Abstract—Code clone detection remains an active topic in software engineering. It is not only common in the programming process, but also leading to low maintainability.

In this paper, we propose a simple and efficient token-frequency-based approach in Java to detect cloned codes. Our tool extracts variables in each method, using bigram to find out similar variable names, as well as counting keywords and symbols. After transforming these pieces of information into vectors, components are set with different weights by machine learning method, then a similarity algorithm is introduced to find out similar codes.

As a counting-based detection tool, it is small but very efficient: 90% in both precision and recall rate. STCD is a new and effective tool in clone detection.