**Steps involved in extraction of text from any image using tesseract.**

1. Preprocessing of image-Preprocessing includes converting image to gray scale, sharpening image , smoothening of image, blurring, shrinking, expanding ,warping of image.

**Note :** text extraction basically depends upon the resolution, source, type of image.

1. Adding domain specific words to the dictionary in order to improve the extraction .

Two approaches have been implemented by me in order to fetch the word from dictionary of custom words.

Suppose we have predicted word as Pword and its equivalent dictionary word as Dword.

Let us assume that Pword is the wrongly predicted word.

**Approach 1:** If first 5 alphabets of Pword (i.e wrongly predicted word) matches the Dword then switch the Pword with Dword.

Ex- Pword is Califin and Dword is California,thus we can correct the incorrect word by using dictionary word.

**Approach 2:** If first 2 and last 2 words of Pword (i.e wrongly predicted word) matches the Dword then switch the Pword with Dword.

Ex- Pword is Neiwork and Dword is Network, thus we can correct the incorrect word by using dictionary word.

**Approach 3:** If first 3 and last 3 words of Pword (i.e wrongly predicted word) matches the Dword then switch the Pword with Dword.

Ex- Pword is estobloshed and Dword is established, thus we can correct the incorrect word by using dictionary word.

**Approach 4:** If first 2 and last 6 words of Pword (i.e wrongly predicted word) matches the Dword then switch the Pword with Dword.

Ex- Pword is repondential and Dword is residential, thus we can correct the incorrect word by using dictionary word.

1. Implementing autocorrect (feature using TextBlob library) over the top of the predicted text enhances the accuracy even more.

Ex- Implementing autocorrect improves the minor mistakes for common words ex- is,am,the,are,these and other common mistakes etc.

**Note:** Code file is attached for the same.