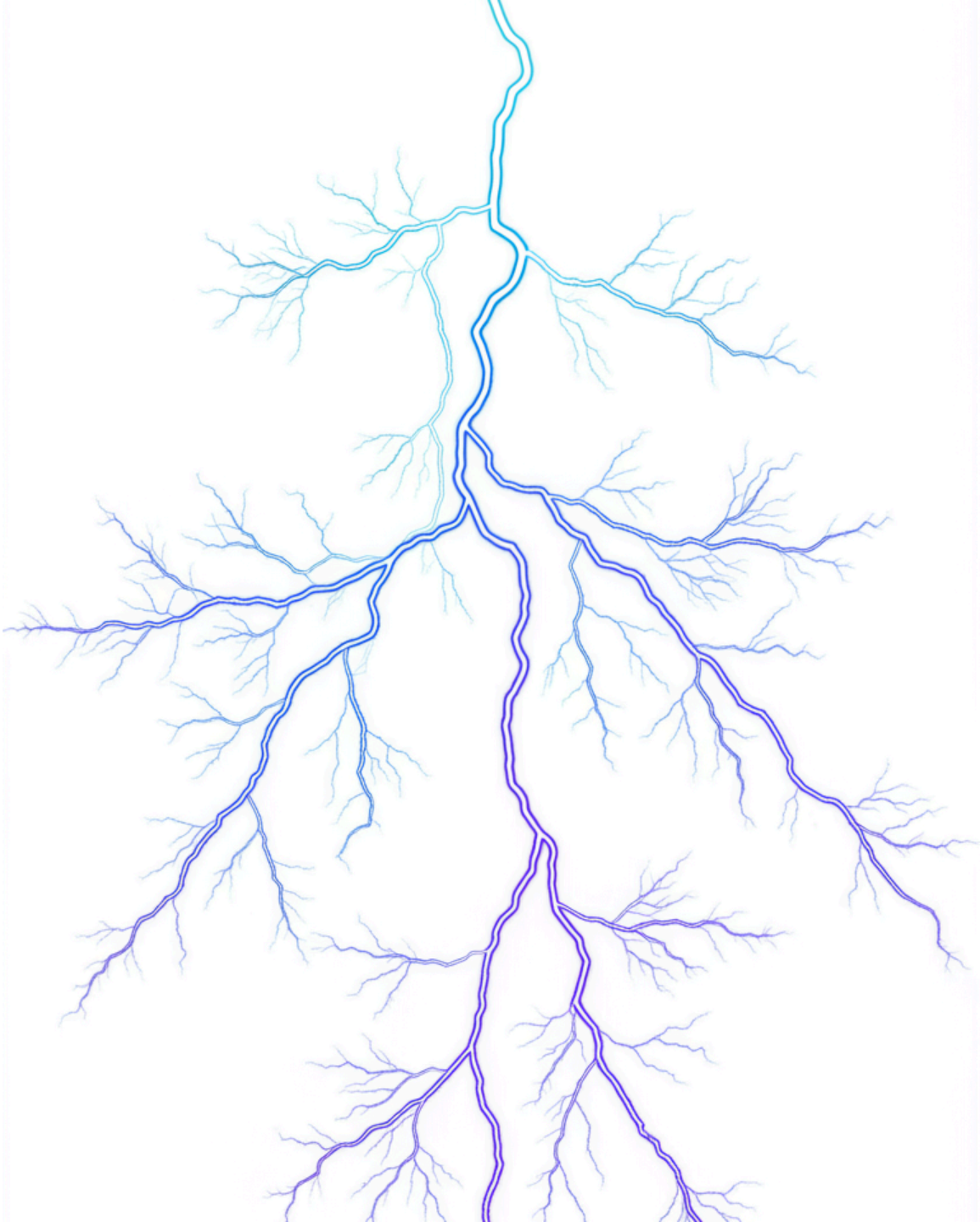
Inteligencia Artificial Generativa Para la Ciencia de Datos

👨‍🏫 Profesor

Juan Camilo Vega Barbosa

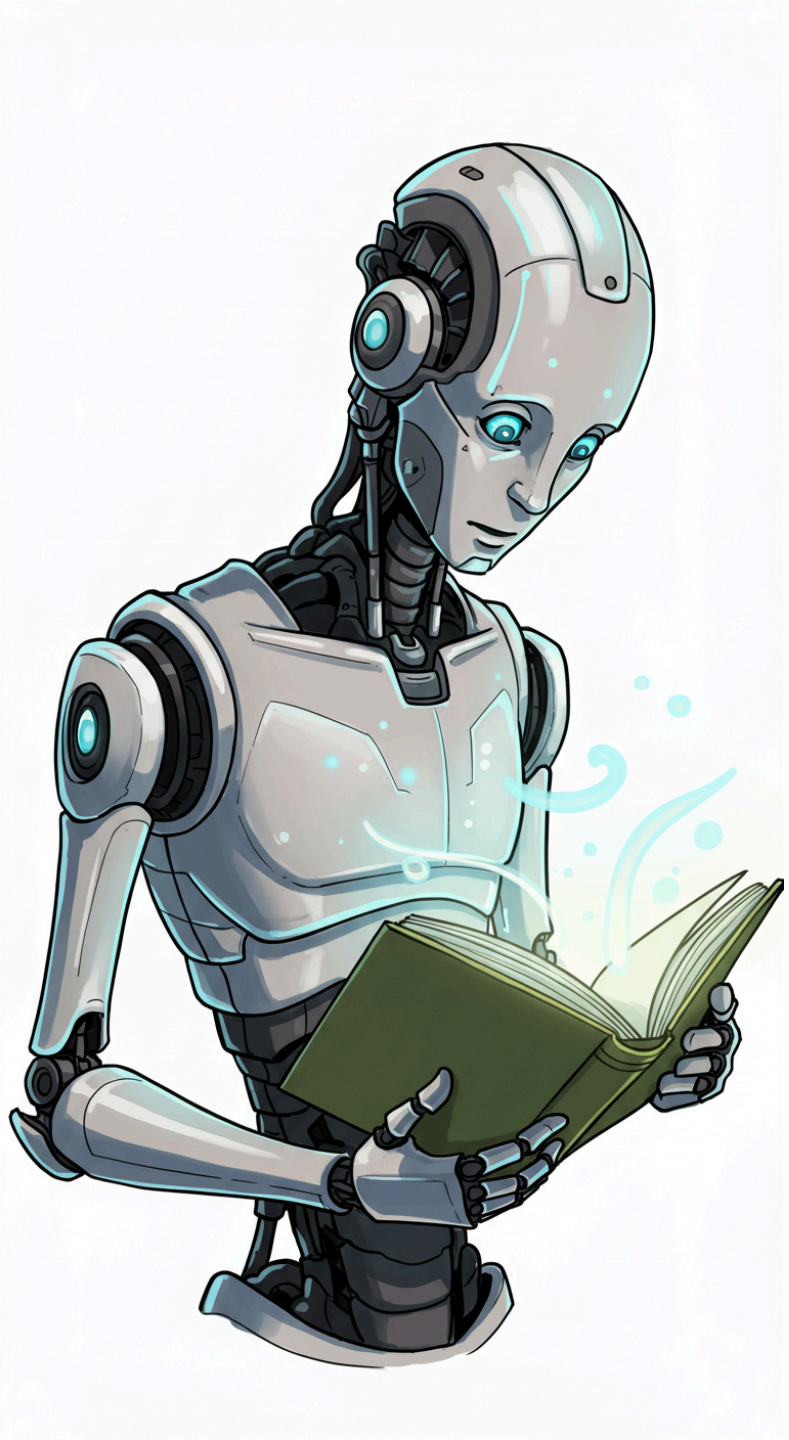
Consultor IA - Ingeniero IA/ML

🔗 LinkedIn



OCR + LLM

Extracción inteligente de datos de documentos



🚀 La revolución del OCR con IA Generativa

El OCR tradicional extrae texto, pero los LLM lo comprenden y estructuran. La combinación de OCR con modelos de lenguaje crea sistemas que no solo leen documentos, sino que los interpretan y organizan según nuestras necesidades específicas.

> Esta fusión permite procesar cualquier documento físico como si fuera una base de datos estructurada, democratizando el acceso a información atrapada en papel o PDFs.

El paradigma evoluciona de "extraer texto" a "entender documentos", creando flujos de trabajo que automatizan completamente el procesamiento de facturas, contratos, reportes médicos y cualquier documento empresarial.

Comparativa de costos: OCR tradicional vs moderno

☁️ Servicios Cloud tradicionales

- **Google Cloud Vision**: $1.50/1000 páginas
- **AWS Textract**: $1.50/1000 páginas
- **Azure Document Intelligence**: $1.50/1000 páginas

🆕 Nuevas alternativas

- **Mistral OCR**: $1.00/1000 páginas
- **Reducto API**: Pricing bajo demanda

🆓 **Tesseract + LLM local:**

- Hardware: $0 (servidores existentes)
- Licencias: $0 (código abierto)
- Procesamiento: Ilimitado

ROI: 95%+ ahorro vs servicios cloud

Tesseract OCR: La base open source

Tesseract es el motor OCR open source más maduro del mundo, desarrollado originalmente por HP y mantenido por Google desde 2006. Su combinación con LLM crea sistemas de extracción de datos prácticamente gratuitos.

Ventajas clave

- **Costo cero**: Sin límites de páginas ni API calls
- **Control total**: Personalización y fine-tuning
- **104 idiomas**: Soporte multilingüe nativo
- **LSTM neural networks**: Precisión moderna

🔗 **Documentación:**
github.com/tesseract-ocr/tesseract

Para empresas: Sistema auto-hospedado sin dependencias externas

Arquitectura Tesseract + LLM

Pipeline híbrido de extracción inteligente:

1. **Análisis del documento ejemplo** → LLM define estructura JSON automáticamente
2. **OCR batch con Tesseract** → Extracción de texto sin costo
3. **Procesamiento con LLM** → Mapeo inteligente a campos estructurados
4. **Validación y salida** → JSON estructurado con metadatos

Componentes técnicos

- **pytesseract**: Wrapper Python optimizado
- **Groq API**: LLM rápido para estructuración
- **pdf2image**: Conversión PDF automática
- **Batch processing**: Escalamiento industrial

Resultado: Sistema que aprende del primer documento y aplica la misma estructura a miles de documentos similares

Casos de uso: Tesseract + LLM

📄 Facturas y recibos

- Extracción de campos clave automática
- Validación de totales y cálculos
- Integración con sistemas ERP

📋 Formularios empresariales

- Procesamiento de aplicaciones
- Digitización de encuestas físicas
- Automatización de workflows

🏥 Documentos médicos

- Historiales clínicos estructurados
- Reportes de laboratorio
- Prescripciones médicas

Ventaja competitiva: Sin límites de volumen ni costos variables. Perfecto para grandes volúmenes de documentos similares.

Mistral OCR: El nuevo estándar

Mistral OCR representa la evolución del OCR con IA nativa, procesando 2,000 páginas por minuto con comprensión contextual de elementos como tablas, ecuaciones y gráficos complejos.

Capacidades avanzadas

- **Multimodal nativo**: Texto, tablas, imágenes, ecuaciones
- **Multilingual**: 100+ idiomas sin configuración
- **Self-hosteable**: Deploy local disponible
- **API simple**: Integración en minutos

🔗 **Documentación:**
mistral.ai/news/mistral-ocr

Pricing: $1.00/1000 páginas
Performance: 2K páginas/minuto

Arquitectura Mistral OCR

Pipeline nativo multimodal:

1. **Upload directo** → PDF/imágenes a API Mistral
2. **OCR inteligente** → Detección automática de elementos
3. **Structured output** → Markdown + JSON estructurado
4. **LLM analysis** → Extracción de insights y metadatos

Ventajas vs competencia

- **94.9% accuracy** vs 88.49% (Gemini 2.0)
- **Incluye images**: Extracción de imágenes embebidas

Ideal para: Documentos complejos con layouts variables, reportes científicos, y procesamiento que requiere alta precisión inmediata

Implementación práctica: Mistral OCR

🔧 Setup mínimo

```
from mistralai import Mistral
client = Mistral(api_key=api_key)

# Upload y procesamiento
uploaded_file = client.files.upload(
    file=pdf_content,
    purpose="ocr"
)

response = client.ocr.process(
    document=DocumentURLChunk(...),
    model="mistral-ocr-latest"
)
```

📊 Outputs múltiples

- **Markdown estructurado** para análisis
- **Imágenes extraídas** como archivos separados
- **JSON metadata** con estructura detectada
- **Full text** para búsqueda y indexación

Automatización completa: De PDF a datos estructurados en una sola llamada API

Reducto: OCR especializado para documentos complejos

Reducto se enfoca en los casos más difíciles: periódicos antiguos, documentos con layouts complejos y estructuras no estándar donde otros sistemas fallan.

Especialización única

- **Agentic OCR**: Auto-corrección con múltiples pasadas
- **Complex layouts**: Multi-columna, tablas anidadas
- **Vision + LLM**: Combinación de modelos especializados
- **Custom schemas**: Definición de outputs específicos

🔗 **Más información:**
reducto.ai

Target: Empresas Fortune 10
Fortaleza: Documentos "imposibles"
Pricing: Enterprise (contactar ventas)

Casos de uso avanzados: Reducto

📰 Periódicos históricos

- Layouts multi-columna complejos
- Calidad de escaneo variable
- Fonts antiguos y deteriorados

📊 Reportes financieros

- Tablas con celdas fusionadas
- Gráficos embebidos
- Múltiples formatos en un documento

🏛️ Documentos legales

- Estructuras jerárquicas complejas
- Anotaciones y marcas manuscritas
- Formatos no estándar

> **Diferenciador:** Cuando Tesseract y Mistral no son suficientes, Reducto maneja los casos extremos con precisión enterprise-grade

Comparativa final: ¿Cuál elegir?

🆓 Tesseract + LLM

Ideal para:

- Grandes volúmenes consistentes
- Presupuesto limitado
- Control total del pipeline
- Documentos con formato predecible

🚀 Mistral OCR

Ideal para:

- Documentos multimodales complejos
- Necesidad de precisión inmediata

💎 Reducto

Ideal para:

- Documentos "imposibles"
- Casos enterprise críticos
- Layouts extremadamente complejos
- Presupuesto enterprise

Recomendación:
Prototipo → Mistral OCR
Producción high-volume → Tesseract
Casos complejos → Reducto

📚 Recursos del curso

🔗 Enlaces esenciales

🛠️ OCR Engines

- Tesseract OCR
- Mistral OCR
- Reducto API

📖 Documentación técnica

- pytesseract Documentation
- Mistral API Docs

🗂️ GitHub del curso

🔳 Script Sesión 5

👨‍🏫 Contacto:

LinkedIn - Camilo Vega

> **Próxima sesión:** Implementación práctica de sistemas OCR híbridos en producción