Extracting Text from Images using Python and Tesseract OCR

Objective:

To extract text from a series of images using the Pillow library for image processing and the pytesseract library for Optical Character Recognition (OCR).

Overview:

processes a list of image files, extracting text from each image and saving the extracted text into a single text file. This can be particularly useful for converting scanned documents or images with text into machine-readable text format.

Steps

1. **Import Libraries:**
   - PIL (Python Imaging Library) from Pillow to handle image file operations.
   - pytesseract for OCR to convert images into text.
   - UnidentifiedImageError to handle errors related to unidentified image files.
2. **List of Image Paths:**
   - A list of file paths for the images to be processed is defined. Each path points to an image file stored on the local machine.
3. **Function to Extract Text:**
   - extract_text(image_path): This function attempts to open an image file and extract text from it using pytesseract. If the image file cannot be identified, it returns an error message specific to that image file.
4. **Extract Text from All Images:**
   - A list comprehension is used to apply the extract_text function to each image in the list, storing the results in the texts list.
5. **Save Extracted Text to a File:**
   - The script opens a new text file in write mode with UTF-8 encoding.
   - It iterates over the extracted texts, writing each text along with its corresponding image number to the file. This ensures the extracted texts are organized and easy to read.
6. **Output Confirmation:**
   - Finally, the script prints a confirmation message indicating that the extracted texts have been successfully saved to the specified output file.