Data Glacier

Your Deep Learning Partner

Week 11 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.

EDA presentation for business users

Introduction to K-means Clustering:

- K-means clustering is an unsupervised learning algorithm used to group data points into clusters.
- It involves choosing k initial centroids, assigning data points to the closest centroid, and iteratively refining the assignments.
- The resulting clusters represent collections of related data points based on their separation from the centroids.

Algorithm for K-means Clustering:

- Steps include calculating the total number of clusters (k), randomly choosing initial centroids, and iteratively updating centroids based on the mean of assigned data points.
- The algorithm converges when centroids stop moving or after a certain number of iterations.

Introduction to K-means++ Clustering:

- K-means++ is an improved version of K-means that enhances centroids initialization.
- It selects initial centroids more intelligently, increasing the likelihood of choosing points distant from existing centroids.

Algorithm for K-means++ Clustering:

- Steps involve randomly selecting the first centroid, sampling subsequent centroids based on the square of their distance from existing centroids, and grouping data using K-means.

- **Introduction to Bisecting K-means Clustering:**
- Bisecting K-means creates a hierarchy of clusters by recursively splitting the largest cluster until the desired number of clusters is obtained.
- It handles non-convex clusters, provides a hierarchical structure, and is less sensitive to initial centroid selection.
- **Algorithm for Bisecting K-means Clustering:**
- Steps include collecting data points into one cluster, iteratively applying K-means to split clusters, and replacing clusters based on SSE (sum of squared errors).
- **Comparison of Silhouette Coefficients:**
- Silhouette coefficient measures how well-separated clusters are and can be used to evaluate clustering performance.
- Comparing the Silhouette coefficients of different methods, we find that k-means consistently outperforms k-means++ and Bisecting k-means for the given dataset.
- **Conclusion:**
- K-means clustering algorithm is the most effective for grouping the dataset based on the Silhouette coefficients.
- Present the Silhouette coefficients obtained for different values of k for k-means, k-means++, and Bisecting k-means.
- Emphasize the importance of proper centroid initialization and the impact on clustering accuracy.

Recommended models for the data set

1. K-means Clustering: The K-means clustering algorithm can be applied to the dataset, as mentioned in the discussion. It is a widely used unsupervised learning algorithm that organizes data points into clusters based on their similarities.

2. K-means++ Clustering: The K-means++ clustering algorithm, an improvement over the K-means algorithm, can also be considered. It enhances the initialization of centroids, resulting in potentially better clustering outcomes.

3. Bisecting K-means Clustering: The Bisecting K-means algorithm, which creates clusters hierarchically, can be another option. It starts with a single cluster and recursively splits the largest cluster until the desired number of clusters is obtained. This algorithm can handle non-convex clusters and is less sensitive to initial centroid selection.

Based on the Silhouette coefficients provided, it seems that K-means and K-means++ consistently outperform Bisecting K-means for this particular dataset. Therefore, K-means and K-means++ are the recommended models to consider for grouping the dataset.