Abstract No: SE-14 **Systems Engineering**

Detecting plagiarism in multiple Sinhala documents

G. A. U. E. Ganepola and *T. K. Wijayasiriwardhane

Department of Industrial Management,
Faculty of Science, University of Kelaniya, Sri Lanka
**thareen@kln.ac.lk*

Abstract

Availability of unlimited information resources over the Internet and the advancement of the Internet search engines such as Google to locate those resources much easily have contributed to an increase of plagiarism. Though there are a number of software tools available for detecting plagiarism in multiple English documents, no such a tool is yet available for the Sinhala language. This paper presents a novel language dependent approach to detect plagiarism in multiple Sinhala documents. It uses stemming, stop word removal and synonym replacement for text preprocessing and term frequency-inverse document frequency (*tf-idf*) and cosine similarity for similarity comparison. A prototype software tool was developed and interlinked with an operational Sinhala WordNet to demonstrate the viability of the proposed approach. The prototype tool was validated against a sample of Sinhala assignments from secondary school students. The assignments were also examined by an expert to determine whether they had actually been plagiarized. When compared the results of the prototype tool against those of the expert judgment, we found that our proposed approach for plagiarism detection in multiple Sinhala documents performs with an accuracy of over 80%.

Keywords: Plagiarism detection, Sinhala language, Sinhala WordNet

Introduction

The word "plagiarism" originated from the Latin word "*plagiarius*" meaning "kidnapper". Oxford dictionary (Oxford, n.d.) defines plagiarism as "the practice of taking someone else's work or ideas and passing them off as one's own". The plagiarism has always been a major issue particularly in academia. The availability of unlimited information resources over the Internet and the advancement of the Internet search engines such as Google to locate those resources much easily have contributed an increase in plagiarism at an alarming rate. According to a study conducted by the Center for Academic Integrity (ICAI, 2017), more than 80% of college students have admitted engaging in plagiarism at least once. Further, a survey by the Psychological Record shows that 36% of undergraduates have admitted plagiarizing the written materials. Moreover, a poll conducted by US News and World Reports has found that 90% of students believe that cheaters are either never caught or have never been penalized appropriately (CheckforPlagiarism, 2017).

Under these circumstances, plagiarism detection has been an important research area (Lukashenko et al., 2007) for many languages. The plagiarism detection approaches are generally classified as language dependent or independent approaches. Though the language independent plagiarism detection tools support many languages, they usually fail due to their inability to take language specific grammar rules and semantics into consideration. In contrast, as they focus on a specific language, the language dependent tools always perform better in detecting plagiarism.