CS 4100

CMOS Report

3/8/24

Abby Leary and Adeline Fitzwater

1. We identified tokens for the following situations: include (matches #include), type (int, char, or void), comparison (= =, !=, <=, >=, <, or >), operator (+, -, \*, /, %, or &), semicolon, curly brace (opening or closing), parenthesis (opening or closing), comma, return (matches the word return and anything in the parentheses if there is any), loop (if, else, while, for, do), variable, string (in single or double quotes), character (in single quotes), include path (path or library name between <>), and/or (&& or ||), number, assignment (=). When each of these patterns are recognized, the corresponding token is returned to the main program where the token is printed (a 0 is added at the beginning of the token if it’s only one digit so each token is two digits). Single-line comments, multi-line comments, and whitespace are also identified, but the scanner does not return anything, so they are not a factor when creating the fingerprints.
2. To create the winnowing algorithm, we first read in each line from the file of tokens, split them into k-grams of the specified length, hashed the value, and stored it in a vector. Each file being parsed has its own vector of hashes, which are stored in a vector of all hash vectors. Then, we find the fingerprints by iterating through the hashes for each file and finding the minimum hash within the window. The fingerprints for each file are stored in a vector which are stored in a vector of every fingerprint vector. From there, we compare the fingerprints for each file to every other file and saves these as a type that includes the file numbers and the percent match. Then, the comparisons are sorted so the highest similarities are at the beginning and printed to the plagiarism report.
3. The report includes all possible combinations of files and how similar they are. It seems like files 3, 9, 53, and 54 are almost exactly the same. Files 24 and 43 are extremely similar to each other. Files 15 and 23 are also very similar. There are many other file pairs that are similar to each other, but these files have a high chance of being plagiarized and should be investigated.