Script Documentation

Assumptions:

- PDF's can contain multiple images.
- Tesseract OCR is used for text extraction.
- Any sort of non-empty key-value pair data is extracted and only the first occurrence of ':' is considered.
- No image preprocessing is done.
- Table extraction is not done as it's a separate project in itself.

Challenges:

For extracting tabular data, I was looking to detect table starting and ending, so as to extract the two information separately. But it turned out that table extraction is a separate project in itself and is beyond the scope of this task.

Working:

The script uses Tesseract OCR, OpenCV and Numpy and pdf2Image.

1. Processing the files:

This is done via the **process_files** function.
For PDFs, first pdf2image is used to convert the PDFs to OpenCV images and then OCR is done. And for image files, they are read as OpenCV images, and then OCR is done.

2. Performing OCR and saving CSV files:

This is done via the **ocr_and_save** function and a helper function called **extract_key_value_pairs**.
A key_value_pairs dictionary is created to store the key-value pairs returned by **extract_key_value_pairs** function.
Then this data is written to a CSV file.

extract_key_value_pairs function uses PyTessaract's image_to_string function to convert the OpenCV images of the documents to a string of extract text. Then new lines are split based on the '\n' character. And then for each line, key-value pairs are extracted based on the first occurrence of ':' and this dictionary of key-value pairs is then returned.