Birla Institute of Technology & Science, Pilani
Work Integrated Learning Programmes Division
First Semester 2024-2025
M.Tech. in AIML

Mid-Semester Test
(EC-2 Regular Paper)

Course No. : AIMLCZG530
Course Title : Natural Language Processing
Nature of Exam : Closed Book
Weightage : 30%
Duration : 2 Hours
Date of Exam : 19-Jan-2024

No. of Pages = 3
No. of Questions = 7

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. [2+2=4 Marks] Introduction

a) What type of ambiguity exists in the sentences below:
"I want to read a book" and "I want to book a flight"
Which NLP technique will help to resolve the ambiguity?
Solution:
Syntactic or grammatical ambiguity. [1 mark]
Can be resolved by doing part of speech tagging for word "book" in both sentences and identifying them as noun in first and verb in second sentence.[1 mark]

b) Name the different preprocessing steps in NLP in the below table for the given sentence
"Natural Language Processing is interesting and one of the best courses!!"

Output after processing of sentence		**pre-processing step**
i.	[Natural Language Processing is interesting and one of the best courses]	Removal of punctuation
ii.	[Natural, Language, Processing, is, interesting, and, one, of, the, best, courses]	???
iii.	['Natural', 'Language', 'Processing', 'interesting', 'one', 'best', 'courses']	???
iv.	['Natural', 'Language', 'Processing', 'is', 'interesting', 'and', 'one', 'of', 'the', 'best']	???
v.	['Natur', 'language', 'process', 'is, interest, and, one, of, the, best]	???

SOLUTION

0.5 marks each

Output after processing of sentence	**pre-processing step**

i.	[Natural Language Processing is interesting and one of the best courses]	Removal of punctuation
ii.	[Natural, Language, Processing, is, interesting, and, one, of, the, best, courses]	Word Tokenization
iii.	['Natural', 'Language', 'Processing', 'interesting', 'one', 'best', 'courses']	Removal of stop words
iv.	['Natural', 'Language', 'Processing', 'is', 'interesting', 'and', 'one', 'of', 'the', 'best']	Lemmatization
v.	['Natur', 'language', 'process', 'is, interest, and, one, of, the, best]	Stemming

Question 2. [4 Marks]

A simple language model needs to choose between two candidate translations:
"The weather is cold"
"The weather are cold"
Using the following bigram probabilities:
P(weather|the) = 0.2
P(is|weather) = 0.5
P(are|weather) = 0.1
P(cold|is) = 0.3
P(cold|are) = 0.3

a) Calculate the bigram probability for each sentence [2 marks]
b) Which translation should be selected and why? [2 marks]

Solution:
a) Sentence 1: P(the weather is cold) = P(weather|the) × P(is|weather) × P(cold|is) = 0.2 × 0.5 × 0.3 = 0.03 [1 mark]
Sentence 2: P(the weather are cold) = P(weather|the) × P(are|weather) × P(cold|are) = 0.2 × 0.1 × 0.3 = 0.006 [1 mark]
b) The model should select "The weather is cold" because:
It has a higher probability (0.03 > 0.006)
This indicates it's more likely to be grammatically correct
The probability is 5 times higher than the second candidate [2 marks]

Question 3. [5 Marks] neural language models

1. Consider following neural network which has 3 outputs and 5 inputs [3+2 marks]

Output 1: Creative: Yes: 1, Not sure:0, No: -1
Output 2: Caring: Yes: 1, Not sure:0, No: -1
Output 3: : Rigid: Yes: 1, Not sure:0, No: -1

Inputs:

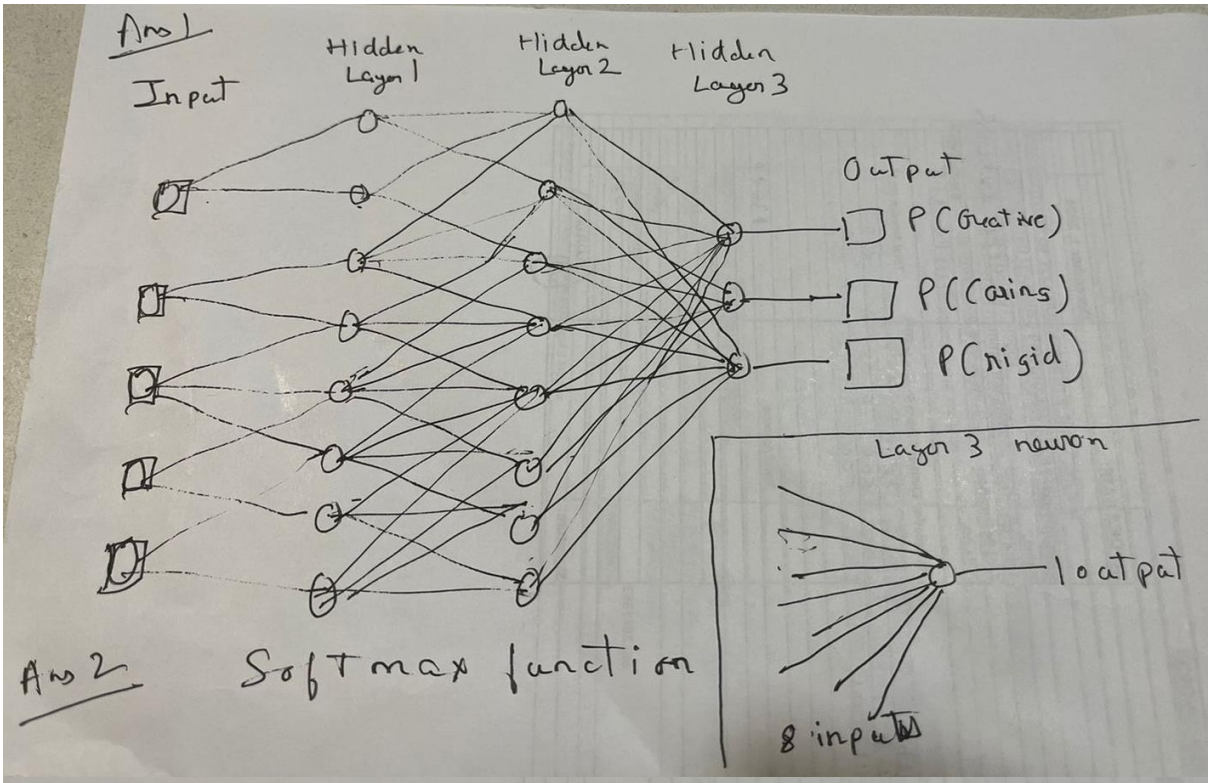
1. Conscientiousness: number b/w 1-5
2. Agreeableness: number b/w 1-5
3. Openness: number b/w 1-5
4. Neuroticism: number b/w 1-5
5. Extraversion: number b/w 1-5

Design a neural network with the above inputs. It shall have 3 hidden layers: layer 1, layer 2 with 8 neurons and layer 3 (classification layer).

All layers to be fully connected

1.1 Draw above network, make suitable assumptions about different layers [3 marks]
1.2 Suggest the act*ivation function to be used for layer 3, draw one* neuron from this layer and mark all the inputs/outputs from it along with the equation of mathematical operation being performed by this neuron, choose suitable function for neuron of this layer [2 marks]

Solution:



Mathematical eqn

$$P(y=1/x) = \frac{e^{x^T w_i}}{\sum_{K=1}^{3} e^{x^T w_K}}$$

where $X = [1, 2, 3]$

1: Creative
2: Caring
3: rigid

Question 4. [4 Marks]

Assume that following is a portion of a vocabulary based on a corpus of text:
Vocabulary = ['international', 'India', 'Islamabad', 'Lanka', 'Delhi', 'Dhaka', 'neighbors', ..., ...]

Based on some scheme like word2vec, the portion of the word embedding matrix M corresponding to the above vocabulary is given below:

M =

1	8	6
3	5	7
2	4	6
4	6	8
1	2	4
6	6	6
6	5	2
…	…	…

(a) Explain the step(s) you will take to extract the embedding vector for 'Lanka', and also write the result: [1 mark]

(b)Derive embedding vectors for between 'Islamabad' and 'Dhaka'? How will you calculate the semantic distance between 'Islamabad' and 'Dhaka'? [2 mark]

(c) Using the word embedding vectors, arrive at embedding vector for the sentence: 'Lanka neighbors India' [1 mark]

Solution

4(a) Explain the step(s) you will take to extract the embedding vector for 'Lanka', and write the result: [1 mark]

The following steps are involved in extracting the embedding vector for any word from the word embedding matrix:

- Derive the one-hot-encoding vector for the target word based on the given vocabulary. For 'Lanka' the one-hot-encoding vector is **[0 0 0 1 0 0 0 …]**
- Take a dot product of this vector with the embedding matrix **M** the resulting vector is the embedding vector for 'Lanka'

[0 0 0 1 0 0 0 …] •

1	8	6
3	5	7
2	4	6
4	6	8
1	2	4
6	6	6
6	5	2
…	…	…

Answer: The embedding vector for 'Lanka' = [4 6 8]

Note: Give 0.5 marks if the steps are adequately explained and 0.5 marks for the correct result.

4(b) How will you calculate the semantic distance between 'Islamabad' and 'Dhaka'? [2 mark]

The 'Cosine Distance' or 'Euclidean Distance' can both be considered as measures of Semantic Distance between two words. Therefore, **any of the following solutions can be accepted**.

'Islamabad' = [2 4 6]

'Dhaka' = [6 6 6]

Based on Cosine Distance

Cosine Distance = 1 – cosine_similarity

Cosine similarity = [2 4 6] **.** [6 6 6] / (|[2 4 6]| x |[6 6 6]|) = (12+24+36)/(7.483 x 10.392) = 0.926

Hence Cosine Distance = 1 – 0.926 = **0.074**

Based on Euclidean Distance

Semantic distance = sqrt((2-6)^2 + (4-6)^2 + (6-6)^2) = sqrt(16 + 4 + 0) = **4.472**

Note: Give 0.5 marks if the steps are correct, even if there are calculation errors

4(c) Using the word embedding vectors, arrive at an embedding vector for the statement 'Lanka neighbours India' [1mark]

The embedding vector for a sentence can be **aggregated** from the embedding vectors of each of the words in the sentence. For the statement 'Lanka neighbours India', the embedding vector can be obtained by either averaging all the constituent embedding word vectors, or by even just summing them up. Either of the following solutions is correct.

'Lanka' = [4 6 8]
'neighbours' = [6 5 2]
'India' = [3 5 7]

Solution 1:

Based 0n averaging the word vectors: [(4+6+3)/3, (6+5+5)/3, (8+2+7)/3] = **[4.333, 5.333, 5.667]**

Solution 2:

Based on summing up the word vectors: [(4+6+3), (6+5+5), (8+2+7)] = **[13, 16, 17]**

Note: Give 0.5 marks if the steps are correct, even if there are calculation errors

Question 5. [5 Marks]

Consider the following word vector embedding's:

i.	**Cat**	[0.1	0.1	0.9]
ii.	**Dog**	[0.7	0.3	0.0]
iii.	**Meow**	[0.0	0.0	1.0]
iv.	**Tree**	[0.1	0.9	0.1]
v.	**Forest**	[0.0	0.8	0.2]

Answer the below: (All answers must be accompanied with detailed steps)

1. Find the word embedding vector representation of the word "Bark" corresponding to the words "Cat", "Meow" and "Dog". **[1 mark]**

2. Given the sentence – "The **dog began to bark at the tree** atop which the cat meowed"

   Train a classifier such that, given the tuple ("bark","dog") where "bark" is the target word and "dog" is the candidate context word, the classifier returns the probability that "dog" is a real context word for "bark".

   - Provide the updated Input weight matrix for the Target Word after one iteration of the Word2Vec algorithm

   Support your answer with detailed steps and rationale on the logic and computation.
   [4 marks]
   The following additional information are provided:
   Use Word2Vec with Skip Gram Classifier with a Single Hidden Layer
   Negative Sampling word is "Forest"
   Activation Function is Sigmoid
   Learning Rate = 0.01
   Precision of computations to be 3 decimal places
   The One Hot Encoded Input Vectors are:

Bark	[1 0 0]
Dog	[0 1 0]
Forest	[0 0 1]

   Initial Embedding Matrix for the Single Hidden Layer

Bark	0.1	0	0.1
Dog	0	0.1	0
Forest	0.1	0	0.1

Initial Embedding Matrix for the Output Layer

Bark	0	0.1	0
Dog	0.1	0	0.1
Forest	0	0.1	0

Solution:

1. Using Parallelogram method: [1 marks]

Bark = Meow – Cat + Dog
= [0.0 0.0 1.0] - [0.10.1 0.9] + [0.7 0.3 0.0]
= [0.6 0.2 0.1]

2.

Step 1 – Forward Propagation (Hidden Layer) [0.5 mark]

- The One Hot Encoded Input Matrix: (I)
[100
010
001]

- Initial Embedding Matrix (W_{input})
[0.1 0.0 0.1
0.0 0.1 0.0
0.1 0.0 0.1]

- Hidden Layer (h) for Target word "bark" = $W^{T}_{input} * I$
= [0.1
0.0
0.1]

Step 2 – Forward Propagation (Sigmoid Output Layer) [1 mark]

- W_{output} (context) *for (dog,forest)*
[0.0 0.1 0.0
0.1 0.0 0.1
0.0 0.1 0.0]

Output Layer = $W_{output}t * h$

= [0.02
0]

Applying Sigmoid Activation,

For Positive Samples: $\sigma(x) = \frac{1}{1+e^{-x}}$

For Negative Samples: $\sigma(x) = \frac{1}{1+e^{x}}$

$\sigma(output\ layer)$ = [0.505
0.500]

Step 3 – Prediction Error [0.5 mark]

Prediction Error = $\sigma(output\ layer)$ – 1-hot encoded vector for context
= [0.505 [1 [-0.495
0.500] -- 0] = 0.500

Step 4 - Backward Propagation (computing ∇W_{input}) step: [1 marks]

Derivative of Loss with respect to Input Word Embeddings for the target word "bark":

$[-0.495^{Transpose}$ [0.1 0 0.1
0.5] 0 0.1 0]
X

= [-0.0495 0.05 -0.0495] { ∇W_{input}}

Step 5 - Updated Weight Matrix by applying Learning Rate [1 mark]

Learning Rate = 0.01 (given)

W^{new}_{input} = [0.1 0.0 0.1] – 0.01 * ∇W = **[0.100 -0.001 0.100]**

This the updated weight matrix for the Target Word after one iteration of the Word2Vec algorithm

Question 6.

Using an HMM tagger to disambiguate the POS tag for the word **"bank"** in the following sentence, given the transition and emission probabilities below:

Went to the bank.

Emission probabilities

	Bank	Went	To	The
VB	0.5	1	0	0
TO	0	0	1	0
NN	0.5	0	0	0
DET	0	0	0	1

Transition probabilities

	VB	TO	NN	Det
VB	0.4	0.1	0.2	0.3
TO	0.2	0.7	0.1	0
NN	0.3	0.2	0.4	0.1
Det	0.1	0	0.2	0.7

Use the provided transition and emission probabilities to disambiguate the POS tag for the word "bank" in the sentence.

Solution:

Went to the bank

Possible taggings are:

i. VB → TO → Det → NN
ii. VB → TO → Det → VB

We are interested in disambiguating "bank" in the above phrase. Hence the computations will be:

P(NN|Det) = 0.2
P(VB|Det) = 0.1
P(bank | NN) = 0.5
P(bank | VB) = 0.5

For tagging (i): P(NN|Det) x P(bank|NN) = 0.2 x 0.5 = 0.1
For tagging (ii): P(VB|Det) x P(bank|VB) = 0.1 x 0.5 = 0.05

Hence, tagging (i) is the preferred POS tag, i.e.
VB → TO → Det → NN

Question 7.

Fill up the Viterbi table for the sentence – "She dances". The tag transition probabilities and word emission probabilities, for the corpus used, are given below:

Tag Transition Probabilities:

	VB	**MD**	**PRP**
VB	0.1	0.2	0.05
MD	0.5	0.3	0.01
PRP	0.4	0.4	0.9
START	0.1	0.2	0.7

Word Emission Probabilities:

	She	dances
VB	0	0.3
MD	0	0
PRP	1	0

Viterbi Table:

	She	dances
VB		
MD		
PRP		

Tags:

- **VB**: Verb Base Form
- **MD**: Modal
- **PRP**: Personal Pronoun

You need to fill out the Viterbi table for the sentence "She dances" using the given transition and emission probabilities.

Solution:

	She	dances
VB	0	0.084
MD	0	0
PRP	0.7	0