Agente de Extracción de Información

Introducción:

El agente de extracción es el primer componente dentro de la arquitectura multiagente del sistema Curador Multiagente de Roadmaps Tech.

Su función principal consiste en leer y procesar documentos PDF o textos planos ubicados en la carpeta data/.

Función técnica:

Este agente identifica el contenido relevante, elimina ruido (como saltos de línea innecesarios o símbolos) y prepara el texto para su posterior segmentación.

En proyectos de Inteligencia Artificial, esta etapa es esencial, ya que una mala extracción puede afectar la calidad de los embeddings generados más adelante.

Tecnologías utilizadas:

Se apoya en bibliotecas como PyPDF2 o PyMuPDF para leer el contenido.

Si los archivos están escaneados, puede usar OCR con Tesseract para extraer texto de imágenes.

Resultado esperado:

Una colección de textos limpios y estructurados que representan el conocimiento base del sistema, listos para ser divididos por el agente de segmentación.