Week 9 Deliverables

- Name Manoj Nagaraja
- Email manojn.7270@gmail.com
- Country United Kingdom
- College University of Liverpool
- Specialization Data Science

- Github Repo Link https://github.com/ManojN7270/Final-Project-week_7-to-week_13.git

Problem description

Requires to implement various clustering algorithms using the Python programming language and apply them to cluster a given dataset. The purpose of this project is to assess the understanding of various clustering algorithms by implementing the algorithms and applying them to text clustering.

Algorithm for K-means Clustering :

- Calculating K, the total amount of clusters.
- Choose K data points at random and assign each one to a cluster.
- Continue working through the previous steps to achieve the ideal centroid, which is the centroid's data point that stays the same. .
- The squared distances between the centroids and the data points are added up and calculated.
- The cluster that contains each data point should be chosen based on its proximity to the other clusters. Calculate the centroids for clusters by averaging all of the data points within the cluster.

Algorithm for K-means++ Clustering :

- At first, decide by random which data point's centroid to use.
- Calculate the distance between each data point and the nearest centroid.
- Select the next centroid from the data points with a probability proportional to the square of how far away the previous centroid is from the current centroid.
- Continue with steps 2 and 3 for selecting k centroids.
- The data can be grouped using the k-means method once the centroids have been seeded with k-means++.

Algorithm for Bisecting K-means Clustering :

- Collect all the data points into one cluster.
- Continue working up until the required number of clusters is reached.
- Apply the k-means clustering algorithm with k=2 to the selected cluster
- Determine the SSE (sum of squared errors) for each cluster that is generated.
- Decide which cluster has the highest SSE, then replace it with the most current clusters.