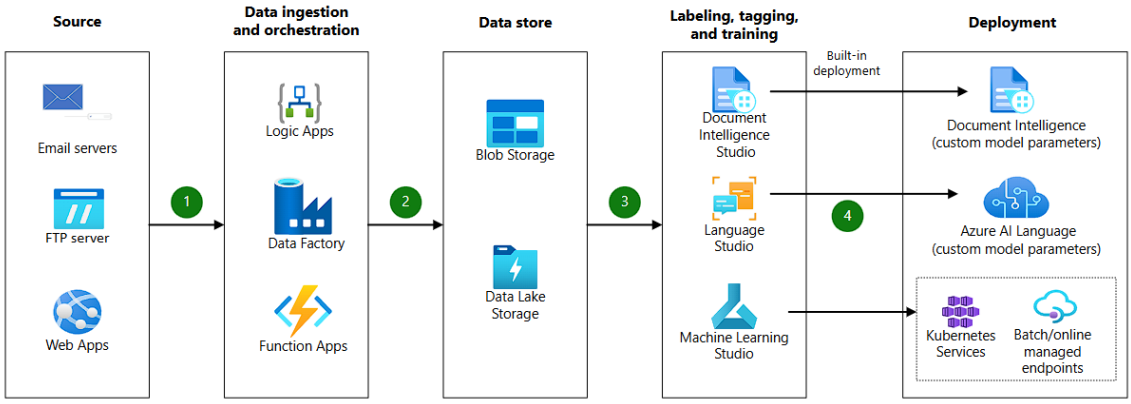
Azure Azure Form Recognizer-Azure Cognitive Services y Azure Machine Learning

Crear un modelo de IA para la extracción de datos sobre currículos (CVs) en PDF de científicos de datos usando Azure puede ser un proceso totalmente mas eficiente .

El enfoque principal será usar servicios en la nube de Azure, como Azure Form Recognizer, Azure Cognitive Services, y Azure Machine Learning, combinados con un pipeline para preprocesamiento y modelado.



1. Recolección de datos y preprocesamiento de los PDFs

- Obtención de los CVs en formato PDF:.

- Preprocesamiento de los PDFs: Los PDFs pueden tener texto incrustado o ser imágenes (escaneados). Si los archivos son imágenes, es necesario aplicar OCR (Reconocimiento Óptico de Caracteres) para extraer el texto. Azure proporciona herramientas como Azure Cognitive Services - Computer Vision para realizar OCR.

2. Uso de Azure Form Recognizer para extracción de datos

- Azure Form Recognizer es una herramienta potente para extraer datos de documentos como formularios y currículos. Ofrece dos enfoques principales:

- Plantillas predefinidas: Si los CVs tienen una estructura estándar (por ejemplo, se repiten ciertos campos como nombre, dirección, experiencia laboral), puedes usar la opción de Plantillas predefinidas de Form Recognizer.

- Entrenamiento personalizado: Si los CVs tienen formatos variados, puedes entrenar un modelo ****customizado**** con Azure Form Recognizer. Para hacerlo:

Etiquetado de datos: Necesitarás etiquetar una muestra de CVs, marcando las secciones relevantes (por ejemplo, nombre, educación, experiencia, habilidades, etc.).

Entrenamiento del modelo: Usando Azure Form Recognizer, puedes entrenar el modelo para que aprenda a extraer los datos clave .

- Ejemplo: Se puede entrenar un modelo para identificar las secciones comunes en un CV de un científico de datos, como su educación, habilidades técnicas, herramientas y lenguajes de programación, proyectos, etc.

3. Extracción y estructura de los datos

Una vez entrenado el modelo, se puede usar el Azure Form Recognizer para extraer los datos. Los resultados se entregarán en un formato estructurado (por ejemplo, JSON), que incluirá los campos identificados y su valor correspondiente (por ejemplo, nombre, fecha de nacimiento, experiencia, etc.).

Consideraciones

- Datos estructurados vs no estructurados: Si los CVs no siguen una estructura fija, la extracción de datos puede ser menos precisa. En este caso, es útil implementar una validación adicional con reglas específicas para asegurar que los datos extraídos sean correctos.

4. Postprocesamiento y almacenamiento de datos

- Después de extraer los datos, puede ser necesario hacer un postprocesamiento para corregir posibles errores en la extracción, como formateo de fechas o valores numéricos.

- Los datos extraídos deben ser almacenados en una base de datos o un almacén adecuado. Puedes usar Azure SQL Database o Azure Cosmos DB para almacenar la información estructurada.

5. Entrenamiento y optimización de un modelo personalizado (opcional)

Si necesitas mejorar la precisión del modelo, puedes utilizar Azure Machine Learning para crear un modelo más avanzado de extracción de datos basado en técnicas de aprendizaje automático o procesamiento de lenguaje natural (PLN).

6. Uso de Azure Logic Apps o Power Automate para automatización

Si deseas automatizar el proceso de recolección y extracción de datos de manera continua, puedes usar Azure Logic Apps o Power Automate para crear flujos de trabajo automáticos. Esto te permitirá:

- Subir automáticamente nuevos PDFs a un almacenamiento en la nube (como Azure Blob Storage).

- Procesar los PDFs de manera automática usando Form Recognizer y extraer los datos.

- Almacenar los datos extraídos en una base de datos o enviarlos a un sistema externo para su procesamiento posterior.

7. Evaluación y mejora continua

Una vez que el sistema esté en funcionamiento, es importante seguir evaluando su precisión y mejorarlo. Puedes:

- Ajustar los modelos de extracción según los resultados que obtengas.

- Incorporar feedback humano para corregir y mejorar los resultados a medida que el sistema extrae más datos de currículos adicionales.

Resumen del flujo de trabajo en Azure

1. Recolección de CVs (PDF).

2. Preprocesamiento (OCR si es necesario).

3. Uso de Azure Form Recognizer para extraer datos de forma estructurada.

4. Postprocesamiento y almacenamiento en bases de datos de Azure.

5. Automatización con Azure Logic Apps o Power Automate.

6. Entrenamiento y optimización con Azure Machine Learning si se requiere mejorar la precisión.

7. Evaluación continua del modelo y mejora del proceso.