Abstract

To make a paper or scientific work can't be denied that the authors need information and references from other people's work. The reference can be obtained through an article, book, paper, or other works that have relevance to the references that required. Over the time, the growth of digital era makes documents represented as digital documents that can easily accessed in the internet, therefore it's makes anyone can do a plagiarism easily. This paper will explained about plagiarism detection using text alignment approach that have 4 process preprocessing, seeding, extension and filtering. The purpose of the text alignment is looking for fragments that reused or be a source of suspicious documents. The dataset that used in this research is from PAN Plagiarims Detection Task that categorized in 5 types : no plagiarism, no obfuscation, random obfuscation, obfuscation translation and summary obfuscation where each type has their own characteristics. In this paper plagiarism detection using sentence similarity with TF-IDF, cosine similarity, dice coefficient and sentence similarity. The final fragment that resulted form that process will evaluate in two ways. First, use case level evaluation that get 0.9768 for f-measure value. Second, use character level evaluation that get 0.7090 for f-measure value. Based on recall and precision value, this research get 11 of 16 best result for recall value and 17 of 19 best result for precision value.

Keywords : plagiarism, text alignment, sentence similarity, cosine similarity, dice similarity, TF-IDF, adaptive parameter.