Gebze Technical University
Department of Computer Engineering
CSE 654 / 484
Fall 2022

Homework 01
Due date: Nov 7th 2022

In this homework we will use edit distance to find similar text sections between documents. Here are the steps of the homework

1. Download the standard textbooks from the Ministry of Education (http://aok.meb.gov.tr/kitap/) for at least 20 textbooks (literature, history, sociology, etc.) Convert them to text documents. Each document should be at most 400 lines. You may truncate the text if it is longer. You should have at least 100 documents of Turkish text. More is better.
2. Insert the same line of text into random postions of random text documents so that some of them have common text between them.
3. Using Smith-Waterman Algorithm, find the common lines between given two text.
4. Define the cost of substitution, deletion, insertion yourself. Write your definitions in your report.

Prepare your report and submit it to the Teams page along with your Jupyter notebook. Ypur report should include the dynamic programming matrix results for some small text matching examples. It should also show the results of text matching examples.