CS 4100/5100

Project 1

Due: Monday, March 4th at 5:00 pm

In this assignment, we will see how using FLEX for lexical analysis of code can be used to detect potential plagiarism in student assignments. In this assignment we will follow a process similar to the process used by MOSS (Measure Of Software Similarity). That process is as follows.

1. Student programs are tokenized, removing all irrelevant features that are often changed to hide plagiarism, such as variable names, comments, whitespace, etc).
2. Create digital "fingerprints" for each program based on the tokenized code.
3. Compare pairs of programs for number of matching fingerprints. A higher number of matching fingerprints indicates potential plagiarism.

Your program will be CMOS, a MOSS like program that works on C code. The program will take in one command line argument, which is the name of the directory containing student examples.

In order to complete this project you will need to complete:

1. cmos.l : A LEX program that can parse and tokenize one student submission
2. cmos.cpp : A C++ program that can read in one file containing all tokenized submissions (tokens.txt) and performs the fingerprint analysis as described in the Winnowing Algorithm paper.
3. A makefile to compile your project
4. A Bash script (provided) that will read through the directory of examples (directory name taken in as an argument), tokenize each one, prepare the token.txt file for CMOS to analyze, and calls CMOS to perform the analysis

Additionally, you should include a brief project report outlining the following aspects of your project. Your report should be converted to a PDF before submission.

1. What regular expressions/tokens did you identify, and how does your program react to each one.
2. A brief description in your own words of how you implemented the Winnowing algorithm
3. The results of your analysis, including identifying any submissions found by the algorithm that you believe may be plagiarism

You will submit all required code on blackboard. You do not need to include the example files.

Group Work Policy

You may work with one partner on this assignment, but you are not required to do so. We will not need this project for a later stage, so you are not tied to that partner for the rest of the semester. Only one group member should submit, and they should include both partners names in all code file, the project report, and as a comment in the blackboard submission.

Late Policy

Late projects are subject to a 20% penalty for up to 24 hours after the due date. After 24 hours, the assignment will no longer be accepted. You are responsible for submitting the correct file for your project submission and for submitting on time.