PLAGIARISM DETECTION ON TEXT BASED ASSIGNMENTS USING VSM

M.A.C. Jiffriya[1*], M.A.C. Akmal Jahan[2], and R.G .Ragel[3]
[1]Post Graduate Institute of Science, University of Peradeniya
[2]Department of Mathematical Sciences, South Eastern University of Sri Lanka
[3]Department of Computer Engineering, University of Peradeniya
[*]macjiffriya@gmail.com

Plagiarism is known as illegal use of others' part of work or whole work as one's own in any field such as art, poetry, literature, cinema, research and other creative forms of study. Plagiarism is one of the important issues in academic and research fields and giving severe concern in academic systems. The situation is even worse with the availability of ample resources on the web. This paper focuses on an effective plagiarism detection tool on identifying suitable intra-corpal plagiarism detection for text based assignments by comparing unigram, bigram, trigram of vector space model with cosine similarity measure. Manually evaluated, labeled data set was tested using unigram, bigram and trigram vector. Trigram shows better results with the labeled data. Because cosine similarity measure using trigram technique focuses on giving more weight for terms that do not frequently exist in the dataset. Even though trigram vector consumes comparatively more time, it is more preferable than the other techniques. Therefore, we present our new tool and it could be used as an effective tool to evaluate text based electronic assignments and minimize the plagiarism among students.

Keywords: VSM, Vector space model, Intra-corpal, Trigram, Cosine similarity