Multilayer Perceptron-based Source Code Classification

I. Mohamed*, B.T.G.S. Kumara, and K. Banujan

Department of Computing and Information Systems, Faculty of Applied Sciences,
Sabaragamuwa University of Sri Lanka, Belihuloya, Sri Lanka.

*ifham547@gmail.com

One of the most crucial stages in the software development life cycle is the implementation stage. Source code is the most critical component in a software application. Developers develop new source code from scratch or reuse old program code functionalities according to project's requirements. Instead of developing source code functionalities, most programmers devote considerable time seeking and searching old source files. Therefore, it is critical to have an effective and efficient way for searching source code functions. Topic modeling is one way for extracting topics from source code. Even though statistical modeling techniques have been used to implement several topic modeling approaches, they possess several limitations. Non-formal code components such as method names, identifiers, and comments are used in this regard. The syntax of a language refers to the rules that define its structure. Without syntax, the semantics of a language are nearly impossible to comprehend. Addressing these concerns, the author used a machine-learning algorithm to predict the source code functionality names. The results are solely dependent on the syntax or algorithm of the source code. This study focuses on three Java project functionalities: primary number, Selection sort, and Fibonacci number. The data set was acquired from the Git open-source repository which is an open-source platform supported by developers worldwide. Four hundred and fifty software projects were analyzed, and 23 variables were considered. The source code components are extracted using the Java parser library, creating an abstract syntax tree to extract the source code features precisely. Then an algorithm is developed to get the count matrices of source code features. The data set was then fed into an Artificial Neural Network machine learning model which yielded 95.4% accuracy rate, 95.5% precision, 95.4% recall, and 95.4% F1-score, with a low error rate of 0.033.

Keywords: Artificial Neural Network, Source Code, Java Parser library, Abstract Syntax Tree