# Similarity Measures for Text Document Clustering

- Accurate clustering requires a precise definition of the closeness between a pair of objects, in terms of either the pairwised similarity or distance.
- A variety of similarity or distance measures have been proposed and widely applied, such as cosine similarity and the Jaccard correlation coefficient. Meanwhile, similarity is often conceived in terms of dissimilarity or distance as well. Measures such as Euclidean distance and relative entropy have been applied in clustering to calculate the pair-wise distances.
- According to the evaluation by the author, Euclidian distance is the worst performer. On average, the Jaccard and Pearson measures are slightly better in generating more coherent clusters, meaning clusters with higher purity scores. Also, Kullback-Leibler (KL) divergence is a solid metric.
-