Informe Técnico: Sistema de Recuperación Multimodal de Información para E-commerce

Estudiante: David Arciniegas

Asignatura: Recuperación de Información

Fecha entrega: 04 de febrero de 2026

1. Descripción del Corpus Utilizado

Para el desarrollo de este sistema de recuperación multimodal, se utilizó el conjunto de datos público "Consumer Reviews of Amazon Products" disponible en Kaggle.

Selección y Preprocesamiento

Originalmente, el dataset contiene miles de registros centrados en reseñas. Sin embargo, para simular un escenario de **e-commerce real** se aplicó un proceso de limpieza antes visto en clase:

1. **Filtrado de Fuentes:** Se utilizaron exclusivamente los archivos `Datafiniti_Amazon_Consumer_Reviews_of_Amazon_Products.csv` y su versión extendida, descartando archivos sin referencias visuales (`1429_1.csv`).
2. **Eliminación de Duplicados:** Dado que el dataset original repite el mismo producto por cada reseña de usuario, se agruparon los registros por el atributo `name` (título del producto). Esto transformó el corpus de una lista de opiniones a un catálogo de productos únicos (SKUs).
3. **Catálogo Resultante:** El corpus final consolidado consta de 79 productos únicos con sus respectivas descripciones textuales y URL de imágenes validadas.
4. **Dominio:** El análisis del corpus revela una predominancia de dispositivos electrónicos de la marca Amazon (Kindle, Fire TV, Echo) y accesorios tecnológicos, con una

ausencia controlada de categorías como "ropa" o "calzado", lo cual fue utilizado posteriormente para pruebas de robustez (negative testing).

2. Explicación del Pipeline Completo

El sistema implementa una arquitectura RAG Multimodal dividida en cuatro etapas secuenciales:

A. Indexación Multimodal

Se utilizó el modelo **CLIP (**`clip-ViT-B-32`**)** para generar los embeddings. La elección de CLIP es fundamental ya que es un modelo entrenado con pares imagen-texto, permitiendo proyectar tanto las descripciones de los productos como las imágenes de consulta en el mismo espacio vectorial de **512 dimensiones**.

- **Almacenamiento:** Los vectores normalizados se indexaron utilizando FAISS con un índice `IndexFlatIP` (Inner Product), que, al trabajar con vectores normalizados, equivale matemáticamente a la similitud del coseno.

B. Recuperación Inicial (Retrieval)

El sistema recibe una consulta (texto o imagen) y la convierte a un vector utilizando el mismo encoder de CLIP. FAISS recupera los Top-K k=20 candidatos más cercanos en el espacio vectorial. Esta etapa prioriza la velocidad y recall sobre la precisión fina.

C. Re-ranking (Reordenamiento)

Para mejorar la precisión de los resultados recuperados, se implementó una estrategia híbrida:

- **Para consultas de Texto:** Se utilizó un modelo Cross-Encoder (`ms-marco-MiniLM-L-6-v2`**)**. A diferencia de los bi-encoders, este modelo procesa la consulta y el documento simultáneamente, capturando relaciones semánticas profundas y asignando un puntaje de relevancia más preciso.
- **Para consultas de Imagen:** Se utilizó la distancia de similitud original de CLIP como score de confianza para reordenar los candidatos, aprovechando la capacidad nativa del modelo para medir la proximidad visual-semántica.
- **Resultado:** Se seleccionan los **Top-5** productos definitivos.

D. Generación Aumentada (RAG) con Contexto

Se construyó un prompt enriquecido que incluye:

1. **Contexto:** Título y descripción de los 5 productos ganadores del re-ranking.
2. **Historial:** Memoria de los últimos turnos de la conversación para permitir refinamientos.

3. **Input Multimodal:** Si el usuario sube una imagen, esta se envía directamente al modelo generativo **Gemini 2.5 Flash** junto con el texto.

El modelo generativo actúa bajo instrucciones estrictas de **Grounding**: solo debe recomendar productos presentes en el contexto recuperado, justificando su elección en base a evidencia visual o textual.

3. Ejemplos de Consultas y Resultados

Se realizaron pruebas funcionales para validar la capacidad multimodal y la robustez del sistema.

Caso 1: Búsqueda Visual Exitosa (Positive Testing)

- **Consulta:** Imagen de un dispositivo "Fire TV Stick"
- **Resultados del Retrieval:** El sistema recuperó en primera posición el *“Fire TV Stick Streaming Media Player Pair Kit”* con un score de similitud de **0.32**.
- **Respuesta Generada (RAG):**
  "La imagen muestra claramente un dispositivo Amazon Fire TV Stick junto con su control remoto... El nombre del producto es una descripción directa y precisa del dispositivo visible."
- **Visualización (UMAP):** La proyección en 2D mostró que el vector de la consulta (estrella roja) se ubicó en la vecindad inmediata de los clusters de dispositivos de streaming (puntos azules).

Imagen 1: imagen cargada Fire TV Stick

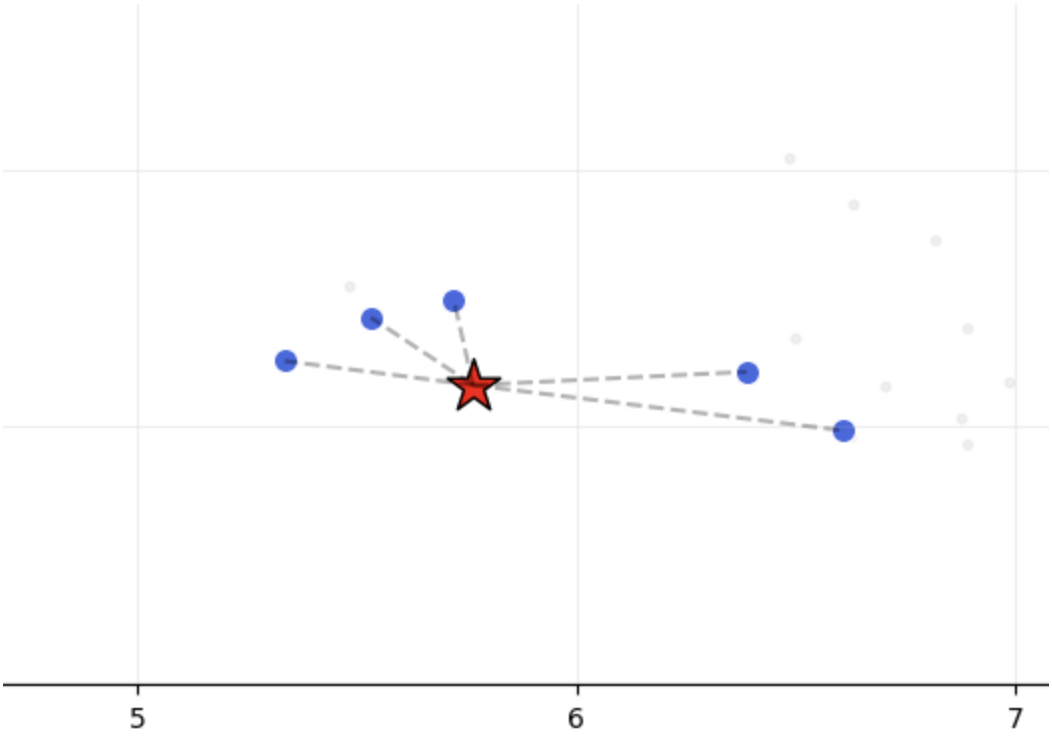

Imagen 2: resultados de gráfica UMAP

Caso 2: Prueba de Robustez / Negativa (Negative Testing)

- **Consulta:** Imagen de unos "Zapatos Deportivos Rojos" .
- **Resultados del Retrieval:** El sistema recuperó objetos con atributos visuales similares (color rojo), como una funda de sartén roja (*Red Handle Cover*) o dispositivos con luces, pero con scores de similitud bajos (< 0.20).
- **Respuesta Generada (RAG):**
  "Revisando el contexto... no hay ningún producto que sea similar a la imagen (zapatillas deportivas). Todos los productos listados son dispositivos electrónicos... Por lo tanto, no puedo recomendar un producto."

4. Análisis Cualitativo

Impacto del Re-ranking

La implementación del re-ranking demostró ser exitosa sobretodo en consultas textuales ambiguas. Mientras que la búsqueda vectorial inicial (FAISS) traía productos vagamente relacionados por palabras clave compartidas, el Cross-Encoder logró filtrar el contenido elevando a las primeras posiciones aquellos productos que respondían a la intención del usuario. En la modalidad de imagen, el uso de los scores de confianza de CLIP permitió

diferenciar entre una coincidencia visual fuerte (el Fire TV) y una coincidencia circunstancial (los zapatos rojos vs. funda roja).

Calidad de las Respuestas y Grounding

El uso de Gemini 2.5 Flash en configuración RAG demostró una alta capacidad de Grounding**.** El sistema exhibió un comportamiento "honesto":

1. **Evitó Alucinaciones:** En la prueba de los zapatos rojos, a pesar de recibir una imagen clara, el modelo priorizó el contexto recuperado (que solo contenía electrónica) y se negó a inventar una recomendación falsa.
2. **Justificación de Evidencia:** En los casos exitosos, las respuestas no solo indicaron el producto, sino que explicaron *por qué* se seleccionó (ej. *"identificable por el botón del Asistente..."*).
3. **Manejo de Contexto:** La memoria de sesión permitió realizar preguntas de refinamiento (ej. *"¿Tiene control por voz?"*) sin necesidad de volver a subir la imagen o repetir el nombre del producto, simulando una experiencia de asistente de ventas real.