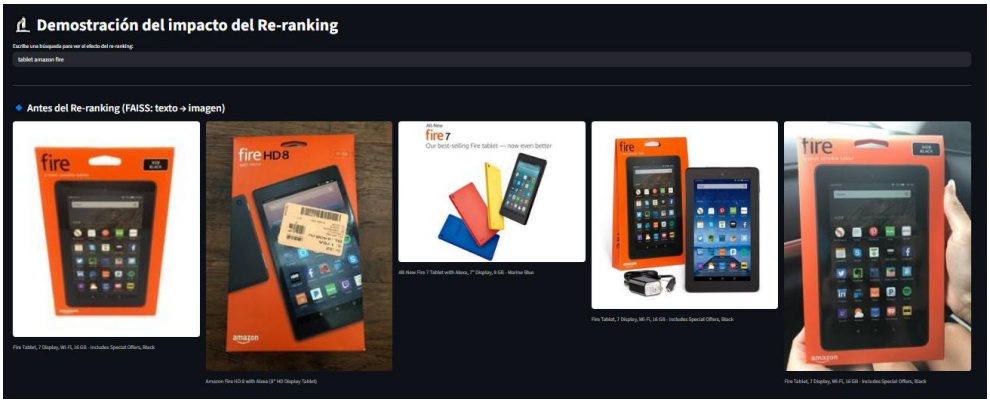
terminan representados como vectores cercanos en el espacio vectorial.

Todas las imágenes del corpus son transformadas en embeddings de 512 dimensiones. Estos vectores se almacenan en un índice FAISS (IndexFlatIP) que utiliza producto interno, equivalente a similitud coseno cuando los vectores están normalizados.

Cuando el usuario realiza una consulta:

- Si es texto → se genera un embedding textual.
- Si es imagen → se genera un embedding visual.
- FAISS recupera los 10 productos más similares de forma extremadamente rápida.

Esta etapa está optimizada para velocidad y cobertura, no para precisión visual perfecta.



4. Re-ranking: mejora de la precisión visual mediante similitud imagen-imagen

La búsqueda con FAISS puede devolver productos semánticamente relacionados, pero no necesariamente los más correctos **desde el punto de vista visual**. Esto ocurre porque la comparación inicial es texto-imagen o imagen-imagen dentro del índice aproximado.

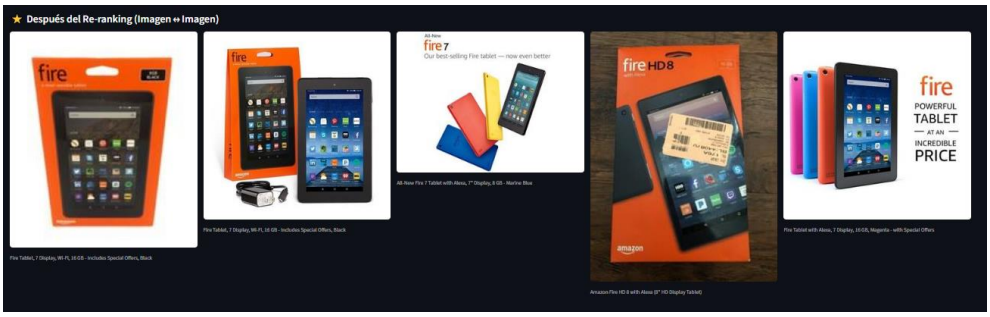
Para mejorar la precisión, se implementa una segunda etapa de re-ranking con una estrategia diferente:

1. Se toman los 10 resultados devueltos por FAISS.
2. Se selecciona la primera imagen recuperada como referencia visual.
3. Se vuelven a cargar todas las imágenes candidatas.
4. Se recalculan embeddings visuales para cada imagen.
5. Se calcula la similitud coseno imagen ↔ imagen entre la imagen de referencia y el resto.
6. Se reordenan los resultados según esta similitud visual exacta.

Este proceso no compara nuevamente contra el texto del usuario, sino que utiliza la primera coincidencia como ancla visual contextual para refinar los resultados.

En la práctica, esto elimina productos que son semánticamente correctos, pero visualmente diferentes, priorizando aquellos que realmente se parecen al objeto recuperado inicialmente.

Este paso mejora notablemente la calidad del primer resultado mostrado al usuario.



5. RAG: Generación Aumentada por Recuperación con Gemini

Una vez identificado el producto más relevante, el sistema utiliza el modelo generativo Gemini para construir una explicación en lenguaje natural.

Es importante destacar que Gemini:

- No busca productos
- No tiene acceso al corpus completo
- Solo recibe el resultado del retrieval

Se le proporciona:

- Lo que el usuario buscó
- El nombre del producto recuperado
- Su categoría

Con esta información, Gemini genera una respuesta que:

- Describe el objeto
- Indica para qué sirve
- Menciona su categoría
- Explica por qué coincide con la búsqueda

Esto convierte el sistema en un asistente explicativo, no solo en un buscador.

6. Ejemplos de consultas reales

Consulta por imagen

El usuario sube la imagen de gato.

El sistema primero describe la imagen subida y luego procede con la búsqueda y explicación.

Consulta por texto

Consulta: “busca más cosas para gatos”.

El sistema encuentra un transportador de mascotas y explica su relación con la búsqueda.

7. Análisis cualitativo del impacto del re-ranking

Durante las pruebas se observó que, sin aplicar re-ranking, los resultados devueltos por FAISS presentaban las siguientes características:

- Aparecían objetos semánticamente relacionados, pero visualmente diferentes.
- Se obtenían productos con formas parecidas, pero pertenecientes a categorías distintas.
- El orden de los resultados no siempre priorizaba el objeto más similar visualmente al recuperado inicialmente.

Al aplicar el re-ranking implementado en el sistema, basado en similitud visual imagen ↔ imagen utilizando la primera imagen recuperada como referencia, se observaron mejoras claras:

- El primer resultado pasa a ser consistentemente el más parecido visualmente.
- Se reduce drásticamente el ruido semántico que introduce FAISS en la recuperación rápida.
- Se prioriza la similitud visual real sobre la similitud aproximada del índice.
- Los productos que se mantienen en las primeras posiciones comparten forma, estructura y características visuales claras.

Esto demuestra que el re-ranking no es un simple reordenamiento, sino un filtro visual contextual que refina significativamente la calidad del resultado final en sistemas multimodales.

8. Análisis cualitativo de la calidad de las respuestas generadas (RAG)

Sin la etapa de generación aumentada por recuperación (RAG), el sistema únicamente mostraría imágenes y títulos de productos, dejando al usuario interpretar por su cuenta la relación con su búsqueda.

Con la incorporación del modelo generativo:

- El sistema describe qué objeto fue recuperado.
- Explica para qué sirve.
- Indica su categoría.
- Confirma explícitamente si coincide con la búsqueda realizada.

Se observó que las respuestas generadas son:

- Coherentes con el producto recuperado.
- Directamente basadas en la información del inventario.
- Claras y comprensibles para el usuario final.
- Consistentes en su estructura y precisión.

Esto demuestra cómo el RAG transforma un sistema de recuperación visual en un sistema explicativo, capaz de traducir resultados técnicos en información interpretativa útil para el usuario.

9. Conclusión

El sistema implementado integra exitosamente recuperación multimodal, re-ranking y generación aumentada por recuperación en un escenario real de e-commerce.

La combinación de CLIP, FAISS y Gemini permite construir un sistema que no solo encuentra productos relevantes, sino que además explica sus resultados, mejorando significativamente la experiencia del usuario.

Este proyecto demuestra una arquitectura moderna y efectiva para sistemas de Recuperación de Información Multimodal.