VICTOR CHERUIYOT
STUDENT, MASTER OF SCIENCE IN INFORMATION TECHNOLOGY

SOFTWARE TEST CASE SELECTION USING NATURAL LANGUAGE PROCESSING AND MACHINE LEARNING

Software testing is a quality control activity which focuses on detecting defects which are then eventually removed. After completion of coding software products are subjected to system testing with the help of different test cases. These tests are essential and necessary to assess the effectiveness of the software. However, it is impossible to completely test any nontrivial module or system. Theoretically it suffers from halting problem: it is impossible to write a program that tests whether every program halts in a finite amount of time.

Practically, executing all test cases involves enormous time and cost. Hence, it is crucial to test a subset of test cases which are important from user's perspective based on the frequency of usage, criticality, and probability of failure. This research work is limited to test case selection for regression testing which retests software that has been changed or extended by new features during software development.

In this research we carried out the following activities to develop a machine learning based test case selection strategy. First, we clean the data by dropping irrelevant features. Then we preprocess the data by encoding categorical features into its numerical values. After that we compute the bag of words for text data such as test case title, which identify the important functionality for test case selection. All these features are fed into machine learning based classifiers, such as Logistic regression, naãve Bayes, Gradient boosting, kNN, Decision tree, and Gaussian process. Experimental results demonstrate that decision tree classifier outperforms other classifiers for software test case selection.

This research demonstrates that machine learning-based approach can reduce the bias and manual labour of domain expert for software regression testing. In the future, we would investigate the feature selection strategy for natural language processing to optimize the whole software test case selection procedure.

Research Advisor: Dr. Baidya Nath Saha