Design and Analysis of Algorithms Plagiarism Checker

Harasees Singh Gill: 100656810

December 1st 2019

Design and Analysis of Algorithms: 43513

For the scope of this project, I chose to write an algorithm using Python3. I did not use any external libraries to do my plagiarism checker. I also did not look online for any information other than syntax and formatting issues as well as available built in functions. The provided algorithm is definitely not the most effective way to do this project as I did not research any algorithms prior.

This project was very abstract because a plagiarism checker is a very complicated algorithm with a lot of mathematics behind it and it is not within the scope of a third year software engineering student’s expertise so I understood that the purpose of this project was to push our ability to think more than to create a perfect plagiarism checker. Due to this understanding, I created this algorithm from scratch without the help of frameworks or reusing code fragments.

This algorithm works by taking in two arguments *text\_original,* and *text\_created*. Text\_original acts as files a.txt to e.txt and text\_created acts as files 1.txt to 100.txt. Each file is then broken down into a list of lines, which is then further broken into a list of words and finally iterate through both lists from index 0 to the length of the lesser list.

In order to deal with commas, periods, and any other delimiters I split the line list at every space and then stored each word without its [-1] (last) index. Every list from the original text is compared against every list from the created text and every time multiple strings are found within an incremental series of indexes a counter that has been initialized as 0 at the beginning of the comparison function is incremented. At the end of each line comparison, the final counter is divided by the length of the list with the smaller length and that result is returned. Once every line of the original file has been compared with every line of the created file, a list of results is averaged and the final number that is the result of that average becomes the level of plagiarism or more accurately the degree of difference from 0 to 1 of the two text files.

This algorithm is run 100 times for each of the 5 original files and each row of results is appended to the output.csv file contained in the parent folder of this file.