Project Proposal

Proposal: To write a program in both C and Haskell in which a user inputs multiple text files and a number $x$, that computes the number of $x$-word sequences between text files and determines whether plagiarism was involved.

My team: I am working individually

Time budget: My project can be flexible in how difficult I want to make it, so my time budget can also be flexible. I could decide to compare only a small number of files, and if I find that to be too easy, I could make it more difficult by increasing the number of files I want to compare with each other. I believe that I can finish this project in five weeks by working about 5-8 hours each week, which means a total of 25-40 hours overall.

Time-line:
1st week: Read in text files and put all the different word sequences of n-words (the user specifies what n is) into some data structure.
2nd-3rd week: Find an efficient way to compare all of the files with each other and record all of the common occurrences between each of the files.
4th week: Come up with a standard for what determines plagiarism. How many similar sequences between two files is it safe to say that one was copied from the other? Output which files meet this standard.
5th week: Work on the class presentation and the written report.

Risks: The biggest risk is that I could aim too high by trying to do too much. On top of having to write the program in two different languages, one of them is Haskell which I am not too familiar with, and probably the hardest programming language of the languages I've learned. I will probably try to mitigate this risk by writing the program in C first, then in Haskell. Hopefully by the time I start writing in Haskell, I will have enough knowledge to write simple, and good code. Like I mentioned before though, I can be flexible in how difficult I make the project once I start and get a good idea of how much time it would take, so I do not think the risks are that great.