Abstract No: SO-03 **Software Intensive Systems**

Plagiarism detection educational tool: A student's assessments similarity checker

J. R. K. C. Jayakody

Department of Computing and Information System,
Wayamba University of Sri Lanka, Dandagamuwa, Sri Lanka
kithsirij@wyb.ac.lk

Plagiarism is very common among students in higher education institutes due to many reasons such as lack of knowledge about the subject, poor academic writing skills or difficulty in meeting a given deadline. The most popular method of plagiarism is to use the online web pages or e-books as it is an easy effort to get the contents from internet, change it and to submit as an original work. Hence, there are bunch of online software tools as well as offline tools exists to detect the plagiarism. However, there are less software tools to identify the copied works among students. Therefore, in this research I developed a plagiarism detection tool to identify the plagiarized assignments or tutorial submitted. Individual assignments and tutorials which had been given to software engineering courses of the Department of Computing and Information System of Wayamba University were used as the dataset. Natural language processing algorithms were developed to derive the statistical features from the assignments such as bag of words, most frequent words, number of words, name entities and paragraphs etc. Moreover, Term Frequency and Inverted Document Frequency (TF-IDF) module was developed to generate a similarity index value among assignments. In addition, Latent semantic analysis module was developed with the word dictionary and vector corpus. Features that were generated and extracted from every module were used to identify the clusters of similar assignments. *K*-mean clustering algorithms in rapid minor were used to identify the clusters. Most of the submitted assignments were identified with number of clusters. Once the clustering results were verified with the students, it was evident that fairly good results were the given by the automatic cluster classification.

Keywords: Plagiarism, Natural Language Processing, Term frequency, Inverted Document Frequency