Birla Institute of Technology & Science, Pilani
Work Integrated Learning Programmes Division
First Semester 2023-2024
M.Tech. in Data Science and Engg.

Mid-Semester Test
(EC-2 Regular Paper)

Course No.	: DSECLZG525
Course Title	: Natural Language Processing
Nature of Exam	: Open Book
Weightage	: 30%
Duration	: 2 Hours
Date of Exam	: 16-July-2023 (FN)

No. of Pages = 3
No. of Questions = 8

Note to Students:
1. Please follow all the *Instructions to Candidates* given on the cover page of the answer book.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.
3. Assumptions made if any, should be stated clearly at the beginning of your answer.

Question 1. Introduction [3 Marks]

a) In the following sentences: "She saw a bat on the tree" and "He was playing with wooden bat", the word "bat" has different meanings. What is this sort of ambiguity referred to as? **[1 mark]**
b) In linguistic morphology, the word "better" getting reduced to the root form of "good" is an example of _____, while the word "courses" getting reduced to the root form of "course" is an example of _____. **[2 marks]**

Question 2. N-gram language models [5 Marks]

"***Erring is human. To err is human. Seeing is believing. To see is to believe.***"

Given above training data, use bigram language model to compare the likelihood of below test sentences.
Note: To handle zero probability, use interpolation with weights 0.3 & 0.7 for unigram and bigram respectively. Don't account for punctuation ie., "." for the modelling and predictions.

a) Test Sentence 1: Seeing is human
b) Test Sentence 2: To believe is human

Question 3. Neural language models [3 Marks]

"***Erring is human. To err is human. Seeing is believing. To see is to believe.***"

For the given above training data construct a neural architecture for language modelling for the following case. Mention the dimension at each and every layer and for the weights between layers.

Case: "Using context window of size 4, design a neural network to predict the next word with candidate vocabulary of size 10. Each vocabulary has a word embedding of dimension 3. Use two internal hidden layers. First one with 6 neurons and second with 4 neurons. Each internal hidden layer's must be use sigmoid activation function. "

If the above needs to be changed to accommodate the net to learn the word embedding as well, do the necessary modification in above case study requirement and pictorially draw the final network (No need to draw all the connection between immediate layers).

Question 4. Vector semantics [1.5+1.5=3 Marks]

Answer the following: with respect to Training Corpus D:

Training Corpus D

document - d1: Today is sunny day! I like sunny days

document - d2: I do not like sunny days. I like chocolate

document - d3: I like both sunny and rainy day

Use the TF-IDF to decide on :

The word "sunny" is discriminative in identifying documents in corpus D.

The word "chocolate" is discriminative in identifying documents in corpus D.

Justify your answer.

Question 5. Word embedding [4 Marks]

Given a training corpus: "I live in Bangalore, India", use the skip-gram negative sampling method and answer the following: The initial embedding matrix has dimensions |v| x 3 and is given as follows:

1	1	0
0	1	0
1	0	1
1	1	1
0	0	1

The initial context matrix has |v| x 3 dimension and given as follows:

1	0	0
0	1	0
0	1	1
0	1	0
1	0	1

a) Generate the training dataset for an input word 'I" and context window of 1 next word and hyper parameter value k=3 for the negative sampling task. Use the information available in the question.

b) Calculate the error for the above dataset during the first iteration of skip-gram training.

Question 6. POS tagging [4 Marks]

Tokenize and tag the following sentences using the Brown Corpus table given,

Sentence 1:" he will chair the session"

Sentence 2: "It is a chair"

Sentence 3: "He will race the car."

Sentence 4: "It is a race."

Write the best sequence of tags for sentence 1 , sentence 2 , sentence 3 and sentence .

Sr.No	Examples	Tag	Description
1	He, she, It	PRP	Personal Pronoun
2	will	MD	modal
3	Chair, race	VB	verb
4	is	VBZ	Verb, 3rd person singular
5	a, the	DT	determiner
6	chair, session, race,car	NN	Noun, singular or mass
7	.	DOT	DOT

Question 7. HMM [4 Marks]

By using Viterbi Algorithm, find whether the word "drive" is VB or NN through computing their probabilities for the sentence, "I love to drive". The transition probabilities and the observation likelihood for this corpus are as follows:

	VB	TO	NN	PRP
<S>	0.019	0.0043	0.041	0.067
VB	0.0038	0.035	0.047	0.0070
TO	0.83	0	0.0047	0
NN	0.0040	0.016	0.087	0.0045
PRP	0.23	0.00079	0.0012	0.00014

	I	Love	to	drive
VB	0	0.0082	0	.07
TO	0	0	0.88	0
NN	0	0.0074	0	0.057
PRP	0.48	0	0	0

Assume the values for the third column of the Viterbi table which corresponds to observation 3 (word is "to") and they are VB: 0, TO: .018, NN: 0, PPSS: 0.
(Hint: Show the computations of the last column of the Viterbi table).

Question 8. Topic Modeling [4 Marks]

a) Explain any one real life application of topic modelling. **[1 Marks]**
b) Draw and explain topic distribution using simplex visualization for the 4 topics: Food, Politics, Technology and Sports **[3 Marks]**