Captura de Datos Automatizada con IA

Proyecto Final de Python 2025

Alumnos
Juan Pablo Jouanny
Lucas Lamiño

Profesor
Nuria Torres

Objetivo del proyecto

El **objetivo** del programa es extraer información numérica de imágenes que contienen tablas utilizando técnicas de Reconocimiento Óptico de Caracteres (OCR), organizar estos datos en un formato tabular (DataFrame) y realizar análisis visuales mediante gráficos que permitan identificar patrones y comportamientos relevantes en los datos extraídos.

Adicionalmente, busca automatizar el procesamiento de información para generar reportes técnicos con los resultados obtenidos, lo que facilita la interpretación y el uso de la información en diferentes contextos industriales o académicos.

Breve explicación del programa

El programa utiliza OCR para extraer datos numéricos de imágenes de tablas, los organiza en un DataFrame y genera gráficos para analizar patrones y comportamientos. También permite crear reportes técnicos automatizados con los resultados obtenidos.

Explicación de cómo funciona la extracción de datos con OCR

El programa emplea OCR (Reconocimiento Óptico de Caracteres) para identificar y extraer datos numéricos desde imágenes. A continuación, se detalla el proceso:

1. **Inicialización del lector OCR**
   Se utiliza la biblioteca EasyOCR, configurada para soportar los idiomas español e inglés, asegurando una interpretación precisa de los textos en la imagen.
2. **Lectura de la imagen**
   El programa carga la imagen especificada y extrae los textos encontrados. EasyOCR devuelve una lista con los textos reconocidos.

3. **Filtrado de datos numéricos**
   - Cada texto extraído es procesado para eliminar caracteres no deseados, como espacios en blanco.
   - Se verifica si el texto puede convertirse a un número válido (reemplazando comas por puntos para valores decimales).

   - Solo los valores numéricos son retenidos y organizados en filas, asegurando que tengan un formato compatible con tablas.
4. **Validación de consistencia:**
   Se verifica que el número total de datos numéricos sea múltiplo de las columnas

esperadas en la tabla. Esto garantiza que los datos extraídos puedan organizarse correctamente en filas y columnas.

5. **Creación de un DataFrame:**
   Los datos extraídos se estructuran en un DataFrame de Pandas, asignando nombres significativos a las columnas.
   Se convierten todas las columnas a tipos numéricos para facilitar su análisis y procesamiento posterior.
6. **Resultado:**
   Si el proceso es exitoso, el programa retorna un DataFrame con los datos organizados. En caso de errores, muestra mensajes descriptivos, como que el archivo no fue encontrado o que hubo problemas durante la extracción.

Cómo usar el programa paso a paso

1. **Preparar la imagen para la extracción**
   Coloca en la misma carpeta donde está el programa el archivo de imagen caldera_table_50_observations (por ejemplo, caldera_table_50_observations.png). Esta imagen contiene la tabla con los datos que se quieren extraer.
2. **Ejecutar el programa**
   Abre una terminal o consola y navega hasta la carpeta donde está el archivo ocr_a_csv.py. Ejecuta el programa con el siguiente comando:
   python ocr_a_csv.py
3. **Extracción automática de datos**
   El programa leerá la imagen caldera_table_50_observations y extraerá automáticamente los datos numéricos de la tabla mediante OCR, sin que sea necesario realizar ninguna acción adicional.
4. **Generación del archivo CSV con los datos extraídos**
   Una vez finalizada la extracción, se generará un archivo llamado caldera_datos.csv en la misma carpeta, que contiene la tabla organizada con los datos extraídos de la imagen.
   Generación del informe y gráficos de análisis
   El programa también generará un informe llamado informe_analisis_caldera.docx que incluye gráficos y análisis visuales de los datos extraídos para facilitar la interpretación.
5. **Verificación de resultados**
   Para revisar los datos, abre el archivo caldera_datos.csv. Para interpretar los resultados y ver las visualizaciones, abre el documento informe_analisis_caldera.docx.

Resultados obtenidos

Luego de ejecutar el programa ocr_a_csv.py sobre la imagen caldera_table_50_observations.png, se generó un archivo CSV llamado caldera_datos.csv

que contiene los datos numéricos extraídos de forma automática. A continuación, se muestra un fragmento de los datos obtenidos:

Presión (bar)	Temperatura (°C)	Caudal (m^3/h)	Nivel de Agua (%)	Consumo de Combustible (L/h)	CO (%)	NOx (%)	Horas Operadas
10.5	150.2	120.0	75.0	40.5	0.12	0.03	5
10.7	149.8	118.5	74.5	41.0	0.11	0.04	6
10.4	150.0	119.8	75.2	40.7	0.13	0.03	7

Estos datos permiten realizar análisis posteriores, como la generación de gráficos de evolución temporal de las emisiones y el consumo de combustible, y el análisis de eficiencia en función del caudal, automatizados dentro del mismo programa.