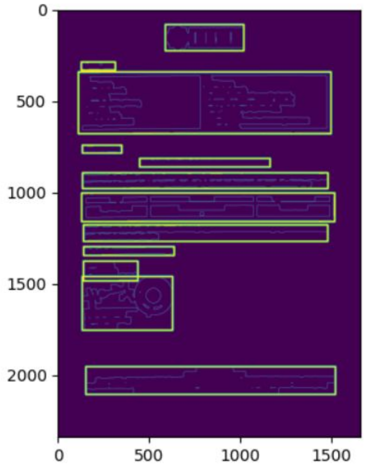
Question 3 Part A Report

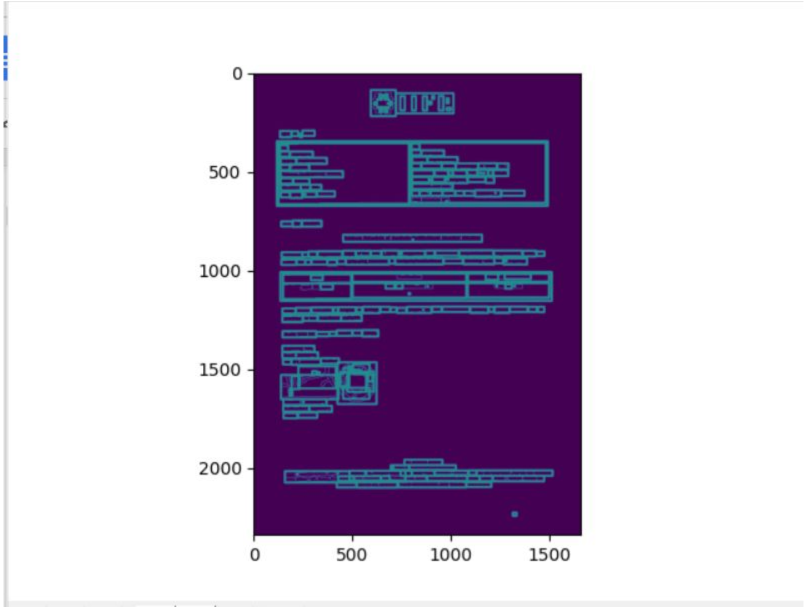
Link to the MongoDB collection :
https://drive.google.com/file/d/1JFPzure7VNb577kg1STRpYBicPdU-T8A/view?usp=sharing
I have read about libraries in python like tabula. However, with the given samples, tabula doesn't work well. For example, in sample 3, it is unable to detect the table as it gets confused due to the vertical line in the middle of the pdf. Other libraries were ineffective as well, and they are only able to detect horizontal and vertical solid lines correctly.
So I decided to use tesseract and opencv to improve on this.
So, this is the higlighting done for each pdf. There are two images per pdf, as one only shows the outer boundaries of all the boxes detected in the pdf and the second image shows all the boxes detected in the pdf.
STEPS used:

1. Dilate all the text so that paragraphs are merged into one box
2. Now mark all the boxes with their outer boundary
3. Use tesseract module of python to read the text within the boxes. Now, for tables , in general, it is noticed that the space between the words is much larger than the text in paragraphs. So using this property and that tesseract can output strings with space preserved as in the document as well as with space truncated, to differentiate between tables and paragraphs, I compared two outputs of tesseract and if they weren't equal, they were added to the tables database.

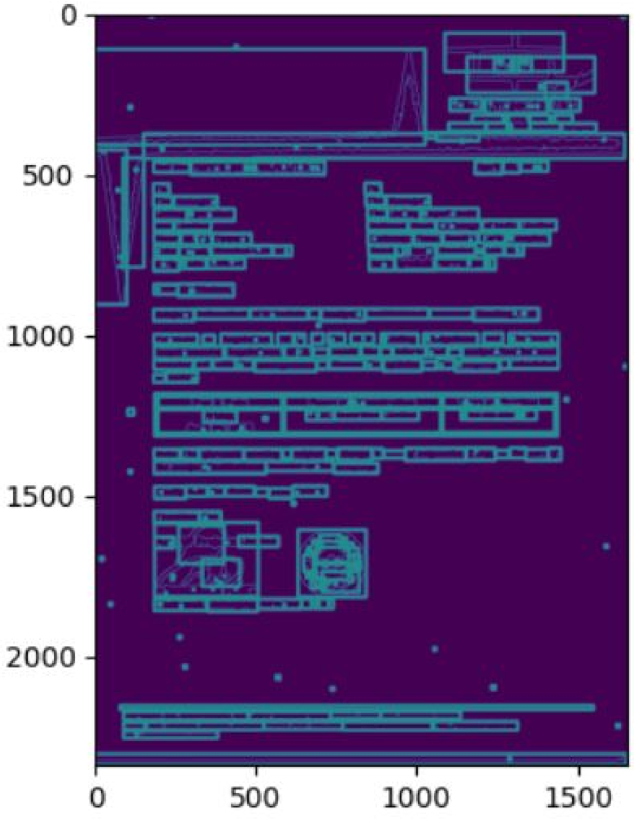
IMPROVEMENTS That can be done:

1. To differentiate between table and other irregular stuff like signatures, and logos with text, we can do the following:
    a. Analyze all the boxes that have the potential to be tables
    b. Now, we have the inner rectangles inside the boxes(2nd image in every sample)
    c. So using an algorithm to analyse the position of the rectangles(like in a table all the rectangles in same row have same x coordinates)
2. Now this table works with normal size text. For bigger text sample either the user should be able to input the text size or there can be libraries that can detect the average size of text.

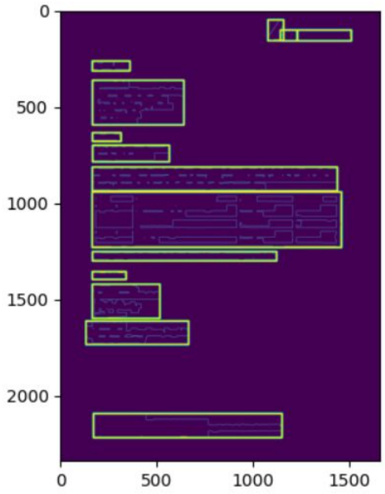

0
500
1000
1500
2000
0
500
1000
1500

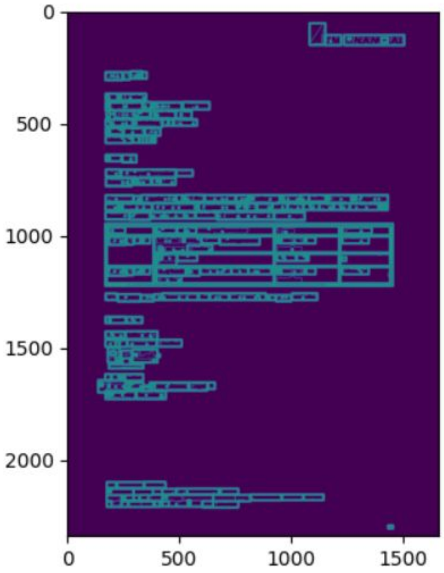

0
500
1000
1500
2000
0
500
1000
1500

SAMPLE

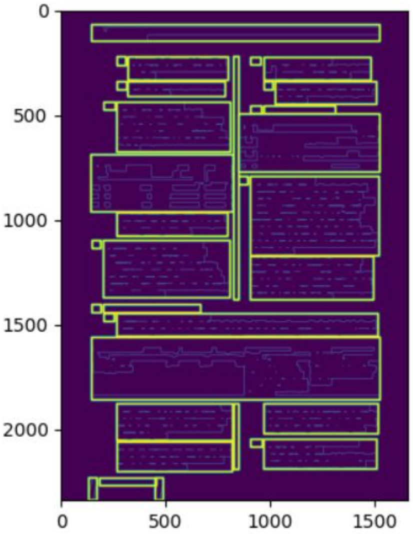


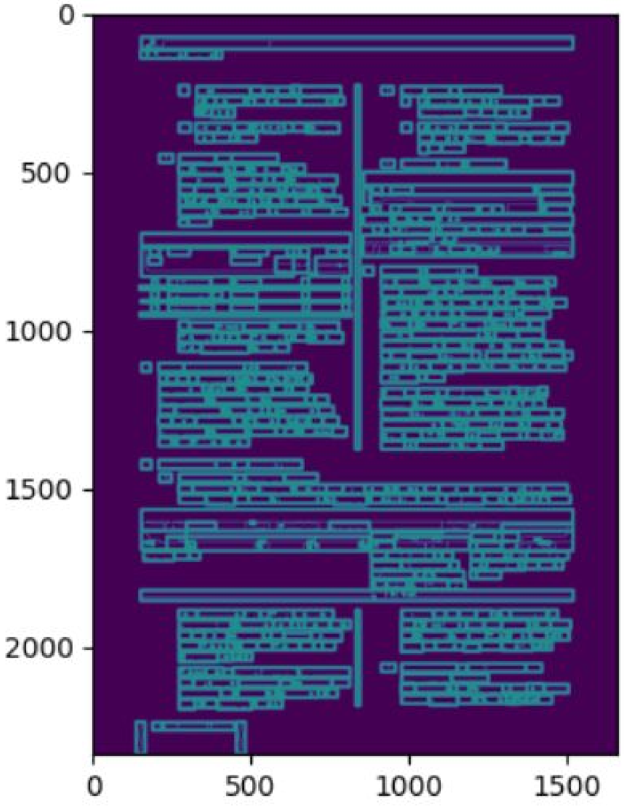
SAMPLE 1


0
500
1000
1500
2000
0
500
1000
1500


0
500
1000
1500
2000
0
500
1000
1500

SAMPLE 2


0
500
1000
1500
2000
0
500
1000
1500


0
500
1000
1500
2000
0
500
1000
1500

SAMPLE 3