Democratización de Inteligencia Artificial para la Gestión de Documentos PDF en Contextos con Recursos Limitados

Joaquin Enrique Rivas Sánchez[1*], Naylin Brizuela Capote [2], Angel Alberto Vázquez Sánchez [3]
Universidad de las Ciencias Informáticas. Carretera San Antonio Km 2 1 Reparto Torrens, La Lisa, La Habana.
joaquiners@estudiantes.uci.cu[1*]. naylinbc@estudiantes.uci.cu[2]. aavazquez@uci.cu[3]
[*] Autor para correspondencia: joaquiners@estudiantes.uci.cu

Resumen
Este artículo abordó la necesidad de desarrollar una herramienta accesible para la gestión y análisis de documentos PDF en entornos con recursos limitados, como el contexto cubano, donde las restricciones de conectividad y métodos de pago internacionales dificultan el acceso a soluciones avanzadas como ChatPDF y Humata.ai. Además, se consideró crucial avanzar en el desarrollo de tecnologías que promuevan la soberanía tecnológica, reduciendo la dependencia de plataformas extranjeras. Para solucionar este problema, se propuso un prototipo basado en la técnica Retrieval-Augmented Generation (RAG), empleando el modelo de lenguaje llama-3.2-3B y all-mpnet-base-v2 para la generación de word embeddings. La principal funcionalidad del prototipo incluye la capacidad de responder preguntas abiertas (open-domain question answering). Adicionalmente, el prototipo demostró su potencial para funcionar eficazmente en entornos con bajos recursos. Se concluye que este prototipo puede ser desarrollado aún más y tiene el potencial de democratizar el acceso a tecnologías avanzadas de inteligencia artificial, lo que fortalecería la capacidad local para gestionar documentos PDF de manera eficiente.

Palabras clave: inteligencia artificial, RAG, procesamiento de lenguaje natural, conectividad limitada, código abierto.

Temática: Inteligencia artificial para la transformación digital.

Introducción

El conocimiento colectivo digitalizado, presente en documentos como noticias, blogs, literatura científica y empresariales, se ha expandido con la revolución digital y el auge de las redes sociales. Los PDF representan una fuente clave de información, pero en países como Cuba, los profesionales enfrentan barreras significativas, como conectividad limitada y falta de acceso a métodos de pago internacionales, que dificultan el uso de herramientas de IA como ChatPDF, Humata.ai o ChatGPT. Estas plataformas, además de requerir conexión constante, suelen operar bajo modelos de pago inaccesibles para muchos cubanos.

Esta investigación busca adaptar soluciones existentes para ofrecer herramientas accesibles que aprovechen la IA en la gestión de documentos PDF. Usando la técnica Retrieval-Augmented Generation (RAG), se abordan problemas como la alucinación y el conocimiento desactualizado en modelos de lenguaje, mejorando la precisión y fiabilidad en tareas como la respuesta a preguntas, el resumen de documentos y la verificación de hechos, integrando información actualizada y relevante.

Los objetivos de este trabajo son:

1. Identificar las limitaciones actuales en el uso de herramientas de IA para la gestión de documentos PDF en el contexto cubano.
2. Adaptar y replicar una solución basada en la técnica RAG para ofrecer una alternativa que considere las condiciones de conectividad así como las limitaciones de infraestructura de nuestro país
3. Evaluar el desempeño de la herramienta propuesta en un entorno con limitaciones tecnológicas, con énfasis en su viabilidad y utilidad práctica.
4. Proponer futuras líneas de desarrollo para extender la funcionalidad de la herramienta, considerando la integración con otras tecnologías emergentes. Esto incluirá recomendaciones para mejorar la capacitación en el uso de herramientas digitales y fomentar un ecosistema que apoye la innovación en la gestión documental.

Desarrollo, Materiales y métodos o Metodología

El prototipo fue desarrollado utilizando Python v3.12 debido a su versatilidad y compatibilidad con bibliotecas clave en inteligencia artificial y procesamiento de texto. Se emplearon herramientas como Giskard v2.15.2 para evaluar el sistema, Faiss-cpu v1.9.0 para realizar búsquedas vectoriales rápidas y precisas, Gradio v5.3.0 para construir una interfaz gráfica interactiva, llama_cpp_python v0.2.90 para cargar y ejecutar localmente el modelo Llama 3.2-3B, y PyTorch v2.4.1+cu124 para soportar operaciones de aprendizaje profundo, optimizando el rendimiento en diversos entornos. La

solución se basó en Retrieval-Augmented Generation (RAG), una metodología que combina la búsqueda de información con la generación de respuestas, utilizando el modelo Llama 3.2-3B-Instruct para comprensión y generación de texto, y all-mpnet-base-v2 para generar embeddings que optimizan el análisis semántico de los documentos. Las pruebas iniciales se llevaron a cabo en Google Colab debido a su facilidad de configuración y acceso a recursos de GPU, mientras que el desarrollo posterior se migró a un entorno local con un equipo Intel Core i7 7700k, 16 GB de RAM, tarjeta gráfica NVIDIA RTX 2070 y sistema operativo Arch Linux, lo que permitió ejecutar tareas de procesamiento y generación sin limitaciones de hardware.

Implementación y desarrollo de RAG

El enfoque de RAG en este prototipo sigue la estructura tradicional, dividida en tres fases principales: indexado de documentos, recuperación de información y generación aumentada. Esta metodología combina las capacidades generativas de los modelos de lenguaje con la recuperación de información desde memoria no paramétrica, lo que permite mejorar la precisión y relevancia de las respuestas, reduciendo significativamente las alucinaciones (Zhang y Kotanko 2024).

A continuación, se describe la estructura de RAG implementada en el prototipo, destacando los pasos clave y las optimizaciones realizadas para su adaptación a un entorno con limitaciones de hardware e infraestructura como el cubano.

Fases de RAG en el Prototipo

1. Indexado de Documentos: En esta fase, los documentos PDF se dividen en pequeños "chunks" que se convierten en vectores de embeddings usando el modelo preentrenado *all-mpnet-base-v2*. Estos vectores se almacenan en una base de datos optimizada con FAISS, junto con metadatos relevantes, para facilitar la identificación rápida de información relevante incluso en grandes corpus (Dai, Olah, y Le 2015).
2. Recuperación de Información: Se buscan los *k* chunks más relevantes mediante búsqueda semántica y un enfoque híbrido que combina BM25 con técnicas de reescritura y expansión de consultas. Esto asegura que las preguntas del usuario se interpreten adecuadamente, mejorando la precisión y relevancia de los resultados obtenidos (Omrani et al. 2024; Chan et al. 2024).
3. Generación Aumentada: El modelo de lenguaje *Llama 3.2 3B* genera respuestas coherentes y precisas utilizando un *prompt* que integra la consulta inicial y los chunks relevantes recuperados. Se aplican técnicas de compresión y reordenamiento para garantizar que solo la información más relevante se incluya en el *prompt*, optimizando así la generación de respuestas contextualizadas (Shi et al. 2024).

Metodologías de prueba

El prototipo fue validado por su capacidad para procesar documentos PDF y generar respuestas coherentes y rápidas, utilizando tres artículos científicos de diferentes áreas: "Attention is All You Need" (Vaswani et al. 2023), Estudio cubano sobre vacunación contra COVID-19 (Toledo-Romaní et al. 2023), Reporte del IPCC 2023 (Calvin et al. 2023).

Se diseñaron 60 preguntas distribuidas en tres conjuntos de prueba para evaluar aspectos como resistencia a preguntas distractoras, desambiguación, ignorancia de contexto irrelevante y manejo de contexto previo. La evaluación utilizó la biblioteca Giskard y el toolkit RAGET para analizar componentes clave del sistema:

- Retriever: Encargado de recuperar información relevante del conjunto de documentos.
- Generator: Utiliza un LLM para generar respuestas basadas en los contextos recuperados.
- Rewriter: Reformula consultas para mejorar su relevancia o adaptarlas al contexto previo.
- Router: Filtra consultas según las intenciones del usuario, optimizando la interacción.
- Knowledge Base: Base de conocimiento que almacena los documentos utilizados para generar respuestas.

Resultados y discusión

Los resultados, presentados en la Tabla 1 y Fig.1, muestran los resultados de las pruebas sobre el prototipo, ChatPDF y Humata.ai en los tres conjuntos de datos probados.

Dataset	Sistema	Generator	Retriever	Rewriter	Routing	Knowledge Base	Overall Correctness
ipcc	Prototipo	40%	40%	40%	100%	0%	40%
	ChatPDF	85%	80%	80%	100%	0%	85%
	Humata	45%	20%	0%	100%	0%	45%
soberana	Prototipo	45%	62.5%	16.67%	100%	0%	45%
	ChatPDF	35%	50%	16.67%	100%	33.33%	40%
	Humata	45%	62.5%	25%	100%	0%	40%

attention	Prototipo	50%	62.5%	33.33%	100%	0%	45%
	ChatPDF	70%	37.5%	58.33%	100%	25.0%	65%
	Humata	40%	37.5%	16.67%	100%	50%	35%

Tabla 1.Resultados porcentuales.

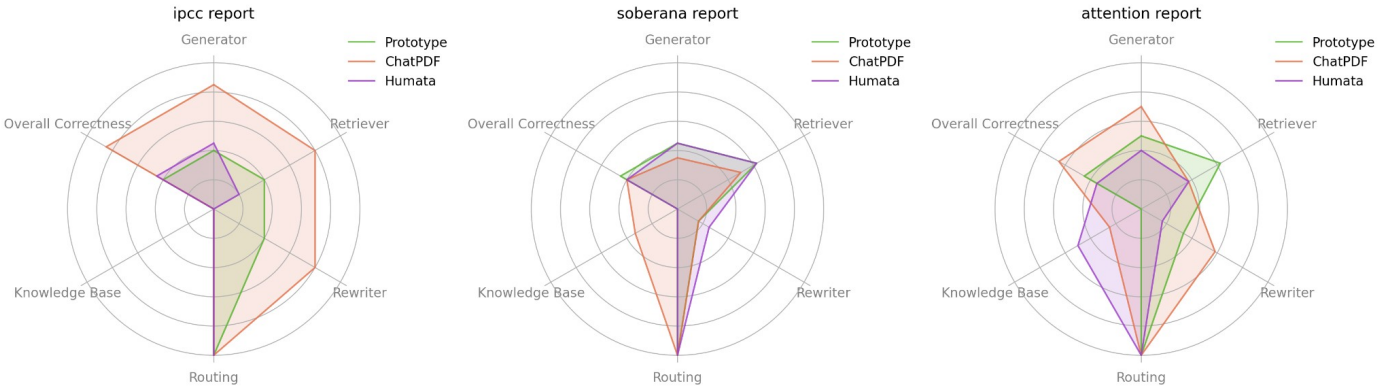

Figura 1. Gráfico comparativo.

A partir de los resultados mostrados se pueden extraer las siguientes conclusiones:
El prototipo muestra un desempeño sólido en la generación de respuestas y la recuperación de información, destacándose especialmente en documentos textuales como "soberana" e "ipcc." Sin embargo, su rendimiento es inferior al de herramientas como ChatPDF y Humata en el manejo de formatos complejos como gráficos, tablas y fórmulas, presentes en datasets como "attention." Esto resalta la necesidad de optimizar el preprocesamiento de formatos más complejos para ampliar su aplicabilidad.
La consulta y reescritura de preguntas requieren mejoras, especialmente en contextos donde herramientas comerciales muestran ventajas claras. A pesar de estas limitaciones, el prototipo tiene el potencial de convertirse en una alternativa localizada eficaz para la gestión de documentos PDF en entornos con recursos limitados.
Futuras líneas de desarrollo incluyen la implementación de modelos más pequeños, como Llama 3.2-1B, optimizados para dispositivos móviles. Esto permitiría una mayor concurrencia de usuarios, mejoraría la accesibilidad y reduciría la carga en los servidores. Además, la incorporación de modelos multimodales ampliaría la capacidad para procesar gráficos, diagramas y formatos más allá del PDF.
El prototipo podría ser integrado en plataformas nacionales de aprendizaje, como entornos virtuales educativos, para apoyar a estudiantes en el estudio de materiales. Asimismo, podría beneficiar a investigadores cubanos al facilitar el análisis de artículos científicos, la generación de resúmenes, la traducción especializada y la extracción de información relevante de documentos PDF.

Conclusiones

El presente trabajo confirmó que es posible desarrollar una herramienta eficiente para la gestión y análisis de documentos PDF en entornos con recursos limitados, como el cubano, utilizando la técnica Retrieval-Augmented Generation (RAG). El prototipo evaluado mostró un rendimiento satisfactorio en términos de precisión y rapidez, validando su viabilidad en escenarios de baja infraestructura tecnológica. Esta herramienta ofrece una alternativa local a soluciones comerciales inaccesibles, promoviendo la soberanía tecnológica al reducir la dependencia de plataformas extranjeras. El prototipo tiene un gran potencial para futuras aplicaciones, como en plataformas educativas nacionales, donde puede facilitar el acceso a recursos de estudio en contextos con limitación de profesores, y en otros sectores como el asesoramiento médico o la gestión de quejas empresariales. Entre las líneas futuras de desarrollo se propone la evaluación con conjuntos de datos más amplios y la incorporación de modelos más pequeños o multimodales que permitan ampliar sus capacidades, incluyendo la interpretación de imágenes y gráficos. Estos avances aumentan la versatilidad del sistema y extenderían su impacto en diversos sectores.

Referencias bibliográficas

Calvin, Katherine, Dipak Dasgupta, Gerhard Krinner, Aditi Mukherji, Peter W. Thorne, Christopher Trisos, José Romero, et al. 2023. «IPCC, 2023: Climate Change 2023: Synthesis Report. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland.» First. Intergovernmental Panel on Climate Change (IPCC). https://doi.org/10.59327/IPCC/AR6-9789291691647.

Chan, Chi-Min, Chunpu Xu, Ruibin Yuan, Hongyin Luo, Wei Xue, Yike Guo, y Jie Fu. 2024. «RQ-RAG: Learning to Refine Queries for Retrieval Augmented Generation». arXiv. https://doi.org/10.48550/ARXIV.2404.00610.

Dai, Andrew M., Christopher Olah, y Quoc V. Le. 2015. «Document Embedding with Paragraph Vectors». arXiv. http://arxiv.org/abs/1507.07998.

Omrani, Pouria, Alireza Hosseini, Kiana Hooshanfar, Zahra Ebrahimian, Ramin Toosi, y Mohammad Ali Akhaee. 2024. «Hybrid Retrieval-Augmented Generation Approach for LLMs Query Response Enhancement». En *2024 10th International Conference on Web Research (ICWR)*, 22-26. Tehran, Iran, Islamic Republic of: IEEE. https://doi.org/10.1109/ICWR61162.2024.10533345.

Shi, Kaize, Xueyao Sun, Qing Li, y Guandong Xu. 2024. «Compressing Long Context for Enhancing RAG with AMR-based Concept Distillation». arXiv. http://arxiv.org/abs/2405.03085.

Toledo-Romaní, María Eugenia, Mayra García-Carmenate, Carmen Valenzuela-Silva, Waldemar Baldoquín-Rodríguez, Marisel Martínez-Pérez, Meiby Rodríguez-González, Beatriz Paredes-Moreno, et al. 2023. «Safety and Efficacy of the Two Doses Conjugated Protein-Based SOBERANA-02 COVID-19 Vaccine and of a Heterologous Three-Dose Combination with SOBERANA-Plus: A Double-Blind, Randomised, Placebo-Controlled Phase 3 Clinical Trial». *The Lancet Regional Health - Americas* 18 (febrero):100423. https://doi.org/10.1016/j.lana.2022.100423.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, y Illia Polosukhin. 2023. «Attention Is All You Need». arXiv. http://arxiv.org/abs/1706.03762.

Zhang, Hanjie, y Peter Kotanko. 2024. «#1506 Uremic Toxicity: Gaining Novel Insights through AI-Driven Literature Review». *Nephrology Dialysis Transplantation* 39 (Supplement_1): gfae069-0657-1506. https://doi.org/10.1093/ndt/gfae069.657.