World Academy of Science, Engineering and Technology
International Journal of Computer and Information Engineering
Vol:9, No:12, 2015

Framework for Detecting External Plagiarism from Monolingual Documents: Use of Shallow NLP and N-Gram Frequency Comparison

Authors : Saugata Bose, Ritambhra Korpal

Abstract : The internet has increased the copy-paste scenarios amongst students as well as amongst researchers leading to different levels of plagiarized documents. For this reason, much of research is focused on for detecting plagiarism automatically. In this paper, an initiative is discussed where Natural Language Processing (NLP) techniques as well as supervised machine learning algorithms have been combined to detect plagiarized texts. Here, the major emphasis is on to construct a framework which detects external plagiarism from monolingual texts successfully. For successfully detecting the plagiarism, n-gram frequency comparison approach has been implemented to construct the model framework. The framework is based on 120 characteristics which have been extracted during pre-processing the documents using NLP approach. Afterwards, filter metrics has been applied to select most relevant characteristics and then supervised classification learning algorithm has been used to classify the documents in four levels of plagiarism. Confusion matrix was built to estimate the false positives and false negatives. Our plagiarism framework achieved a very high the accuracy score.

Keywords : lexical matching, shallow NLP, supervised machine learning algorithm, word n-gram
Conference Title : ICLT 2015 : International Conference on Language and Technology
Conference Location : Istanbul, Türkiye
Conference Dates : December 21-22, 2015

Open Science Index, Computer and Information Engineering Vol:9, No:12, 2015 publications.waset.org/abstracts/39355.pdf