ABSTRACT

COMPARISON BETWEEN DBSCAN AND ACCELERATED HIERARCHICAL DBSCAN FOR TEXT CLUSTERING UNSTRUCTURED QUESTIONNAIRE DATA*

Nelson Halim
15/380922/PA/16730

Questionnaires are often used for market research. Considering that the questionnaire data is in the form of text data and that large amounts of these data are produced, it becomes difficult to read the responses from questionnaires line per line. Text clustering can aid in getting an overview of the data. This research therefore aims to obtain a suitable clustering algorithm for text data. The text data has to be converted into a structured form using GloVe model, then the two clustering algorithms are implemented - DBSCAN and accelerated HDBSCAN*. Different sample sizes of the data are obtained from the same dataset to get sample sizes of 5,000, 6,000, 7,000, 8,000, 9,000 and 10,000 responses. These samples are clustered using the two different algorithms where evaluation of the clustering result is done by comparing the silhouette coefficient, Calinski-Harabaz index and run time for both the clustering algorithms to see which algorithm performs better. From this research, it is found that DBSCAN results in higher silhouette coefficient and higher Calinski-Harabaz index, showing that DBSCAN produces better clustering. The time taken to cluster for DBSCAN is longer than accelerated HDBSCAN*. DBSCAN is therefore a better clustering algorithm for text data compared to accelerated HDBSCAN*.

Keywords: Text clustering, DBSCAN, HDBSCAN*, GloVe