MSc Project Report

INTRINSIC PLAGIARISM DETECTION IN SINHALA LANGUAGE DOCUMENTS

Charith Thiwanka Amarasinghe

2021

A report submitted as part of the requirements for the degree of

MSc Big Data Analytics at Robert Gordon University, Aberdeen, Scotland

Abstract

The research study is conducted on intrinsic plagiarism detection in Sinhala language documents. There are considerably low number of studies done on the plagiarism detection and authorship verification for Sinhala language. This research proposes an anomaly detection-based approach classify text portions based on anomalous behavior when compared to the neighboring context for the featured extracted using word embedding based approach. In the study multiple feature extraction methods and anomaly detection algorithms and supervised algorithm were used to conduct a series of experiments to identify the combination which perform best for the Sinhala languages. Study uses paragraph level features to distinguish segments with anomalistic behavior. Proposed solution was able classify plagiarized content with an accuracy of 85% with a f1-score of 0.40.

Key words: Data mining, Text analytics, Anomaly detection, Plagiarism detection