Title: Uncovering highly obfuscated plagiarism cases using fuzzy semantic-based similarity model

Author/Authors: Salha Mohammed Alzahrani, Naomie Salim, V. V. Palade

Abstract: Highly obfuscated plagiarism cases contain unseen and obfuscated texts, which pose difficulties when using existing plagiarism detection methods. A fuzzy semantic-based similarity model for uncovering obfuscated plagiarism is presented and compared with five state-of-the-art baselines. Semantic relatedness between words is studied based on the part-of-speech (POS) tags and WordNet-based similarity measures. Fuzzy-based rules are introduced to assess the semantic distance between source and suspicious texts of short lengths, which implement the semantic relatedness between words as a membership function to a fuzzy set. In order to minimize the number of false positives and false negatives, a learning method that combines a permission threshold and a variation threshold is used to decide true plagiarism cases. The proposed model and the baselines are evaluated on 99,033 ground-truth annotated cases extracted from different datasets, including 11,621 (11.7%) handmade paraphrases, 54,815 (55.4%) artificial plagiarism cases, and 32,578 (32.9%) plagiarism-free cases. We conduct extensive experimental verifications, including the study of the effects of different segmentations schemes and parameter settings. Results are assessed using precision, recall, F-measure and granularity on stratified 10-fold cross-validation data. The statistical analysis using paired t-tests shows that the proposed approach is statistically significant in comparison with the baselines, which demonstrates the competence of fuzzy semantic-based model to detect plagiarism cases beyond the literal plagiarism. Additionally, the analysis of variance (ANOVA) statistical test shows the effectiveness of different segmentation schemes used with the proposed approach.