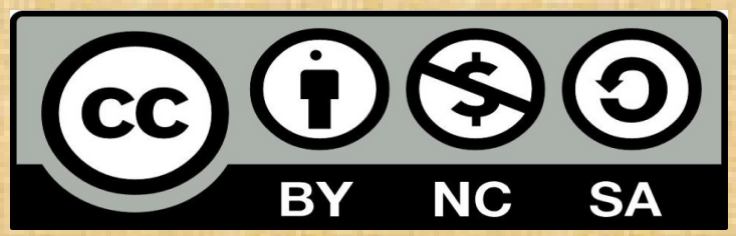
OCR

OPTICAL CHARACTER RECOGNITION



Why I choose This Topic?

- I Love Three Things:
  Automation, Automation and Automation
- Suppose You have typed a document and taken it's printout..
- Now your document is deleted..

No way to recover it..

You have just printout..

Then OCR will help you to recover your document.

How OCR is Used ??

- Have you ever used Google Dork ?
- If yes ? Then This dork ?

filetype:pdf

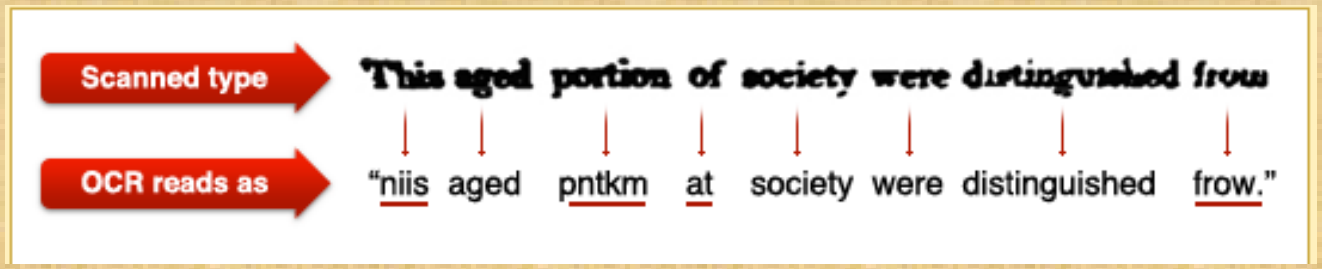
Then how google can read contents from the pdf file?

Answer is OCR(Optical Character Recognition)

Let’s See How Google uses OCR

Google Drive uses it,They says,

In Google Drive, we take your uploaded images or PDF files, scan the file, and use computer algorithms to convert the file into a Google document.



History

- About 1914 Edmund Fournier d'Albe developed the Optophone, a handheld scanner that when moved across a printed page, produced tones that corresponded to specific letters or characters.

OCR – How It Works

- OCR is a complex technology that converts images with text into editable formats.
- This technology is widely used in many areas and the most advanced OCR systems can handle almost all types of images, even such complex as scanned magazine pages with images and columns or photos from a mobile phone.

Steps For OCR

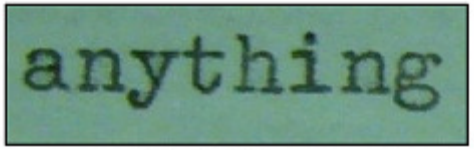
- Loading image as bitmap from given source.
- Detecting the most important image features like resolution and inversion.

  (size , background colors, text colors)
- Image can be skewed or it can have a lot of noise.. (algorithms are applied to improve quality)
- Many OCR algorithms require bi-tonal image,image must be converted to black-white image. This process is called "binarization“ .
- Line Detection
- Page Layout Analysis
- Broken- character analysis.

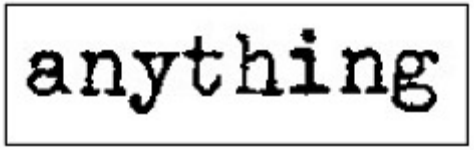
Some Steps..

Original image

Image after Deskew algorithm

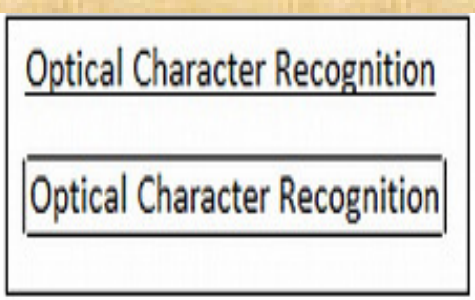


Original image



Correct binarization

Incorrect binarization



Original image

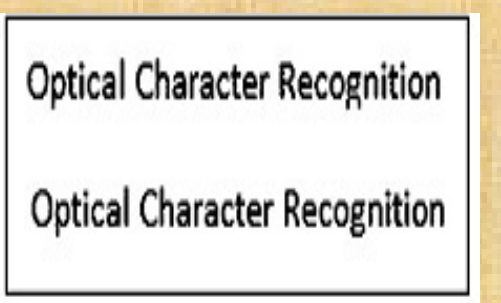


Image with removed lines

OCR softwares

- CuneiForm
- GOCR
- Ocrad
- OCRFeeder
- OCRopus
- tesseract-ocr
- TOCR (paid one – This one I am using)

End

- You can Bypass captcha Using OCR..
- Use It To Make Web Automation Projects..

- Credits:
- My Mom , Dad
- Ashish Mistry