National University of Computer and Emerging Sciences, Lahore Campus

	Course Name:	NLP	Course Code:	CS4063
	Degree Program:	BS-CS	Semester:	Spring 2023
	Exam Duration:	60 Minutes	Total Marks:	82
	Paper Date:	28-02-2023	Weight	
	Sections:	ALL	No of Page(s):	
	Exam Type:	Midterm I		

Student : Name:_____ **Roll No.**_____ **Section:**_____

Instruction/Notes: Attempt all questions. Programmable calculators are not allowed.

Q1. You are given the following training corpus. (1+2+2+5+2) 12

<s> i want to eat thai food </s>
<s> we ate pakistani food </s>
<s> i ate apples </s>
<s> they ate thai food </s>

a) Calculate the probability of the following test sentence. Include </s> in your counts just like any other token.
<s> i ate thai food </s>
   i) Unigram Model
   ii) Bigram Model
   iii) Trigram Model
   iv) Bigram model with linear interpolation $\lambda_1 = 0.7$, $\lambda_2 = 0.3$

b) Calculate the perplexity of the test sentence using bigram model.

<s> I ate Chinese food </s>

i. Unigram Model

$= P(I) * P(ate) * P(Chinese) * P(food) * P(</s>)$

$= \frac{2}{21} * \frac{3}{21} * \frac{2}{21} * \frac{3}{21} * \frac{4}{21} = 3.52$ (4)

ii. Bigram Model

$= P(I|<s>) * P(ate|I) * P(Chinese|ate) * P(food|Chinese) * P(</s>|food)$

$= \frac{2}{4} * \frac{1}{2} * \frac{1}{3} * \frac{2}{2} * \frac{3}{3} = \frac{1}{12} = 0.83$ (2)

iii. Trigram Model

$= P(I|<s><s>) * P(ate|<s>I) * P(Chinese|I\ ate) * P(food|ate\ Chinese) * P(</s>|Chinese\ food)$

$= \frac{2}{4} * \frac{1}{2} * \frac{0}{1} * \frac{1}{1} * \frac{2}{2} = 0$ (2)

iv. Bigram language model with linear interpolation ... 0.7 0.3

$\lambda_1 = 0.7, \lambda_2 = 0.7$

$= (\lambda_1 P(Bigram) + \lambda_2 P(Unigram))$

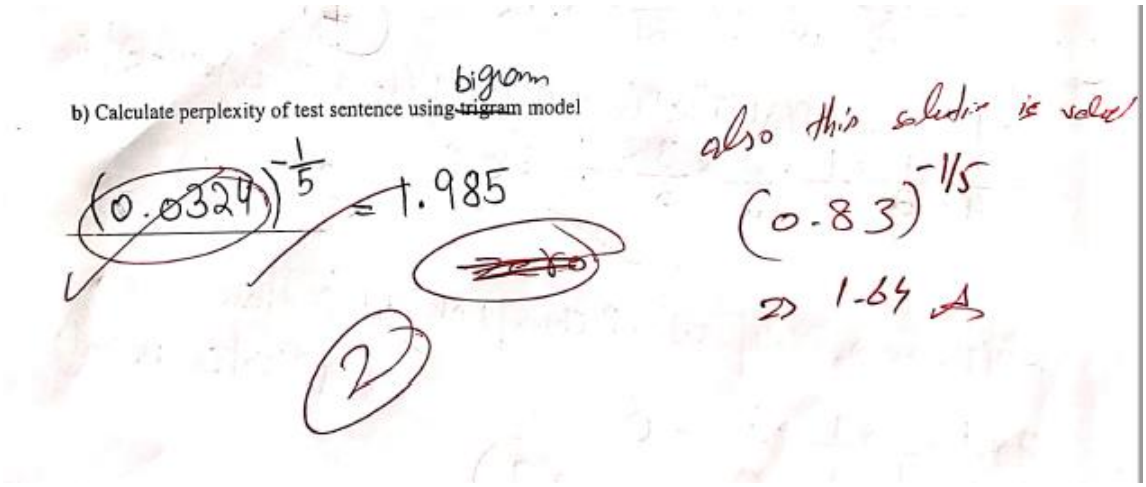
$= \left[0.7\left(\frac{2}{4}\right) + (0.3)\left(\frac{2}{21}\right)\right] * \left[(0.7)\left(\frac{1}{2}\right) + (0.3)\left(\frac{3}{21}\right)\right] * \left[(0.7)\left(\frac{1}{3}\right) + (0.3)\left(\frac{2}{21}\right)\right]$

$\left[(0.7)\left(\frac{2}{2}\right) + (0.3)\left(\frac{3}{21}\right)\right] + \left[(0.7)\left(\frac{3}{3}\right) + (0.3)\left(\frac{4}{21}\right)\right]$

$= (0.35 + 0.028)(0.35 + 0.042)(0.23 + 0.028)(0.7 + 0.4$

$(0.7 + 0.057)$

$= 0.0324$ (3)



Q3: You are given two documents:
Doc 1 = the car is in the parking and the bike is in the garage (13)
Doc 2 = the truck is driven on the highway and the tractor is in the farm parking (15)

a) Compute the normalized term frequency and un-smoothed logarithmic inverse document frequency for the given corpus. You can use log of base 10 for the calculation of IDF. Fill the table based on your calculations: (15+10)
IDF: log(N/n)

Term	Count (doc1)	Count (doc2)	TF (doc1)	TF (doc2)	IDF	TF*IDF (doc1)	TF*IDF (doc2)

Term	Count (doc 1)	Count (doc 2)	Tf (doc 1)	Tf (doc 2)	IDF	TF*IDF (doc 1)	TF*IDF (doc 2)
the	4	4	0.30769231	0.26666667	0	0	0
car	1	0	0.07692308	0	0.301	0.02315385	0
is	2	2	0.15384615	0.13333333	0	0	0
in	2	1	0.15384615	0.06666667	0	0	0
parking	1	1	0.07692308	0.06666667	0	0	0
and	1	1	0.07692308	0.06666667	0	0	0
bike	1	0	0.07692308	0	0.301	0.02315385	0
garage	1	0	0.07692308	0	0.301	0.02315385	0
truck	0	1	0	0.06666667	0.301	0	0.02006667
driven	0	1	0	0.06666667	0.301	0	0.02006667
on	0	1	0	0.06666667	0.301	0	0.02006667
highway	0	1	0	0.06666667	0.301	0	0.02006667
tractor	0	1	0	0.06666667	0.301	0	0.02006667
farm	0	1	0	0.06666667	0.301	0	0.02006667

b) From TF*IDF vectors calculated in part a, compute Euclidean Distance and Cosine Similarity (write the formulae too).

Euclidean Distance:

0.0634

Cosine Similarity:

0

Q4: Answer the following. (10)

I. Which of the following is a type of Minkowski distance? _____
   (a. **Hamming**, b. Levenshtein, c. Jaro)

II. Damerau-Levenshtein allows __4__ edit operations, while Hamming allows __1__ operations. (0, 1, 2, 3, 4, 5)

III. Root of the word "antidisestablishmentarianism" is establish and the lemma is disestablishmentarianism

IV. What are the derivational morphemes in each of these words "entitled", "replays", "cities", "undid", "realism", and "higher"? **(en-, re-, none, un-, -ism, none)**

V. Mention all of the following expressions that the regex /bla+..?!/ will select?
blaa! blat! black! bla?! bla.?! bla+?!

a. bla! **b. blaa! c. blat! d. black! e. bla**?! f. bla+..?! g. bla..?! **h. bla.?! i. bla**+?! j. bla+.?!

Q5. Find the Levenshtein distance between PAYMENTS and APARTMENTS. Use the same algorithm and weights as discussed in the class (i.e. cost(Insertion)=1, cost(Deletion)=1, cost(Substitution)=2). (20)

	#	A	P	A	R	T	M	E	N	T	S
#	0	1	2	3	4	5	6	7	8	9	10
P	1	2	1	2	3	4	5	6	7	8	9
A	2	1	2	1	2	3	4	5	6	7	8
Y	3	2	3	2	3	4	5	6	7	8	9
M	4	3	4	3	4	5	4	5	6	7	8
E	5	4	5	4	5	6	5	4	5	6	7
N	6	5	6	5	6	7	6	5	4	5	6
T	7	6	7	6	7	6	7	6	5	4	5
S	8	7	8	7	8	7	8	7	6	5	4

Minimum Edit Distance: _____ 4

- Show the optimal alignment between the sequences and one possible minimal edit sequence (a sequence of inserts *I*, deletes *D* and substitutions *S*) that would result in an optimal conversion from PAYMENTS to APARTMENTS.

```
     P   A     Y   M   E   N   T   S
     |   |     |   |   |   |   |   |
A    P   A  R  T   M   E   N   T   S

I    S   S  I  S   S   S   S   S   S
```

Name: _____ Reg #: _____ Section: _____

National University of Computer and Emerging Sciences, Lahore Campus

Course:	**Natural Language Processing**	**Course Code:**	**CS 535**
Program:	**BS(Computer Science)**	**Semester:**	**Spring 2018**
Duration:	**60 Minutes**	**Total Marks:**	**20**
Paper Date:	**26-Feb-18**	**Weight**	**13%**
Section:	**ALL**	**Page(s):**	**4**
Exam:	**Sessional 1 Solution**		

Q1) You are given the following corpus: [5 + 5 = 10 Marks]

<s> I am Sam </s>
<s> Sam I am </s>
<s> I am Sam </s>
<s> I do not like green eggs and Sam </s>

a) Calculate the probability of following test sentence using trigram language model with linear interpolation. Include <s> and </s> in your counts just like any other token.
λ_1 = trigram weight, λ_2 = bigram weight, λ_3 = unigram weight,
λ_1 = 0.5, λ_2 = 0.3, λ_3 = 0.2

<s> I like green eggs </s>

P (like | <s> I) = (0.5) (0) + 0.3(0) + 0.2 (1/25) = 1/125
P(green | I like) = 77/250
P(eggs | like green) = 101 / 125
P(</s> | green eggs) = 4/ 125

P (<s> I like green eggs </s>) = (1/125) (77/250) (101 / 125) (4 / 125) = $6.37 * 10^{-7}$

Name: _____ Reg #: _____ Section: _____

b) Calculate the probability of P(Sam | am) using Kneser Ney smoothing from the corpus given above. d = discounting factor = 0.5

P(Sam | am) = (2 – 0.5) / 3 + 1/3(2/14) = 0.55

Q3) a) Let S = { a , b, c } the sample space. Suppose you are given following bigram probabilities P(b | a) = 0.125, P(a | c) = 0.25, P(c | c) = 0.25, P(a | a) = 0.25, P(c | b) = 0.125, P(a | b) = 0.25, P(<s> | c) = 0.1, P(<\s> | c) = 0.1, P(b | <s>) = 0.1, P(<\s> | a) = 0.1

Can you compute P(b | c) from the information given. If yes what is P(b | c) ? [2 + 3 = 5 Marks]

Yes, P (b | c) = 1 – (0.25 + 0.25 + 0.1 + 0.1) = 1 – 0.7 = 0.3

b) What is perplexity of bigram distribution from part (a) if computed against following data

<s> b c a <\s>

= ¼ (log (0.1) + log (0.125) + log (0.25) + log (0.1)) = -2.91
Perplexity = $2^{-l} = 2^{2.91} = 7.52$

Name: _____ Reg #: _____ Section: _____

Q4) Detect most probable real word spelling error from following sentence. Use Noisy Channel Model with bigram probabilities as given in Table 1. Probabilities of 1 character spelling mistakes are also given in Table 2. Assume all words with 1 edit distance from the words in test sentence are given in Table 1. α = Probability of word being correct = 0.95. **Show all calculations** [5 Marks]

<s> Three off the </s>

	Three	**off**	**the**	**there**	**tree**	**of**	**then**	**</s>**	**<s>**
Three	0.0001	0.0002	0.003	0.003	0.003	0.21	0.003	0.003	0.003
off	0.003	0.003	0.0003	0.003	0.003	0.003	0.003	0.003	0.003
the	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.01	0.01
there	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003	0.003
tree	0.003	0.003	0.003	0.003	0.004	0.003	0.003	0.001	0.003
of	0.003	0.003	0.22	0.003	0.003	0.003	0.003	0.003	0.003
then	0.003	0.003	0.003	0.003	0.075	0.003	0.003	0.003	0.003
<s>	0.004	0.003	0.003	0.01	0.003	0.003	0.003	0.003	0.003
</s>	0.003	0.003	0.003	0.003	0.04	0.003	0.003	0.003	0.003

Table 1: Bigram Probabilities. P(of | Three) = 0.21

x | w	**P (x | w)**
re | er	0.03
Th | T	0.02
ff | f	0.05
e | ew	0.05

Table 2: x is spell error and w is correct

Solution:

P (<s> Tree off the </s>) = $5.4 * 10^{-13}$

P (<s> Three off the </s>) = $2.7 * 10^{-2}$
P (<s> Three of the </s>) = $9.24 * 10^{-8}$

The probability of “Three of the” is highest so spelling error is “off” instead of “of”

Name: _____ Reg #: _____ Section: _____