Proyecto: Enterprise AI Agent Platform - Día 3 Resumen del Trabajo Realizado Durante el Día 3, se avanzó en la construcción del módulo de ingestión de documentos para el sistema RAG. Este módulo permite recibir archivos PDF, TXT y DOCX, extraer su contenido, limpiarlo y dividirlo en chunks que serán posteriormente utilizados para generar embeddings y alimentar el vector store. Componentes Implementados 1. DocumentIngestor Controla el flujo principal de ingestión. Detecta el tipo de archivo según el mimetype. Deriva el archivo al extractor correspondiente. Aplica limpieza de texto. Genera chunks del texto procesado. 2. Extractores **PDF:** Implementado con pdfjs-dist. Lee cada página y extrae el texto.
TXT: Lee directamente el contenido del archivo UTF-8.
DOCX: Implementado con la librería mammoth para convertir el .docx en texto plano.
3. Utilidades **cleanText:** Limpieza básica del contenido, remoción de saltos raros y normalización.
chunkText: Divide el texto en porciones más pequeñas (chunks), útiles para embeddings.
4. Endpoint de ingestión Se añadió un nuevo controlador y ruta para procesar documentos vía API: POST /api/ingest
Resultados de Prueba PDF: Se extrajo texto correctamente, con advertencias mínimas (fuentes no definidas). TXT: Procesamiento correcto, lectura directa del archivo. DOCX: Funcionamiento completo utilizando mammoth. Conclusión El módulo de ingestión quedó funcional y preparado para integrarse con la capa de embeddings y el posterior almacenamiento vectorial. El proyecto avanza hacia un pipeline RAG completo.