For my project proposal, I want to do a text summarization for essays with similar topics and combine them into a single essay to summarize all the files in the cluster. Basically, reduce all the meaningless words, such as, stopwords from each essay and take all the meaningful words from each essay and combine them into one essay. I am not trying to add all the sentences together, I am comparing each sentence within the same position and taking the most meaningful one. The analysis that I want to find, is to take all the meaningful text from each essay and see if they can be combine into one essay to summarize the cluster in a way that makes some sort of sense.

To actually do this, I want focus on each paragraph one at a time. In each essay, the format of the essay is usually the same, the beginning paragraph is usually the introduction that gives you the main idea and supporting arguments of the essay. The middle paragraphs are the support arguments. The final paragraph is the conclusion. So, the format of the combine single essay will be the same. Each paragraph will get rid of their stopwords and count how many times each word appears in the text and weighted them. The calculation for the weighing is to find the highest frequency word and divide each word with the highest frequency word. This is how I will find the meaningful text, by adding up all the weighed words in each sentence to get a sum for the sentence. Compare the sentence with another sentence in the same position, in another essay and taking the highest sum sentence. The output will printout the combine essay. It will tell you when one essay ends and another begins.