Plagiarism Detection in Verilog Hardware Description Language Programs using Machine Learning

Anusha Gupta, Master of Science in Computer Science - Data Science
University of Dublin, Trinity College, 2020

Supervisor - Prof. John Waldron

In this dissertation, we study how we can apply machine learning algorithms to detect plagiarism in Verilog Hardware Description Language programs. It presents a novel technique which can identify similarities between VHDL programs and gives the most similar files among all the files trained. The second part of the research focuses on evaluating the trained machine learning models and identifying which is the most accurate among those in providing the results of plagiarized files. The training as well as testing datasets used in this study consists of the code metrics which are annotated in the form of tokens representing the structure of each VHDL program. Several machine learning algorithms were compared in this research to identify the best performing algorithm, among which Stochastic Gradient Descent classifier came out to be the most accurate with 82.3% accuracy. Lastly, the results of this algorithm were used to detect the similar files and identify plagiarism within them.