COURSE: COMP 4411

PROJECT 2 – DETECTING SIMILARITIES WITHIN STRINGS IN CLOJURE

NAME: DHAVAL THANKI

STUDENT NUMBER: 0871347

Objective

The objective of this project is to compare strings to see the similiarity between different strings. This has a wide variety of applications, including plagiarism checkers etc.

Note: most of my results are displayed on Jupyter labs in the attached PDF file. I used the clojupyter kernel to run my code on the notebook because it was easier to use as a incremental way to test and debug issues.

Algorithm 1:

Jaccard index

The Jaccard index will quantify the number of words that are similar between the two documents, it is done by taking the number that fall into the intersection of the two inputs and divided it by the total number of words present, it’s the fastest way to find similarity within textual documents.

Algorithm 2:

Hamming distance

This is the measure of the minimum substitutions required in order to match the two given inputs exactly, this tells us how close the two strings are to one another.

Results:

The results comprise of a combination of the sample data, as well as a larger data set procured from the New York Times. The results of my test can be viewed via the PDF file attached here.