INFORMATICS INSTITUTE OF TECHNOLOGY

INFORMATICS INSTITUTE OF TECHNOLOGY

In Collaboration with

UNIVERSITY OF WESTMINSTER

Unsupervised Hybrid Approach for Extractive Summarization

A Project Proposal by

Mr. Suvendran Nirahulan

Supervised by

Ms. Dileeka Alwis

Submitted in partial fulfilment of the requirements for the BEng/BSc in Software Engineering degree at the University of Westminster.

Date: February 2023

ABSTRACT

The ever-increasing volume of information on the internet can be overwhelming, making it difficult to keep up with the data deluge. Article summarization offers a viable solution to this problem, involving the condensation of large amounts of text into shorter versions while preserving the essential information. Researchers have explored various article summarization approaches, but most rely on supervised learning algorithms, which require labelled data that is not always available. Therefore, this project proposes an unsupervised hybrid approach to article summarization that combines K-means clustering and Latent Semantic Analysis (LSA) techniques.

The hybrid approach combines K-means clustering, LSA, and feature scoring to generate article summaries. The researchers implemented a hybrid summarization algorithm that first pre-processes the input text by tokenizing, lemmatizing, and removing stop words. The algorithm then applies K-means clustering to group similar sentences together and LSA to extract important topics. Finally, it scores the sentences based on their similarity to the extracted topics, relevance to the article's keywords, named entity recognition (NER) scores, sentence length, numeric scores, term weight scores, and position scores. The top-scoring sentences are then selected to form the summary.

The researchers evaluated the unsupervised hybrid approach by testing ClusterFuseBERT on various articles and comparing the results to existing supervised approaches. They used the Rouge score, a standard metric for evaluating summarization techniques. The experimental results demonstrated that the hybrid approach outperformed existing supervised approaches in terms of Rouge scores. The researchers also noted that the unsupervised approach can be used on any new dataset without requiring labelled data. Overall, the unsupervised hybrid approach proposed in this project shows promising results for generating effective article summaries.