# CS 581 Spring 2024 Written Assignment #02

Due: **Sunday, February 11, 2024, 11:59 PM CST**
Points: **30**

## Instructions:

1. Use this document template to report your answers. Name the complete document as follows:

   LastName_FirstName_CS581_WA02.doc or pdf

   **ONLY PDF or MS Word file formats will be accepted.**

2. Submit the final document to Blackboard Assignments section before the due date. No late submissions will be accepted.

## Objectives:

1. (10 points) Demonstrate your understanding of Minimum Edit Distance algorithm.
2. (10 points) Demonstrate your understanding of the N-gram language modeling.
3. (10 points) Demonstrate your understanding of an HMM POS tagger.

## Problem 1 [10 pts]:

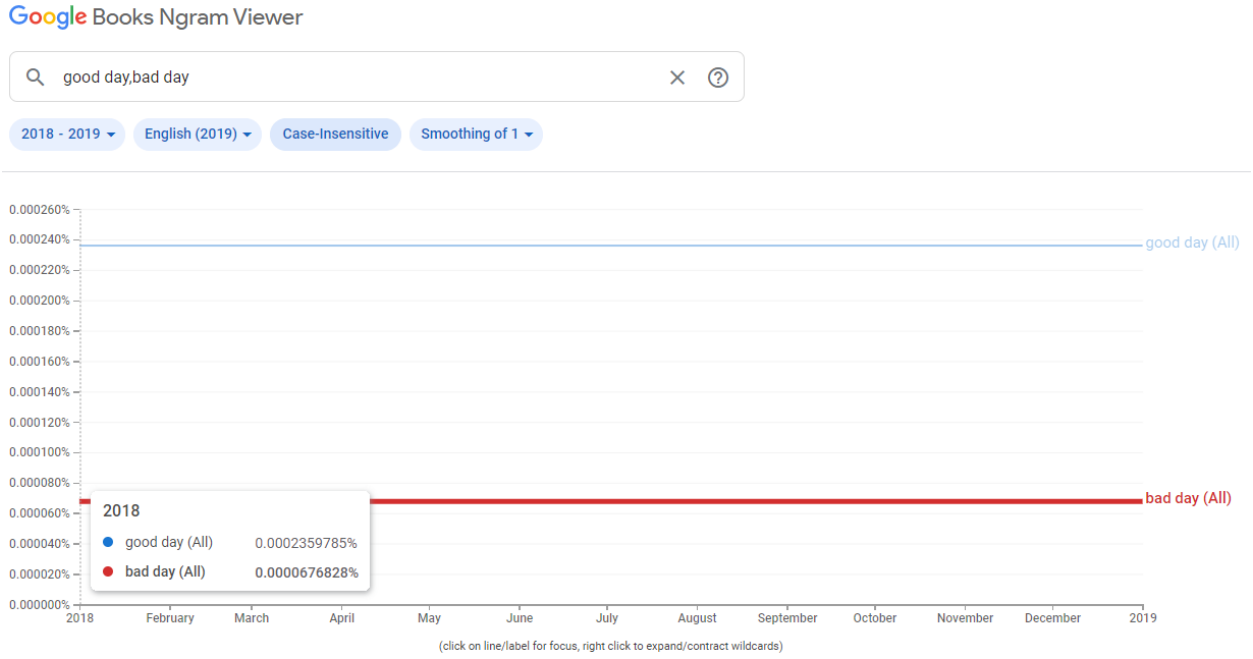
What is the **Minimum Edit Distance** between words STALK and FABLE (assume that insertion / deletion cost is 1, substitution cost is 2)? Populate the table below to find the MED. Include back pointers.

| | | | | | | |
|---|---|---|---|---|---|---|
| **K** | | | | | | |
| **L** | | | | | | |
| **A** | | | | | | |
| **T** | | | | | | |
| **S** | | | | | | |
| # | | | | | | |
| | # | **F** | **A** | **B** | **L** | **E** |

## Problem 2 [10 pts]:

Your task is to calculate probabilities of selected sentences in English using a language model (based on Google Books N-gram corpus). Use the Google N-Gram Viewer website https://books.google.com/ngrams to collect all necessary data (NOTE: Google provides N-gram PERCENTAGES - those are NOT COUNTS! and not exactly probabilities!) and calculate sentence probability.

Figure 1: Notes:
- assume that probability of any bigram starting or ending a sentence is 0.25.
- use the settings shown below (2018 probabilities, English (2019), case insensitive, Smoothing of 1)




A) Probability of a sentence:

*Today is a good day*

| **Relevant bigram probabilities [1 pt]:** |
|---|
| |
| **Probability of a sentence formula [2 pt]:** |
| |
| **Probability of a sentence (calculations and value) [2 pt]:** |
| |

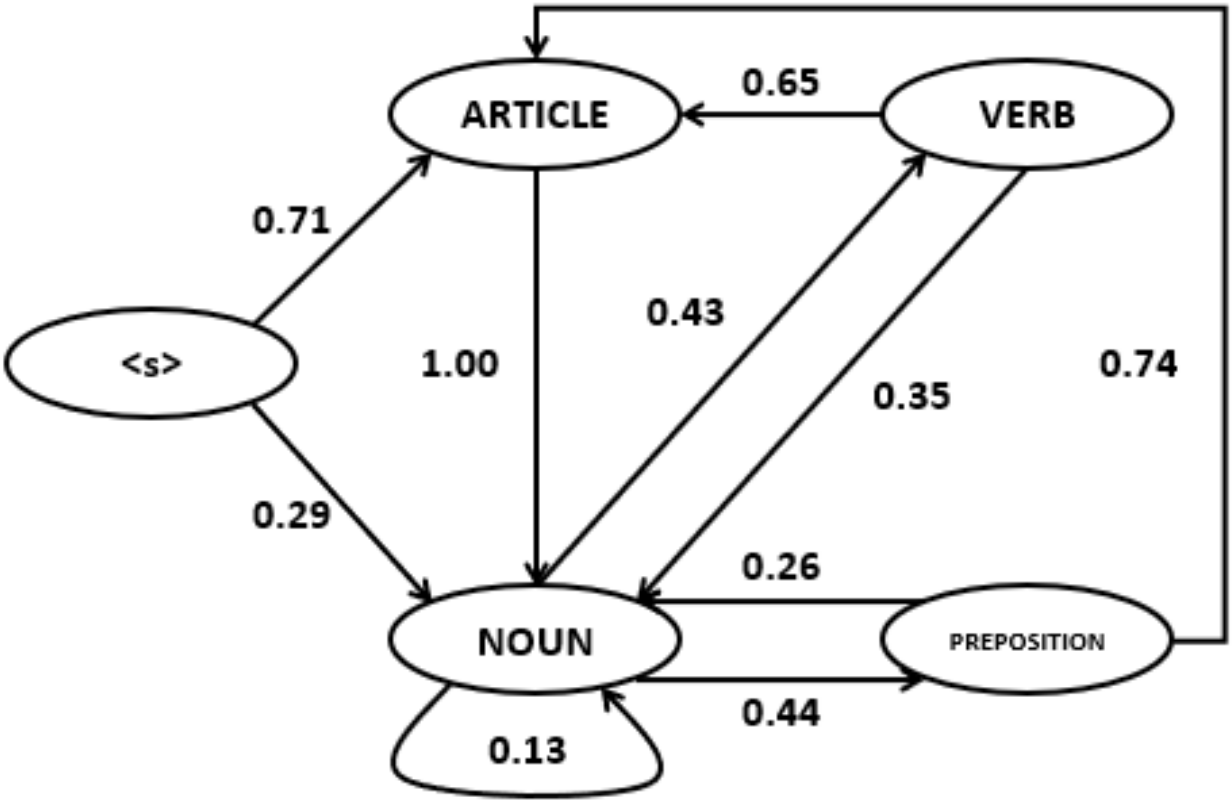
B) Probability of a sentence:

*Today is a bad day*

| **Relevant bigram probabilities [1 pt]:** |
|---|
| |
| **Probability of a sentence formula [2 pt]:** |
| |
| **Probability of a sentence (calculations and value) [2 pt]:** |
| |

# Problem 3 [10 pts]:

Given the following Hidden Markov model (transition probabilities shown; emission probabilities to be determined by you using corpus $C$ data) based on corpus $C$:




And the following table of selected word counts from some corpus $C$:

| Word/Tag | N | V | Art | P | Total |
|---|---|---|---|---|---|