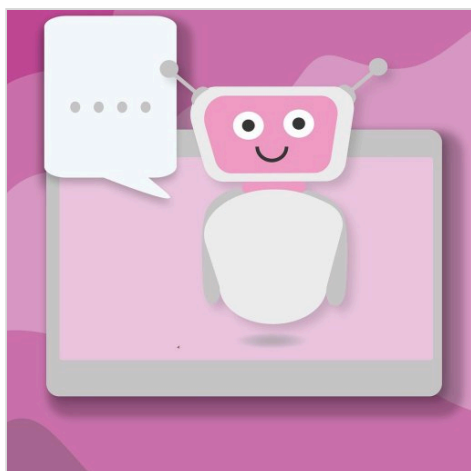


L U M B R E



Extraer y estructurar: del PDF al dato útil

Materia: IA Aplicada, Vibe Coding y Negocios Digitales

07/10/2026

Antes de empezar

Un PDF es texto atrapado. Antes de preguntar o analizar, hay que liberar los datos y ponerlos en columnas. Aquí empieza la magia con documentos.

QUÉ APRENDERÁS

- Entender cómo convertir documentos sueltos (PDF, Word, correos) en datos estructurados y comparables, y por qué la estructura es la base de todo lo demás.

Del párrafo a la columna



Del archivo olvidado al conocimiento consultable: la IA lee y estructura tus documentos.

VIDEO

Extrae información de documentos con Mistral OCR

Del PDF al dato útil: extracción con IA en la práctica. Verás la OCR y estructuración que el curso vuelve sistema consultable.



Escanea para verlo en Lumbre

https://www.youtube.com/watch?v=l5Yx0erB_i4

Una empresa vive llena de información **atrapada**: facturas en PDF, contratos en Word, cotizaciones en correos. El dato existe, pero está encerrado en texto libre, y con texto libre no puedes filtrar, sumar ni comparar. El primer paso para trabajar con documentos es **extraer** los datos que importan y **estructurarlos** en campos con nombre.

Pasar de un PDF de factura a una fila con fecha | proveedor | total | IVA es lo que convierte un archivo en algo consultable. Todo lo demás (buscar, reportar, analizar) se construye sobre esa base.

Primero define los campos

El error es "a ver qué saca la IA". Primero decides **qué campos necesitas** (proveedor, fecha, total, impuestos) y **para qué los usarás** (sumar gastos por mes, detectar pagos duplicados). La extracción es un medio: si no sabes qué pregunta harás después, no sabes qué campo extraer.

Cada tipo de documento pide sus campos; la estructura se diseña según la pregunta de negocio.

Fuente	Qué atrapado hay	Cómo se estructura
Factura en PDF	Monto, fecha, RFC, proveedor	Una fila por factura con esas columnas
Correo de cotización	Cliente, productos, precio	Registro con cliente, ítems y total
Contrato en Word	Partes, vigencia, monto	Ficha de contrato con campos clave
Encuesta en imagen	Respuestas manuscritas	Tabla de respuestas por pregunta

1. Empareja cada término con lo que significa en el flujo de documentos.

1.	Texto libre	-----
1.	Campo estructurado	-----
1.	Extracción	-----
1.	Estructuración	-----

Opciones: Palabras sueltas que no puedes filtrar ni sumar · Un dato con nombre y tipo (fecha, total) · Sacar el dato específico del documento · Ordenar los datos extraídos en columnas

2. Una dueña quiere "saber a qué proveedor le compra más cada mes". ¿Qué necesitas primero?

- ☐ Un chatbot que converse ameno sobre compras.
- ☐ Extraer de cada factura al menos: fecha, proveedor y total, y guardarlos como filas.
- ☐ Un PDF que mezcle todas las facturas del año.
- ☐ Pedirle a la IA que adivine los montos.

3. Si extraes los datos pero los guardas como texto sin campos, ya podrás sumar y filtrar con facilidad después.

- ☐ Verdadero
- ☐ Falso

4. Toma un documento real de un negocio (una factura, un pedido, un reporte). Escribe los 4 o 5 campos que extraerías y, al lado de cada uno, para qué pregunta de negocio serviría.

Si un campo no responde a ninguna pregunta, quizá no hacía falta extraerlo.

Siguiente paso: *Clasificación documental y OCR* — cómo ordenar montañas de archivos por tipo y leer lo que solo existe como imagen.

Referencia curricular

Materia:	IA Aplicada, Vibe Coding y Negocios Digitales
-----------------	---

Versión interactiva completa (con actividades, videos y autoevaluación):



<https://empresa.lumbre.lat/object/iav-c09-r01-0a01>

Material educativo generado desde Lumbre · lumbre.lat