## Problem 1: Segmentation of a given Document

For this problem, I used OpenCV and tesseract-OCR to detect and extract text.

1. Then using imread function the image was read.
2. After that, the image is converted to gray scale and threshold is applied to it.
3. Then I dilated the image to help in better text extraction as the text expands after dilation.
4. Then I found all the contours of the image and made a .txt file in which we will be storing our text.
5. Then I looped through all the contours, extracted the text from the image using an OCR, and stored it in the text file.

Problem 2: Clustering of Documents and its Optimization

For this problem, I have used Glove embedding (200D) and converted all the words in the text to a vector form.

1. Then I calculated the mean of every sentence and got a 200-dimension vector of every sentence.
2. Then I calculated the mean of all sentences in an article and hence got a 200-dimension vector of every article.
3. Then I converted this to a dataframe and found the appropriate number of clusters using the inertia method.
4. Using KMeans, I clustered the articles and added the ‘cluster’ column to the dataframe.
5. The code is in the repository.