TEXT EXTRACTION PROJECT

Introduction:

Text extraction is an essential task in various fields such as document management, data analysis, and information retrieval. With the advancement in technology, automated text extraction has become a need for many businesses to save time and reduce errors. In this mini-project, we aim to develop a text extraction system using Python, OpenCV, Tesseract, and other relevant libraries.

Overview:

The text extraction system developed in this project extracts text from images and converts it into machine-readable text. The system can handle images with various types of text, including handwritten text, printed text, and text with different font styles and sizes. The extracted text is then saved in a file for further processing or analysis.

Method:

The text extraction system developed in this project follows the following steps:

Image Acquisition: The first step is to acquire the image containing the text. The image can be acquired from various sources, such as a scanner, camera, or downloaded image.

Preprocessing: The acquired image is preprocessed to enhance the quality of the text. The preprocessing steps include noise reduction, thresholding, and image smoothing.

Text Detection: The preprocessed image is then processed for text detection. The system uses OpenCV's text detection algorithm to identify the location of the text in the image.

Text Recognition: Once the text location is identified, the system uses Tesseract OCR to recognize the text. Tesseract OCR is an open-source OCR engine that can recognize various fonts and text sizes.

Post-processing: The recognized text may contain errors due to noise, low-quality image, or complex text layout. To minimize these errors, the recognized text is post-processed by

applying text normalization techniques such as spelling correction, punctuation removal, and stop-word removal.

Output: The final step is to save the extracted text into a file format such as txt, doc, or pdf.

Relevant Libraries:

The following libraries are used in this project:

OpenCV: OpenCV is an open-source computer vision library used for image and video processing.

Tesseract OCR: Tesseract OCR is an open-source OCR engine that can recognize various fonts and text sizes.

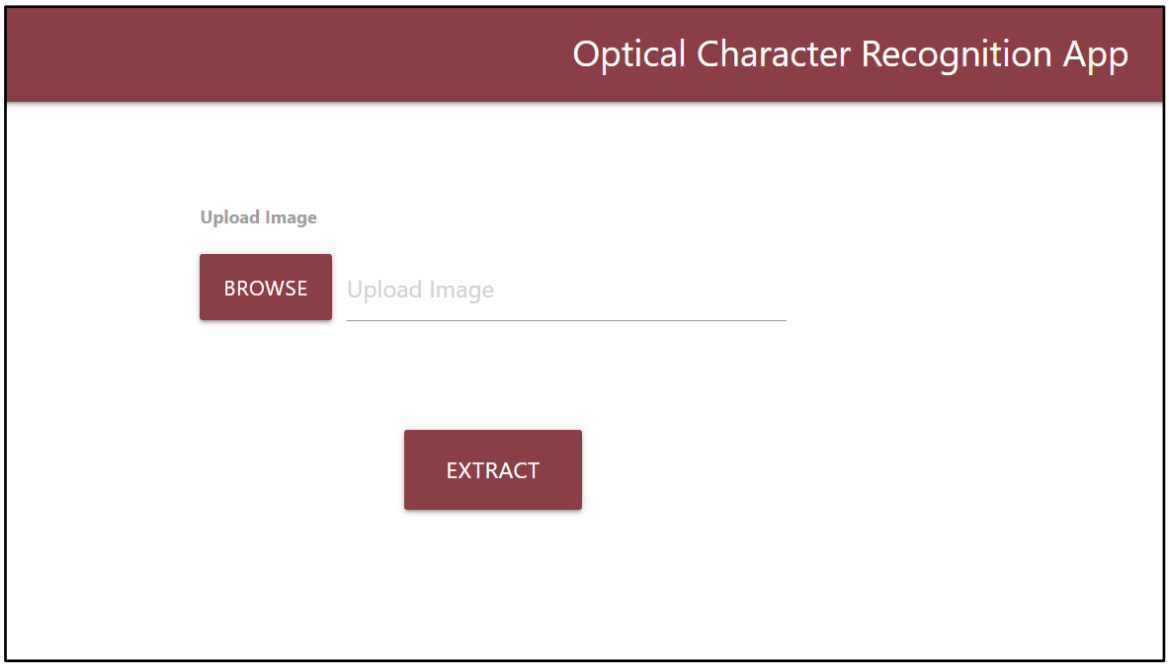
Pytesseract: Pytesseract is a Python wrapper for Tesseract OCR, making it easier to use in Python-based projects.

Numpy: Numpy is a numerical computing library in Python used for array operations.

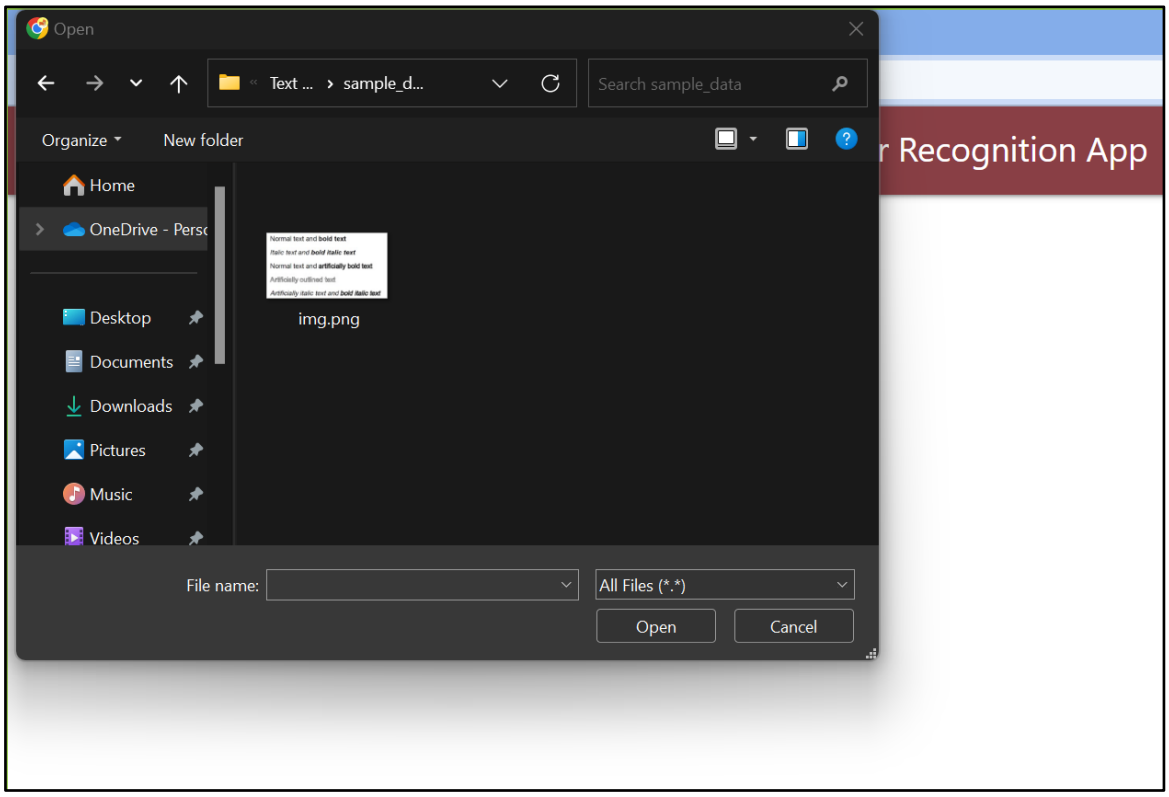
Pandas: Pandas is a data manipulation library in Python used for data analysis.

Conclusion:

The text extraction system developed in this mini-project can extract text from various types of images and convert it into machine-readable text. The system uses OpenCV's text detection algorithm and Tesseract OCR for text recognition. The extracted text can be saved in a file format for further processing or analysis. The use of Python and relevant libraries has made the development of this system efficient and effective. The system can be further improved by incorporating deep learning techniques for better accuracy and performance.



Initial Page of the application



Upload Section for the extraction

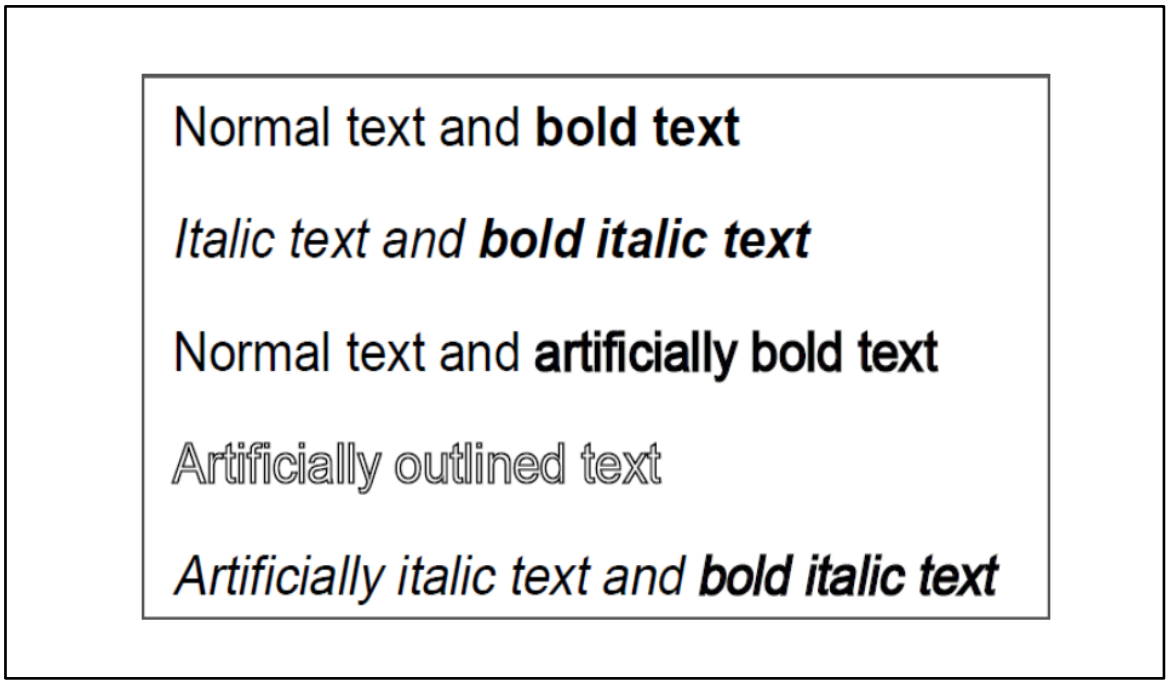


Image Uploaded

Extracted

Normal text and bold text

Italic text and bold italic text

Normal text and artificially bold text

Artificially outlined text

Artificially italic text and bold italic text

Text Extracted