Volume 8 • Issue 4 • October-December 2017

Plagiarism Detection in Marathi Language Using Semantic Analysis

Namrata Mahender C., Department of CS and IT, Dr. B.A.M. University, Aurangabad, India

ABSTRACT

In this article, the authors have proposed a method to detect plagiarism in the Marathi language by using semantic analysis. Nowadays, plagiarism is a challenging task in educational and research fields. Currently, there are some tools available to detect the plagiarism on the basis of similarity of words. But there is no tool available to detect the plagiarism semantically. In this article, the authors have applied preprocessing to a database i.e. tokenization, removed stop words and punctuations, for the goal of calculating the frequency of words. Then searching the same word or synonyms of words in wordnet to detect the semantic plagiarism. It is useful for many researchers who are working in this domain.

KEYWORDS

Plagiarism Detection, Semantic Plagiarism, Tokenization, Word net

1. INTRODUCTION

Plagiarism is defined as "the re-use of someone else's prior ideas, processes, results, or words without explicitly acknowledging the original author and source" (Barrón-Cedeno & Rosso, 2010). There are mainly two methods of plagiarism detection i) Extrinsic or External plagiarism detection and ii) Intrinsic or internal plagiarism detection. Intrinsic plagiarism detection analyses the input document only to find some parts which are not written by the same author without performing comparisons to external corpus. External plagiarism detection needs a reference collection of documents which are assumed to be genuine. A suspicious document is compared to all the documents in this collection to find duplicates or near duplicates fragments in source documents (Mahdavi et al., 2014). Semantic similarity plays an important role in natural language processing, information retrieval, text summarization, text categorization, text clustering and so on. Many semantic similarity measures have been proposed. In general, all the measures can be grouped into four classes: path length-based measures, information content-based measures, feature based measures, and hybrid measures (Meng, Huang & Gu, 2013).

1.1. Types of plagiarism

- **Copy and Paste Plagiarism:** This refers to directly picking up content from a viable source and put it in one's own research paper with or without citing the source appropriately or providing credit to the original author and declaring the work to be one's own (Weber-Wulff, 2010).