Student Name

CS 585 Spring 2024 Written Assignment #02

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

Points: **30**

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]

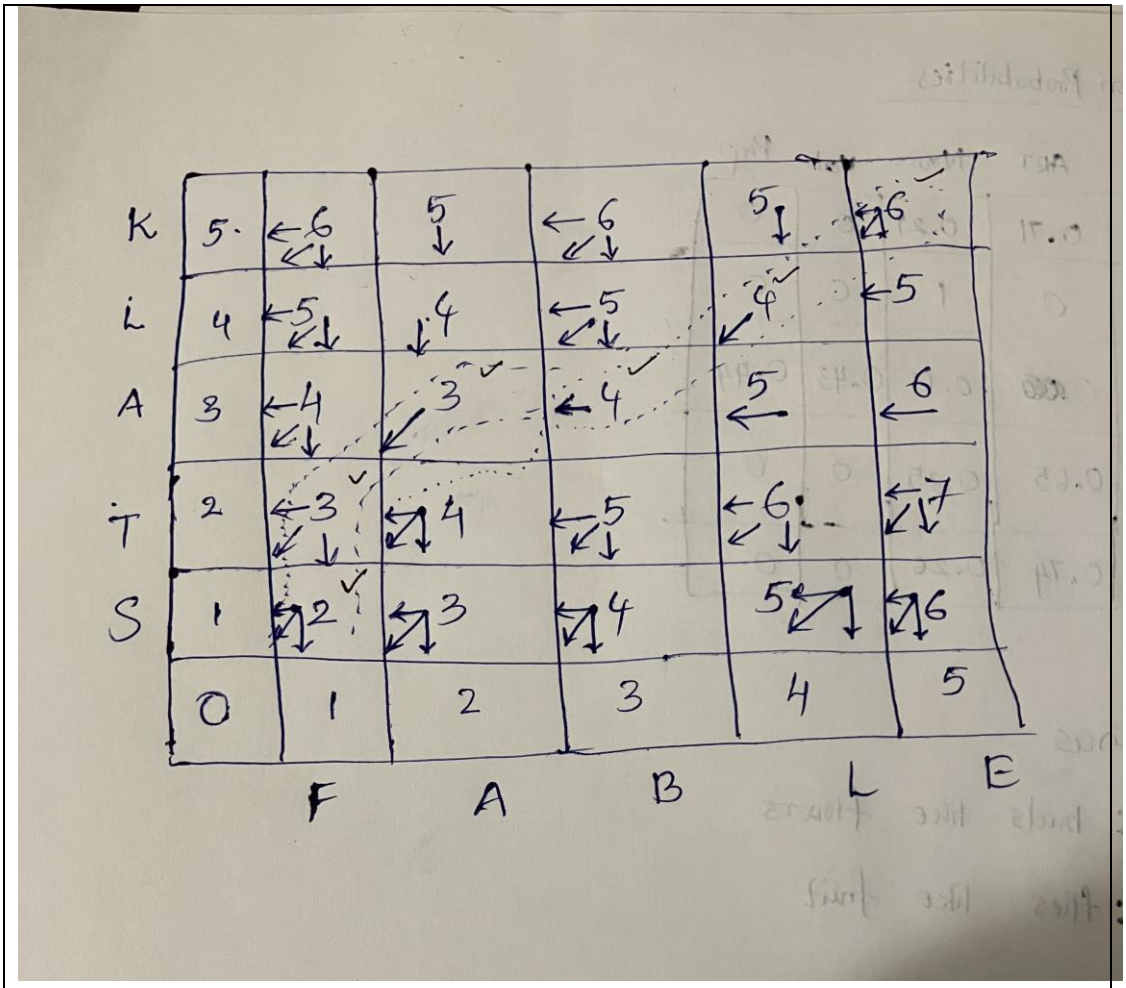
What is the **Minimum Edit Distance** between words STALK and FABLE (assume that insertion / deletion cost is 1, substitution cost is 2)? Populate the table below to find the MED. Include back pointers.

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

Solution:

K	**5**	**6**	**5**	**6**	**5**	**6**
L	**4**	**5**	**4**	**5**	**4**	**5**
A	**3**	**4**	**3**	**4**	**5**	**6**
T	**2**	**3**	**4**	**5**	**6**	**7**
S	**1**	**2**	**3**	**4**	**5**	**6**
#	**0**	**1**	**2**	**3**	**4**	**5**
	#	**F**	**A**	**B**	**L**	**E**

Minimum Edit Distance:



Minimum Edit Distance = 6

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.

NOTES:

- assume that probability of a 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) **[5 pts]** Probability of a sentence:

Today is a good day

Relevant bigram probabilities [1 pt]:
(<s>, today) – 0.25 (Today, is) – 0.00039373768/100 = 0.399376*10^-5 (is, a) – 0.0562696642/100 = 0.005626 = 0.562696642*10^-5 (a, good) – 0.013680092/100 = 13.6860092*10^-5 (Good , day) – 0.0002368249/100 = 0.2368249*10^-5 (day, </s>) – 0.25
Probability of a sentence <u>formula</u> [2 pt]:
P(<s>, Today) * P(Today, is) * P(is, a) * P(a, good) * P(Good, day) * P(day, </s>)
Probability of a sentence (calculations and value) [2 pt]:
(250 * 0.3993768 * 56.2696642 * 13.6860092 * 0.2368249 * 250) *10^-5 = **4.552*10^-5**

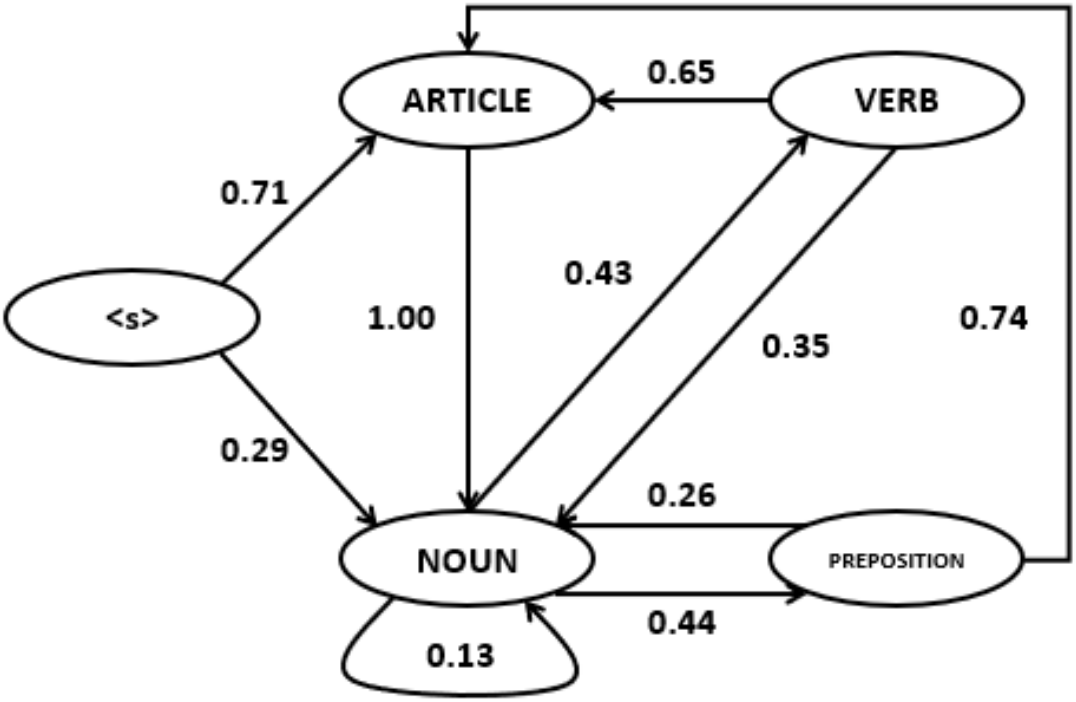
B) **[5 pts]** Probability of a sentence:

Today is a bad day

Relevant bigram probabilities [1 pt]:
(<s>, Today) = 0.25 (today, is) = 0.3993768*10^-5 (is, a) = 56.2696642*10^-5 (a, bad) = 1.963926 *10^-5 (bad, day) = 0.0680989 *10^-5 (day, </s>) = 0.25
Probability of a sentence formula [2 pt]:
P(<s>, Today) * P(Today, is) * P(is, a) * P(a, bad) * P(bad, day) * P(day, </s>)
Probability of a sentence (calculations and value) [2 pt]:
(250 * 0.3993768 * 56.2696642*1.963926*0.0680989 *250) *10^-5 = **0.188*10^-5**

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
flies	21	23	0	0	44
fruit	49	5	1	0	55
like	10	30	0	21	61
a	1	0	201	0	202
the	1	0	300	2	303
flower	53	15	0	0	68
flowers	42	16	0	0	58
birds	64	1	0	0	65
others	592	210	56	284	1142
TOTAL	833	300	558	307	1998

Using the approach presented during the lecture, decide which sentence, S1:

birds like flower

or S2:

flies like fruit

is **more likely to be labeled with a sequence of tags N, V, N**. Show all your work.

Answer:
Second word is most likely to be labelled with sequence of tags because of the probability of 2nd sentence is greater than probability of first sentence.