Proyecto OCR en Python

Autor: Luis Carlos Pacheco Ramirez

Requerimientos funcionales:

- **R1 Carga de archivos PDF con INE escaneada:**
  - El programa debe permitir al usuario cargar archivos PDF que contengan imágenes escaneadas de credenciales de identificación (INE), se utilizaran las librerías Pillow y pdf2image para este proceso.
  - Las imágenes deben ser extraídas del PDF para su posterior procesamiento.

- **R2 Preprocesamiento de imágenes:**
  - El sistema debe aplicar técnicas de preprocesamiento a las imágenes antes de ejecutar el OCR, para esto se usará la librería OpenCV.
  - Las técnicas pueden incluir:

- Redimensionamiento: Ajustar el tamaño de las imágenes para una mejor legibilidad.

- Recorte: Eliminar áreas no relevantes o ruido alrededor de la INE.

- Binarización: Convertir la imagen a blanco y negro para facilitar la detección de texto.

- Eliminación de ruido: Reducir artefactos o imperfecciones en la imagen.

- Ajuste del tamaño de la letra: Engrosamiento o adelgazamiento del contenido de la imagen para hacer más legible el texto

- Enderezamiento de la imagen: Predicción de orientación y acomodo de la rotación de la imagen para mejor detección de caracteres

- Adición de "Bounding boxes": Creación de cajas de filtrado de texto

- **R3 Extracción de texto mediante pytesseract:**
  - Utilizar la biblioteca pytesseract para extraer el texto de las imágenes preprocesadas.
  - Configurar correctamente los parámetros de pytesseract para obtener resultados óptimos.

- **R4 Identificación de campos relevantes:**
    - El programa debe ser capaz de identificar y extraer información específica de la INE, como:

Nombre completo.

Clave de Elector.

Fecha de nacimiento.

Domicilio.

- **R5 Validación de resultados:**
    - Verificar la precisión del OCR comparando los resultados extraídos con los datos reales de la INE.
    - Implementar mecanismos para corregir errores o solicitar intervención manual si es necesario.

Requerimientos NO funcionales:

- **RN1 Eficiencia:**
    - El sistema debe procesar las imágenes y extraer el texto de manera rápida y eficiente.
    - El tiempo de respuesta debe ser aceptable incluso para archivos PDF grandes.

- **RN2 Precisión:**
    - El OCR debe ser altamente preciso en la extracción de texto.
    - Minimizar errores en la interpretación de caracteres y palabras.

- **RN3 Documentación:**
    - Documentación completa y detallada del programa, incluyendo:

- Instrucciones de instalación y uso.

- Descripción de las funcionalidades y opciones disponibles.

- Especificaciones técnicas del programa.

- Ejemplos de uso y casos de prueba.

- **RN4 Mantenimiento y actualizaciones:**
  - El programa debe ser fácil de mantener y actualizar.
  - Se debe ofrecer soporte técnico para solucionar problemas y responder preguntas.
  - El programa debe ser compatible con diferentes versiones de Python y sistemas operativos.
  - Asegurar que las bibliotecas utilizadas sean estables y bien mantenidas.

- **RN5 Licenciamiento:**
  - El programa debe tener una licencia clara y permisiva que permita su uso y distribución.

- **RN6 Seguridad:**
  - El programa debe proteger la privacidad de los datos personales contenidos en la INE.
  - Evitar fugas de información o acceso no autorizado.

- **RN7 Consideraciones adicionales:**
  - Soporte para diferentes formatos de INE (versiones antiguas y nuevas).
  - Capacidad para procesar imágenes con diferentes niveles de calidad.
  - Implementación de técnicas de aprendizaje automático para mejorar la precisión del OCR.
  - Integración con otras aplicaciones o sistemas.

- **RN8 Usabilidad:**
  - Si se implementa una interfaz gráfica, esta debe ser intuitiva y fácil de usar.
  - Proporcionar retroalimentación adecuada al usuario durante el proceso de carga y procesamiento.

- **RN9 Rendimiento:**
  - Evaluar y optimizar el uso de recursos (CPU, memoria) para un rendimiento óptimo.
  - Minimizar la carga en el sistema durante el procesamiento.