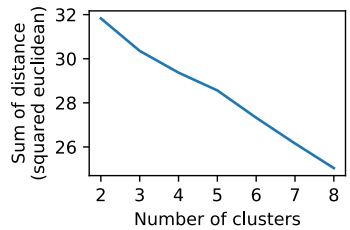
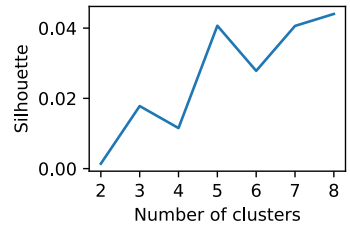
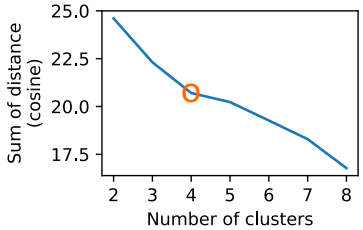
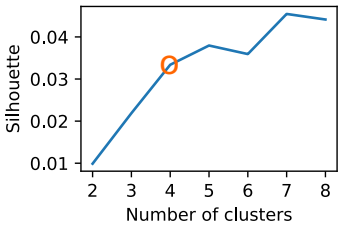
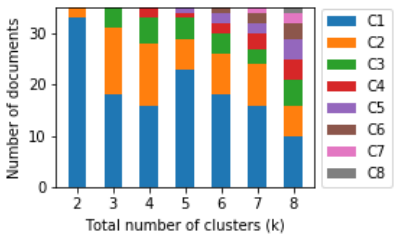
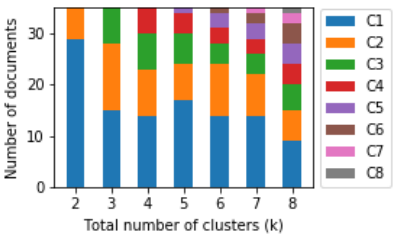
CS5344 Big data analytics technology - Lab 2 Task B

Jesisca Tandi (A0185994E)

Table below shows the resulting clusters from k-means clustering of documents (represented by normalized TF-IDF vectors) using two different distance measures: Euclidean and Cosine distance.

Distance measure	Euclidean	Cosine
Error scores with varying ***k*** (2 to 8)	 	 
Number of documents in each cluster with varying ***k*** (2 to 8)		

Optimal number of clusters

In k-means clustering, the optimal number of clusters, ***k***, is typically determined using the *elbow method*; when the reduction of within-cluster distance is no longer substantial if we increase ***k*** further. In addition, silhouette score gives some estimate of the separation between the resulting clusters; it ranges from -1 to 1 (1 indicates samples are far away from other clusters, i.e. good separation and -1 indicates samples might have been assigned to the wrong cluster).

In this exercise, there is no clear "elbow" in the resulting clusters computed with Euclidean distance. On the other hand, results from k-means clustering with cosine distance suggest that k=4 might be optimal, as shown by the initial decrease of error scores from k=2 to 4, and then it plateaus afterwards. In addition, when k=4, silhouette score is relatively comparable to when ***k*** is increased further (5 to 8), i.e. there is not much gain of increasing ***k*** further.

Comparison between Euclidean and Cosine as a distance metric in k-means clustering

Euclidean distance is highly affected by the absolute value or magnitude of the data; the score is not normalized and hence it is more prone to outlier. Preprocessing steps must be applied prior to clustering

to ensure that input data are in comparable scales. Failure to normalize the data will result in clusters which are not meaningful.

On the other hand, cosine distance is bounded, more stable, and less affected by magnitude of the data since it normalizes the input. It might be able to pick up "true" correlations better as compared to Euclidean.

In this exercise, clustering with Euclidean distance had difficulties finding the optimal boundaries because the data points are almost disconnected from one another (large distance). This is shown by the small number of documents in each new cluster as we increase ***k*** further. For example, when k=5, there are 2 clusters (c4 and c5) with only 1 document as their member.

However, it is worth highlighting that the number of data points (i.e. documents) is very small and the number of features (bags of word) is large. With this small sample size, it might be challenging for clustering methods to determine clusters that are representative.