Project Report of Extracting details from a set of images

- It is one of the challenging task with full of joy and logic.
- In this project we are using the ocr(Optical Character Reader) technology to extract the data from the images like text and symbols and in different position in it.
- As part form this, we are also using OpenCV python library with NumPy to perform the image loading, converting in different forms like(BGR_To_GRAYSCALE), sizes etc.
- We are using pandas library to convert all data in structured form and export into csv file.

In the First Step we iterate on each and every images in the folder convert into grayscale, remove unnecessary part and noise from image so that we can create bounding boxes and extract data more accurately and creating bounding boxes on every required area.

In Second Step we divided all the bounding boxes into to category based on their cropped images heights to identify that whether it is from text category or symbols category

In addition of Second Step, there is little constraints that all the symbols are coming in one bounding box and few are on separate. So this problem can be solved by dividing the bounding box that contains multiple symbols into the widths of the bounding box that contains only single symbols. For example, width=220.

In Third Step, we pass the cropped images that contains text to the ocr pytesseract that extract the text from images and return in string form and the cropped images that contains symbol to the similarity_check function that will return the name of that files whose matched mostly.

In addition of Third Step, when we pass the symbol's cropped images to our own defined similarity_check function. In this function the passed image matches with all the images of symbols folder and calculate difference and error. After that it will returns symbols_ID of that symbols whose has least error.

In the Fourth Step or last step, we perform some basic operation to clean our extracted data(text and symbols_IDs) and make a pandas data frame to export into the csv file.

Requirements to run the files

1.OpenCV python Library (pip install cv2)
2.Numpy python Library (pip install numpy)
3. Pytesseract python Library (pip install pytesseract)
4. Tesseract-OCR which is optical character recognition engine (download from https://tesseract-ocr.github.io/tessdoc/Downloads)