Plagiarism Checker

Formally, given a set of text documents , try to determine whether or not, or the degree to which, any pair of documents contains heavily copied portions of text. If one document shares an inordinate number of phrases in common with another, it's likely the document is plagiarized (or is the result of the all-too-often occurrence of finding out that one student "helped" another by giving him a copy of his work, only to have the other turn it in as his own).

You have to define your own measure of similarity and explain it in the report. Also you have to describe the time and space complexity required by your approach.

You will be graded based on how you choose the measure of similarity and working time complexity of your program.

Note : You are only expected to check text similarity, and not the program control flow similarity.

Input :

A) Set of documents [upto 25] against which you have to check the plagiarism. Size of each document will vary.(in order of few thousands of words) [CORPUS FOLDER]
B) A document to be tested for plagiarism

You will be required to write a single make file.
Command to run the executable : ./plagChecker <SPACE> <LOCATION_OF_TEST_FILE> <LOCATION_OF_CORPUS_FOLDER>

Output format:

For each text document in CORPUS FOLDER, print a line as follows:
<TEXT_DOCUMENT_NAME> <SPACE> <SIMILARITY_PERCENTAGE>

Implementation : Any implementation in C programming language but you can not implement a neural network.

This is the initial problem description, minor changes might occur for implementation purposes.