Practical 1
2CS501
19BCE248

AIM:- Use pytesseract library in Python for optical character recognition from
(i) an image file
(ii) a multi-page pdf file.

What is OCR?
OCR stands for "Optical Character Recognition". It is basically used for extracting text from digital images. Used in applications like google lens.

What is pytesseract?
Pytesseract is an open source library provided by google to achieve OCR in an efficient manner.

Task Performed:

- In the image to text recognition only a single function (image_to_string(image)) was enough to achieve the required task.
- In the second half for reading pdf files a new library named fitz was used to just zoom in- zoom out and other functionalities like no. of pages.
- All the pdf pages were converted to images and then using the above function I was able to extract the text.

Conclusion:
These libraries are very handy for overcoming various tasks in such a convenient way with few lines and code and provides a good abstraction level for beginners.