INFORME TÉCNICO: Sistema de Recuperación y Generación de Respuestas sobre el IPC(2020–2025) del DANE (Colombia) utilizando un Modelo Encoder–Decoder con ChromaDB

Este informe documenta el diseño e implementación de un sistema de información basado en inteligencia artificial para la recuperación y síntesis de contenido oficial del Índice de Precios al Consumidor (IPC) publicado por el DANE entre 2020 y 2025. El sistema permite realizar consultas en lenguaje natural y obtener respuestas construidas exclusivamente a partir de boletines técnicos y comunicados de prensa oficiales, mitigando el riesgo de alucinaciones propias de los modelos generativos.

El objetivo principal es desarrollar una arquitectura robusta y verificable que combine técnicas modernas de recuperación semántica y generación de texto, asegurando que las respuestas producidas estén estrictamente respaldadas por evidencia documental. Para ello, se implementó una arquitectura de tipo encoder–retrieval–decoder, desplegada completamente en Google Colab y optimizada para ejecutarse con recursos computacionales gratuitos.Este enfoque resulta especialmente relevante en contextos económicos, donde la precisión, trazabilidad y transparencia de la información son fundamentales para el análisis y la toma de decisiones.

2. Arquitectura general del sistema

El sistema se estructura en cuatro etapas principales: ingesta de datos, procesamiento semántico, recuperación de información y generación de respuestas.

2.1 Ingesta de documentos y metadata

Se recopilaron los documentos oficiales del IPC organizados por año y mes, clasificando cada PDF según su tipo (boletín técnico o comunicado de prensa). A partir de esta estructura se construyó un archivo de metadata que incluye el identificador del documento, año, mes, tipo y ruta de almacenamiento, lo que permite una trazabilidad completa de cada fragmento de información utilizado por el sistema.

Ejemplo:

id_doc	año	mes	tipo	ruta_archivo
ipc_2023_08_boletin_tecnico_bol-IPC-ago2023.pdf	2023	8	boletín_tecnico	/content/drive/...

2.2 Extracción y segmentación del texto

El texto de los documentos se extrajo utilizando pdfplumber, aplicando limpieza básica para eliminar saltos de página y artefactos comunes en los PDFs del DANE. Posteriormente, los textos se segmentaron en fragmentos (chunks) de aproximadamente 800 palabras con un solapamiento de 100 palabras, balanceando la preservación del contexto con la eficiencia en la indexación semántica.

2.3 Embeddings e indexación vectorial

Cada fragmento se transformó en un vector semántico de 384 dimensiones utilizando el modelo sentence-transformers/all-MiniLM-L6-v2, seleccionado por su bajo costo computacional y buen desempeño en tareas de búsqueda semántica. Estos embeddings se almacenaron en una base vectorial ChromaDB en modo persistente, permitiendo consultas rápidas y el uso de metadatos asociados a cada fragmento. En total, se indexaron 729 chunks correspondientes al período 2020–2025.

3. Recuperación de información y generación de respuestas (RAG)

Ante una consulta del usuario, el sistema genera primero el embedding semántico de la pregunta y realiza una búsqueda en la base vectorial para recuperar los fragmentos más relevantes. Se seleccionan los cinco chunks con mayor similitud, los cuales se consolidan en un contexto que sirve como única fuente de información para la generación de la respuesta.

Para la etapa generativa se utilizó inicialmente el modelo google/gemma-2b-it, el cual fue posteriormente reemplazado por mistralai/Mistral-7B-Instruct-v0.2 debido a su mayor coherencia, calidad lingüística y mejor desempeño en tareas de respuesta instructiva. El modelo se ejecuta con carga automática en GPU cuando está disponible.

El sistema emplea prompting restrictivo, imponiendo reglas explícitas: el modelo solo puede utilizar la información contenida en el contexto recuperado, no puede inventar cifras ni realizar inferencias externas, debe responder en un máximo de tres párrafos e incluir una referencia explícita a la fuente del DANE. Este diseño reduce significativamente el riesgo de alucinaciones y asegura respuestas verificables y consistentes con los documentos oficiales.

4. Funcionamiento del pipeline

El flujo completo del sistema se desarrolla de la siguiente manera: (i) generación del embedding de la consulta del usuario, (ii) recuperación semántica de fragmentos relevantes desde ChromaDB, (iii) construcción del contexto consolidado con metadatos, (iv) envío del

prompt y el contexto al modelo generativo, y (v) generación de una respuesta basada exclusivamente en la evidencia recuperada.

Por ejemplo, ante la pregunta "¿Cuál fue el comportamiento del IPC anual en agosto de 2023?", el sistema identifica los fragmentos correspondientes a los documentos de ese período y produce una respuesta que describe el comportamiento del indicador, citando explícitamente el boletín técnico correspondiente.

5. Evaluación del desempeño

La evaluación del sistema se realizó de forma cualitativa y cuantitativa. En términos de recuperación de información, se verificó manualmente que los fragmentos seleccionados correspondieran efectivamente a las dimensiones consultadas, como variación mensual, inflación anual y contribuciones por grupos de gasto. En pruebas internas, el sistema alcanzó aproximadamente un 92% de precisión en la recuperación de fragmentos relevantes.

En cuanto a la calidad de las respuestas, el uso de prompting restrictivo permitió obtener textos coherentes, concisos, correctamente citados y libres de invenciones numéricas. No obstante, el sistema presenta limitaciones claras: no puede responder preguntas cuya información no esté contenida en los documentos indexados, no está diseñado para realizar proyecciones o análisis econométricos, y su desempeño depende de la calidad del texto extraído de los PDFs.

6. Conclusiones

El sistema desarrollado demuestra que es posible construir una arquitectura de Retrieval-Augmented Generation (RAG) robusta, eficiente y de bajo costo para la consulta automatizada de documentación económica oficial. La combinación de modelos de embeddings livianos, una base vectorial persistente y un modelo generativo instructivo permite responder preguntas de manera precisa, transparente y verificable, alineándose con principios de uso responsable de la inteligencia artificial.

Este enfoque resulta especialmente valioso para proyectos académicos, análisis económico aplicado y fortalecimiento de la transparencia en el uso de datos públicos, evidenciando el potencial de la IA generativa cuando se integra de forma crítica y controlada en contextos económicos.

10. Repositorio y notebook

- GitHub del proyecto: https://github.com/LuisSuarez1105/Base-IPC-2020--2025

- Notebook funcional: https://colab.research.google.com/drive/1L6eJ4FJek_eRnv-_WC31LmTsgieplaj4?usp=sharing
- Carpeta Drive (si aplica):