Fundamentals of Big Data Analytics

Quiz 1's Solution

List down some of the possible conditions for producing two different dendrograms using an Agglomerative Clustering algorithm with the same dataset.

There are several conditions that can lead to producing two different dendrograms using an Agglomerative Clustering algorithm with the same dataset. Some of these conditions include using different linkage methods (e.g. single linkage, complete linkage), using different distance metrics (e.g. Euclidean distance, Manhattan distance), and using different preprocessing techniques (e.g. normalization, standardization) on the data before clustering.

List down the different types of the data along with examples.

- **Text Data**: blogs, webpages, tweets, documents, emails
- **Multimedia Data**: image, audio, video
- **Time Series Data**: Sensor data, Stock market data, Forex rates, Temporal tracking (GPS), Smart Meters Data (AMI)
- **Sequential Data**: Bio-sequences, Discretized music and audio data, Textual sequences
- **Streams**: Real time algorithms, fire-alarms, sensor based parameter controls in industrial manufacturing
- **Graphs and Homogeneous Networks**: Facebook, web-graph, twitter, co-authorship graphs (bibliometrics), citation networks
- **Graphs and Heterogeneous Networks**: Authors and conferences, transportation

What should be the best choice for number of cluster based on the following results. Please explain the reason as well.

6 cluster because we can see a very slow change in the value of WSS after k=6, so you should take that elbow point value as the final number of clusters.

The major change in the slope of curve also occurs at k=6.

Compare Hierarchical Clustering and k-Means Clustering in terms of flexibility and scalability.

Hierarchical clustering is more flexible than k-means clustering because it does not require the selection of a predefined number of clusters in advance. It can be used to build a cluster hierarchy, which can be visualized as a dendrogram. On the other hand, k-means clustering is more scalable than hierarchical clustering because it can handle large datasets more efficiently.