Integración de Agentes RAG para la Generación Automática de Preguntas desde Documentos PDF en Moodle

Mustapha Bouleili
CE Inteligencia Artificial y Big data
ITICBCN, CAT, ES
2024_mustapha.bouleili@iticbcn.cat

Lautaro Fleitas
CE Inteligencia Artificial y Big data
ITICBCN, CAT, ES
2024_lautaro.fleitas@iticbcn.cat

Abstract

Creación de un Agente RAG para generación automatizada de cuestionarios educativos mediante procesamiento multimodal de PDF, utilizando modelos cuantizados (Mistral-7B) e integración con Moodle vía API REST.

1. Estado del arte y Motivación

Actualmente, existen soluciones fragmentadas para tareas como extracción de texto desde PDF, generación de preguntas y gestión de plataformas de e-learning. Sin embargo, no existe una solución unificada que combine estos elementos bajo un mismo sistema modular.

Con un Agente que es capaz de usar herramientas para a partir de un PDF, procesarlo y generar preguntas tipo test en diferentes formatos dependiendo de la demanda del usuario.

La motivación principal ha sido comprender a fondo cómo funciona un agente RAG, y demostrar su aplicación práctica como herramienta educativa, autónoma y escalable.

2. Desarrollo: Agente RAG Modular con *LangGraph*

- 2.1 Implementación del Agente RAG

El agente RAG se implementó de forma modular con *LangGraph* y su flujo de trabajo incluye la conversión de archivos PDF a texto utilizando *pdfplumber,* la segmentación del contenido mediante técnicas de *chunking*, la vectorización e indexación del contenido en Pinecone usando como modelo de *embedding* a *all-MiniLM-L6-v2* para la recuperación de contexto relevante a través de un modelo de lenguaje la generación automática de preguntas basadas en ese contexto, el modelo que hemos usado como LLM es *Nous-Hermes-2-Mistral-7B-DPO*, para la creación de cuestionarios que se suben automáticamente a Moodle mediante su API este flujo ha sido creado mediante tools las cuales se van pasando el estado de las herramientas anteriores para que se vea el correcto funcionamiento

- 2.2 Herramientas y Librerías Utilizadas

LangGraph se usó para estructurar el flujo del agente *pdfplumber* se utilizó para extraer texto desde archivos PDF Pinecone funcionó como motor de vectorización y búsqueda semántica Gradio proporcionó una interfaz gráfica para la interacción con el usuario *FastAPI* permitió implementar un servidor local para gestionar los *endpoints* la API de Moodle facilitó la integración con la plataforma educativa y se utilizaron modelos LLM cuantizados para garantizar eficiencia en entornos con recursos limitados.

- 2.3 Infraestructura Técnica

Está optimizado de forma para que se pueda ejecutar en entornos de bajo recurso, en este caso *Colab*. Por eso usamos el modelo () cuantizado a 4 bits usando *BitsAndBytes*. Para la recuperación basada en el contexto hemos usado *SentenceTransformer* (*all-MiniLM-L6-v2*), y los vectores se almacenan y se consultan en Pinecone.

El flujo completo, desde la conversión de PDF a texto, segmentación, indexación, consulta y generación de preguntas tipo test, se gestiona mediante un grafo de estados (*StateGraph*), asegurando modularidad y automatización.

La integración con Moodle se realiza mediante su API REST, lo que permite subir automáticamente las preguntas generadas al espacio privado del docente.

Se ha desplegado una instancia local de Moodle utilizando docker-compose con la imagen *Bitnami/Moodle*, exponiendo el servicio a través de un túnel seguro mediante *Ngrok* para habilitar el acceso remoto a la API.

Por último, se construyó una interfaz con Gradio, lo que permite interactuar fácilmente con el sistema sin requerir conocimientos técnicos.

3. Resultados

Como resultados obtenidos ha generado buenas preguntas basadas en el contexto, pero si se generan demasiadas preguntas puede generar que el modelo pierda contexto o alucine. Y el archivo se sube bien en el área privada del Moodle.

Tema	Número de archivos	Número de páginas	Tiempo de ejecución
IOT	5	176	93 s
Web Scraping	4	161	84 s
Programación R	10	330	129 s
MongoDB	1	28	41 s

Basándose en estos datos obtenidos mediante varias pruebas a nuestro Agente, hemos determinado que lleva un tiempo aproximado de 1,69 segundos por página procesada segundos

4. Trabajo Futuro

El sistema logró generar preguntas tipo test e integrarse exitosamente con Moodle. En futuras versiones se plantea su integración con *Google Cloud* para alojarlo en la nube, escalarlo y aprovechar servicios como *Google Forms* y *Gmail*. Esto permitiría generar y enviar formularios automáticamente, ampliando su uso más allá de Moodle. También se prevé el uso de modelos más avanzados (como GPT-4, Claude o Gemini) para mejorar la calidad de las preguntas. Estas funciones no se implementaron por limitaciones técnicas y de acceso, pero siguen en consideración para próximas fases.

5. Conclusiones

El sistema cumple con su objetivo: genera preguntas tipo test a partir de documentos PDF y las sube correctamente al área privada de Moodle. A pesar de ello, existen limitaciones. El tiempo de ejecución varía según el tamaño de los archivos y el estado del índice en Pinecone. En ocasiones, especialmente en la primera ejecución, las preguntas pueden ser poco coherentes, aunque esto mejora posteriormente. Además, si el índice contiene temas variados, pueden generarse preguntas fuera de contexto. La subida a Moodle funciona de forma estable y sin fallos.

Bibliografía

Universitat Oberta de Catalunya. (2025). *Estilo APA.*
https://biblioteca.uoc.edu/es/pagina/Estilo-APA/

Sanchez, C. [Carlos]. (2023). Cómo citar ChatGPT. *Normas APA (7ma edición)*.
https://normas-apa.org/referencias/como-citar-chatgpt/

OpenAI. (2023). *ChatGPT* (versión 14 de marzo) [Large language model].
https://chat.openai.com/chat

DeepSeek. (2023). *DeepSeek* (versión 2025-03-24) [Large language model].
https://www.deepseek.com/

Hugging Face. (2025). *Hugging Face Agents Course.*
https://huggingface.co/agents-course

Singh, A., Ehtesham, A., Kumar, S., Talaei, T., (2025) *Agentic Retrieval-Augmented Generation: A Survey on Agentic RAG. https://arxiv.org/abs/2501.09136#*

Hugging Face. (2025). *Gradio* (Versión 5.31.0) [Software].

Google LLC. (2017). *Google Colaboratory* [Software].

Harrison, C. [Chase], y LangChain Team. (2022). *LangChain* (Versión 0.3.24) [Software].

LangChain Team. (2025). *LangGraph* (Versión 0.3.1) [Software].

Nous Research. (2024). *Nous Hermes 2 - Mistral 7B - DPO* [Software].

Pinecone Systems Inc. (2021). *Pinecone* [Software].

Reimers, N. [Nils], y Gurevych, I. [Iryna]. (2020). *Sentence-Transformers: Embeddings semánticos multilingües* [Software].

Investigar es fácil. (24 de abril de 2022). *Cómo referenciar un video de YouTube con Normas APA 7ma edición- Investigar es fácil.* [Archivo de video]. Youtube.
https://www.youtube.com/watch?v=0TZW1lnEZZQ

José Maria Labarta. (29 de enero de 2024). *Cómo instalar Moodle usando Docker compose.* [Archivo de video]. Youtube.
https://www.youtube.com/watch?v=DpfCPihqKHs

VMware. (2025). *bitnami/moodle* [Software].
Ngrok. (2025). Ngrok [Software].