Informatics Institute of Technology

In Collaboration with

University of Westminster, UK

Sinhala Multi Document Similarity Detection Tool

A dissertation by

Achala Yasas Piyarathna

Supervised by

Mr. Guhanathan Poravi

Submitted in partial fulfillment of the requirements for the

BSc (Hons) in Computer Science

Department of Computing

May 2019

Abstract

Word plagiarism simply refers to using someone else's work without attribution whether it's intentional or not. There are number of software tools to detect plagiarism in various domains. Almost all of these tools are available for English language, but similar tools for Sinhala language is not yet available.

Language independency is a crucial factor that affects the accuracy of similarity detection. There are many attempts of developing language dependent similarity detection tools for languages like Hindi, Chinese, Malayalam, Arabic and Persian. Most of these tools outperforms the available language independent commercial plagiarism detection tools as well. Sinhala language being similar to these languages and also being the official language of Sri Lanka along with Tamil, the need of a comprehensive similarity detection tool is present. Due to the complexity of the language itself the available language independent tools produces very poor results on plagiarism.

This research's main objective is to address the need of a similarity detection tool for Sinhala language to detect similarity in multiple documents. A novel algorithm has been developed to detect the similarity among multiple documents. The proposed system mainly consists of two stages as text pre-processing and similarity detection. A prototype of a multi document Sinhala similarity detection tool has been developed and introduced for demonstration. Sinhala language resources used in this project were taken from the Language Technology Research Laboratory of University of Colombo and Natural Language Processing Research group of University of Kelaniya.

Testing and validations have been carried out by collecting random text samples of school students and which were examined by experts. And the prototype's plagiarism calculation of these datasets was cross referenced by the experts and the actual plagiarized content was identified. The developed prototype has been successful in identifying the plagiarized content with a high accuracy.

Keywords: Plagiarism Detection, Natural Language Linguistics, Natural language Processing, text pre-processing