**Birla Institute of Technology & Science, Pilani**

**Work Integrated Learning Programmes Division**

**First Semester 2023-2024**

**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 : 21-01-2024\_FN

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. [3+2+2=7 Marks]**

1. Find the ambiguity of the below sentences and justify your answer **[3 marks]**

a) The tank is full of water. I saw a military tank.

Soln:

Semantic ambiguity

b) Before the professor left the stage, the play begins

Syntactic ambiguity

c) She is looking for a match

Lexical ambiguity

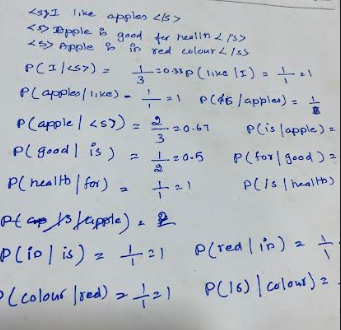
1. Given is the following toy corpus. Calculate all the bigram probabilities. **[2 marks]**

**<s> I like apples </s>**

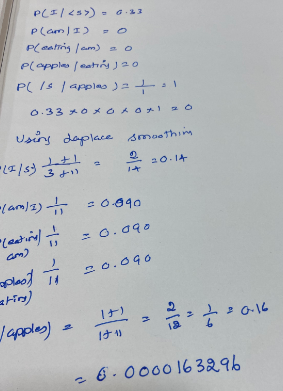
**<s> Apple is good for health</s>**

**<s> Apple is in red colour </s>**

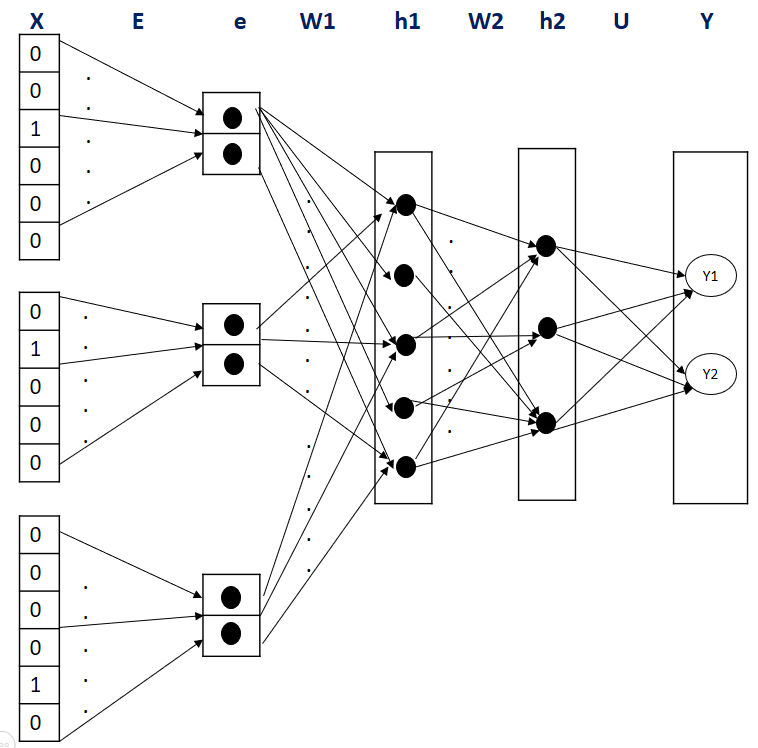
**Soln:**



1. For above training data in (i), Calculate the probability of below sentence using raw bigram probabilities and using Laplace smoothing**, <s> I am eating apples</s>**  **[2marks]**



**Question 2. [4+3 = 7 Marks]**

1. Study the below neural network designed for learning word embedding along with other NLP application stated below and answer the following questions.

If the input layer ‘X’ denote the one hot encoding of the vocabulary, “e” is the embedding layer, “h1”,”h2” are hidden layers and “Y” is the output layer emitting continuous valued output, identify no more than 3 issues/error in the architecture and suggest modification to suit the below use case requirement. If there are no corrections required, then mention “No Error” explicitly.

Use Case: Neural network in language modelling for sentence completion

Given a training corpus with below vocabulary each vectorized with four dimensions, and following test sentence phrase, the neural network, should have predictive ability to identify next best word to fill in the blank of the test sentence, by analyzing context window with five tokens.

Vocabulary: {he, she, bat, tree, wooden, park, playing, saw, was, a, on, the, with, in, morning, evening} Test Sentence: “on a morning he was playing in the \_\_\_\_\_\_\_\_”

**Solution & Marking Scheme:**

1 Mark: Each for below error & suggestions:

* No.of.one-hot-encoding vectors must be 5 but is three in number
* Output needs |V|=16 no.of.neuron for Y-Pred
* One-hot-encoding size is not consistent 🡪 Change to size of vocabulary = 16
* Embedding size ‘E’ should be 4 in dimension but its two

1. The number of times each word appears in different documents is given in the table below.

Calculate the TF-IDF value for each term in D1. **[1.5 mark]**

Find the word embedding for each term using TF-IDF value. Find which words are closest using TF-IDF word embedding. **[0.5 mark]**

Which documents are more similar to each other? **[0.5 mark]**

What is the disadvantage of using the TF-IDF values for the word embedding’s? **[0.5 mark]**

|  |  |  |  |
| --- | --- | --- | --- |
|  | D1 | D2 | D3 |
| NLP | 10 | 0 | 0 |
| is | 50 | 66 | 89 |
| extremely | 20 | 22 | 12 |
| interesting | 30 | 32 | 11 |
| course | 20 | 0 | 0 |

**Solution**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | D1 | N | df | idf |
| NLP | 10 | 3 | 1 | 0.477 |
| tf | 1.04 |  |  |  |
| tf-idf | 0.496 |  |  |  |
| is | 50 | 3 | 3 | 0 |
| tf | 1.70 |  |  |  |
| tf-idf | 0 |  |  |  |
| extremely | 20 | 3 | 3 | 0 |
| tf | 1.32 |  |  |  |
| tf-idf | 0 |  |  |  |
| interesting | 30 | 3 | 3 | 0 |
| tf | 1.49 |  |  |  |
| tf-idf | 0 |  |  |  |
| course | 20 | 3 | 1 | 0.477 |
| tf | 1.32 |  |  |  |
| tf-idf | 0.63 |  |  |  |

Find the word embedding for each term using TF-IDF value.

NLP [0.496,0,0]

is [0,0,0]

extremely [0,0,0]

interesting[0,0,0]

course [0.63,0,0]

Find which words are closest using TF-IDF word embedding. **[0.5 mark]**

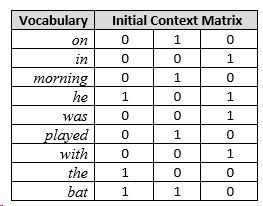
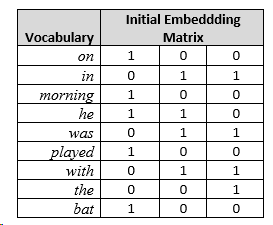
Which documents are more similar to each other? **[0.5 mark] D2 and D3**

What is the disadvantage of using the TF-IDF values for the word embedding’s? **[0.5 mark] - Sparsity**

**Question 3. [1.5+3+1.5=6 Marks]**

Given a training corpus: “*played in the morning*”, use the skip-gram negative sampling method and answer the following: The initial embedding matrix and initial context matrix has dimensions |v| x 3 and is given as follows:

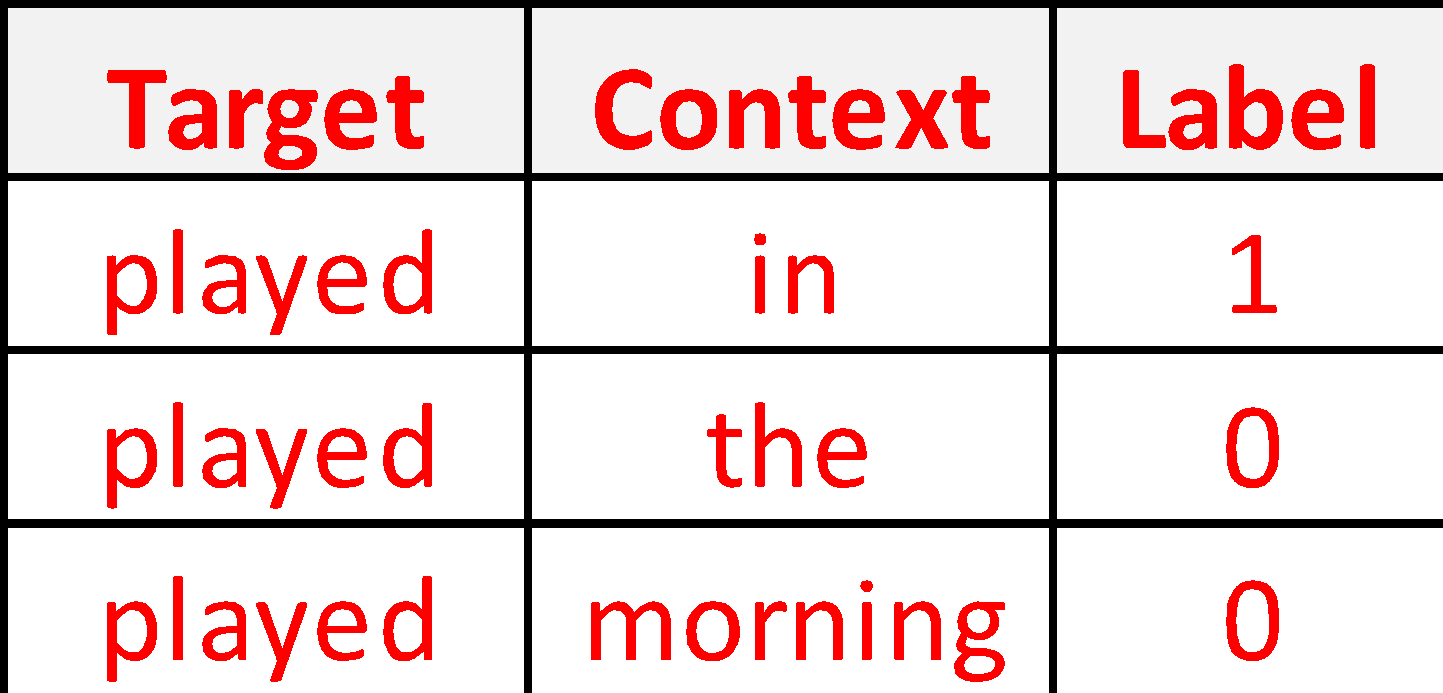
***Note:*** *No need to update or show any weights other than necessary for below questions. Follow only the approach as discussed in class. i.e., Simplified Skip gram negative sampling with binary classification model. Round all the calculations to exactly two decimal places.*



1. Generate the training dataset for an input target word “*played*” and context window of 1 next word and hyper parameter value k=2 for the negative sampling task. Use the information available in the question.
2. Calculate the error for the above dataset for only the first iteration of skip-gram training, with only one hidden layer.
3. Explain in no more than 40 words, why skipgram algorithm training was modified from multiclass to binary classification task.

Solution & Marking Scheme:

1. Training data: 1.5 mark

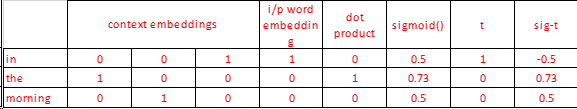


1. Error:

Dot product = 1 mark

Sigmoid = 1 mark

Error = 1 mark



1. 0.5 mark Faster/Ease of Computation.

1 mark: Instead of updating all weights, 1 iteration updates only the context word & target word’s embedding for a training instance

**Question 4. [5 Marks]**

Find the appropriate POS tag using statistical model with bigram assumption, for the word “cook” in the sentence,

“He will cook the food”

|  |  |  |
| --- | --- | --- |
| Word to Tag combination | | |
| 1 | p(cook/VB) | 0.0056 |
| 2 | p(cook/NN) | 0.0072 |
| 3 | p(the/DT) | 0.014 |
| 4 | P(food/NN) | 0.00047 |

|  |  |  |
| --- | --- | --- |
| Tag to Tag combination | | |
| 1 | p(VB/MD) | 0.78 |
| 2 | p(NN/MD) | 0.14 |
| 3 | p(DT/VB) | 0.075 |
| 4 | p(DT/NN) | 0.026 |

Note:

NN- Noun

VB-Verb

MD-Modal

DT-Determiner

**Solution:**

**p(VB/MD)\* p(DT/VB)\* p(cook/VB) =** 0.0003276

**p(NN/MD) \* p(DT/NN) \*p(cook/NN) =** 0.000026208

**The most appropriate POS tag is VB**

**Question 5. [5 Marks]**

By using Viterbi Algorithm, fill the Viterbi table for the sentence, “He will fight”. The tag transition probabilities and the word likelihood for this corpus are as follows:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Tag transition probabilities | MD | NN | VB | PRP |
| MD | 0.000008 | 0.31 | 0.46 | 0.0056 |
| NN | 0.000096 | 0.209 | 0.658 | 0.00068 |
| VB | 0.001 | 0.05 | 0 | 0.008 |
| PRP | 0.08 | 0.02 | 0.001 | 0.00001 |
| START | 0.008 | 0.000934 | 0.05677 | 0.08 |

|  |  |  |  |
| --- | --- | --- | --- |
| Word likelihood probabilities | he | will | fight |
| MD | 0 | 0.8 | 0 |
| NN | 0 | 0.2 | 0.4 |
| VB | 0 | 0 | 0.6 |
| PRP | 1 | 0 | 0 |

|  |  |  |  |
| --- | --- | --- | --- |
| Viterbi Table | HE | WILL | FIGHT |
| NN |  |  |  |
| VB |  |  |  |
| MD |  |  |  |
| PRP |  |  |  |

Note:

|  |
| --- |
| PRP: PERSONAL PRONOUN  MD:MODAL  VB:VERB BASE FORM  NN:NOUN, SINGULAR OR MASS |

**Solution**

|  |  |  |  |
| --- | --- | --- | --- |
| Viterbi Table | HE | WILL | FIGHT |
| NN | 0 | 0.00032 | 0.00063488 |
| VB | 0 | 0 | 0.00141312 |
| MD | 0 | 0.00512 | 0 |
| PRP | 0.08 |  | 0 |