

Duplicate code occurs in source files for different reasons. In many cases the motivation for copying the code is laziness of a programmer, or an attempt to use an alien source code. Over the years, multiple methods for detection of the duplicate source code have been developed. Approaches vary in the ways they analyze the code, focusing on different representations of the program. Methods based on the analysis of the syntactic properties of the source code often use abstract syntax trees. By examining the tree representation instead of the textual representation of the code, these methods are able to detect duplicate code that underwent formatting changes as well as changes to the names of identifiers. Duplicate code fragments are discovered by identifying the subtrees of the same shape. After the suspicious parts of the tree are identified, further examination of AST properties determines to what extent the code was copied. In this work we develop a system for duplicate code detection based on AST comparison.