Ahmed Alhaj
CMSC 476
Phase 5: documents clustering

Document clustering has been searched for use in a number of different areas of text Mining, natural language processing and information retrieval. Document clustering has been investigated heavily for improving the precision or recall in information retrieval systems.

First of all, In this phase are build on top of each other, so with regard to the similarity matrix, I used Token Vector structure from phase 4. So basically each document is converted in weighted Token Vector which is a just vector of word. And then I loop over whole document vector set and compare each document vector with the whole vector set using the cosine similarity. Following the requirement if the cosine similarity between each pair of document vector less than 0.4 they label as not similar each similar. So I wrote the similar documents into a file as pair of file.

The main goal of agglomerative or hierarchical document clustering is to create a hierarchical tree of clusters whose leaf nodes represent the subset of a document collection. Moreover, this method can be further classified into agglomerative and divisive approaches, which work in a bottom-up and top-down fashion, respectively. An agglomerative clustering iteratively merges two most similar clusters until a terminative condition is satisfied. On the other hand, a divisive method starts with one cluster, which consists of all documents, and recursively splits one cluster into smaller sub-clusters until some termination criterion is fulfilled.