Project - Plagiarism Detector

Problem Statement

The aim of this project is to design, develop, test and implement a plagiarism detector which would show a similarity score (percentage) between any two uploaded files. The goal is to calculate the similarities in the input files so that instructors can assess whether two or more assignments are copied from one another and take the required actions based on it.

Proposed Solution

We aim to build an interactive tool for the plagiarism detector. The project will be implemented using HTML, CSS and Java. The primary actors of this application would be the instructors, with secondary actors being teaching assistants.
We will be implementing the detector primarily for Java programming language and additionally for HTML/CSS (tentative). The project will be carried out in 3 phases.

Below is an overview of the features we plan to implement -
- The actors would have to login with valid credentials to access the application features.
- On logging in, they will be presented with the following options to choose from-
  a) View results of previous runs.
  b) Upload two files from file system to compare
  c) Upload a file and select one existing file saved in system to compare
  d) Run the plagiarism detector against the two files
  e) View similarities in code due to which plagiarism was detected
  f) Download the report for plagiarism score
  g) Save uploaded documents for later use
- The actors can choose any one of the options mentioned above and run the application to get the results.

This tool/application can help them decide whether there are similarities between any two submitted files and if there are potential cases of plagiarism.

Future scope:

Every project has areas it can improve upon and since we plan to adopt an agile process while developing the application requirements are subject to changes and new enhancements can be added. The following is the future scope for our project:
- Ability to run the detector against a web URL, so as to check for code snippets taken from the web without proper citations.
- Ability to run a detection algorithm by uploading multiple of files at once.