INFORMATICS INSTITUTE OF TECHNOLOGY

INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

News-Clust (An Automated News Clustering Model)

A dissertation by

Mr. Pasindu Sandaruwan

Supervised by

Prof. Prasad Wimalaratne

Submitted in partial fulfillment of the requirements for the BSc (Hons) Computer Science degree at the University of Westminster.

April 2023

Abstract

Natural Language Processing and Machine Learning are the two main fields of study involved in the rising academic topic of text clustering. Unsupervised machine learning, which includes clustering, is more difficult to implement and evaluate than its supervised counterparts. It is challenging, especially when dealing with a dynamic domain such as online news information, due to the unpredictable nature of cluster labels or cluster count. Without this background information, the necessary knowledge should be learned from the data set itself. Therefore, the goal of this research is to develop a suitable strategy that considers all these issues and clusters online news items more precisely. Due to the longer news item content, the main issue was a lot of noise. To increase the precision and efficiency of the clustering module, attention was focused on reducing this noise and extracting important information. To determine the most effective methodology and appropriate set of parameters for a noisy and lengthy text corpus like ours, text preprocessing and feature extraction strategies were empirically tested under various parameter combinations. The feature matrix's dimensions were decreased by the application of singular valued decomposition, greatly enhancing the output of the clustering module.

The ideal number of clusters was discovered using the K-means Elbow technique and Silhouette Score curves. Using data with ground truth information, the optimized model was evaluated, and the caliber of the created clusters was determined using extrinsic techniques like Jaccard Score and Adjusted Random Index. By automatically pulling sports news from CNN, BBC, and Aljazeera news sites, the data set for the experiment was created.

The created clusters for the calculated optimal number of clusters (k), when evaluated with the whole data set with no ground truth knowledge, had a Silhouette Score over 0.4 and a Sum of Squared Errors below 60. Although most of the clusters showed a meaning in accordance with most of the items assigned to them, some overlapping clusters caused the cluster quality measures to fall short of expected values. When put to the test on a sample data set with non-overlapping clusters, the model produced reliable results. For this sample, SSE was lowered to a value less than 10 and the silhouette score was about 0.9. Extrinsic methods were used to compare the generated clusters for this sample with the data from the real world, and the calculated value for the Adjusted Random Index is a sign of a good outcome which was 0.969.