Detecting Plagiarism

Fritz, Caleb

This experiment sought to discover whether a certain set of software techniques, including "string comparison", "bag of words", and "fingerprinting", were more effective than one another. The hypothesis stated that the fingerprinting technique would be the most successful, due to the fact that it is among and if not the most popular plagiarism detection technique available. These techniques were then thoroughly researched and implemented by the researcher, using the computer programming language, Java. After the software was successfully built, a control program was chosen on the basis of accuracy and the disuse of any of the three techniques being tested in order to avoid bias. The control program chosen was the freeware application "WCopyfind", known for its accuracy and usability. Four plagiarism methods, from the most trivial to the most obfuscated, involving verbatim plagiarism, copy and paste plagiarism, shake and paste plagiarism, and paraphrasing plagiarism, were used to provide a percentage of similarity between a reference document and 50 documents implementing each level of plagiarism based on the particular plagiarism detection technique. For accuracy, a base document of the same subject as the reference document was formed before the reference document was actually found in order to create a condition in which document matches may occur due to the sharing of topic, as would be in a real situation. Despite the hypothesis, the bag of words technique was closest to the control program's averaged similarity.