PLHC-DS REPORT (CS60050)

VISHAL RAVIPATI 20CS10076

The results obtained by the clustering algorithm are subject to random fluctuations because the initial centroids are chosen randomly.
It has been observed that k = 3 is most often the best k chosen by k-means clustering algorithm, followed by k = 4, with k = 5 and k = 6 being seen very rarely, if ever.

The similarity coefficients for k-means too are subject to the randomness of the algorithm and vary significantly, usually from 0.45-0.52 (better than the sklearn k-means model, which has been found to give a score of 0.4-0.47.
The similarity coefficients for top-down clustering are much more stable, and have been found to vary around 0.45-0.46 when best_k = 3, sometimes going to 0.5 when best_k = 4

Note: The below execution times have been reported on a local system (Dell Inspiron 5501, 8GB RAM, Intel Core i7, 10th gen) on a CPU (GPU wasn’t used).

The cell executing k-means model executes the algorithm 4 times and completes in 1.0 seconds, giving an average time of 0.25 seconds for the k-means algorithm (other operations being negligible).

The cell executing hierarchical top-down clustering is much slower, with the cell executing the model once being completed in 1.5 s, which is equal to the model execution time (other operations being negligible).