Libraries for Text Extraction from Image

1. PYTESSERACT

- I tried this library on structured , unstructured and semi-structured data and have different insights.

- This library when used on structured data , does not extract a lot of data , can said only 10% of the data or text is extracted , so we cannot use this library on structured data.

- This library when used on unstructured data , extract almost 100% of the text , so we can use this library On unstructured data.

- This library when used on semi-structured data , extract a lot of data , can be said 80% , so can be considered .

2. PYOCR

- This library when used on structured data , only extract some amount of data , can be said as 15% , so we cannot use it for structured data.

- This library when used on unstructured data , extract 100% of the data , so can be used in unstructured data.

- This library when used on semi-structured data , extract almost all the data , can be said 95% , so can be used In semi-structured data extraction.

3. Open cv2

- This library when used on structured data , extract only 15% of the data , so can not be considered.

- This library extract all the data in unstructured data , can be said 100% .

- This library when used on semi structured data, does not work very well , i.e. extract only 50% of the data.

4. Tried converting the given dataset images into pdf and then extracting the text.

I converted the png image to pdf and then tried pdfminer Library to extract text but , no text was extracted , can be said 0% . so this didn’t work , but when I gave a unstructured normal pdf , then it extracted all its data , i.e. 100% of the text.