**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

No. of Pages = 3

# No. of Questions = 5

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. [4+2+1=7 Marks]**

1. Consider the training set:                    [**2 +2=4marks**]

*The Ali Baba and the Forty Thieves*

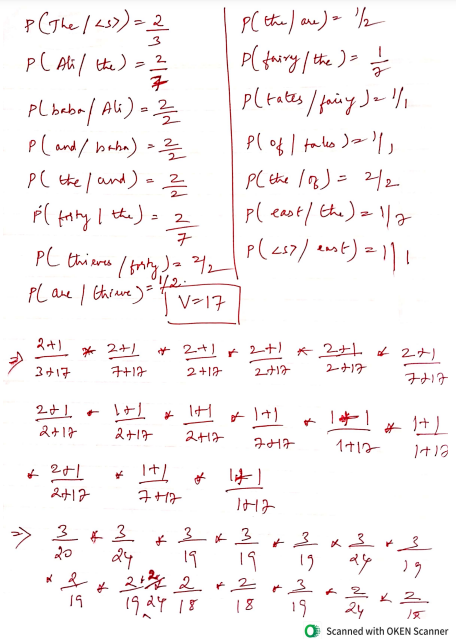
*These are the fairy tales of the east*

*The stories of the Ali Baba and the Forty Thieves are translated in many languages*

Compute the probability of the sentence using the bigram model and add 1 smoothing for the below sentence. Include start and end symbol in your calculations.

*The Ali Baba and the Forty Thieves are the fairy tales of the east*

*Soln:*

**

*A piece of paper with red writing

Description automatically generated*

1. Why the below sentence is ambiguous. Explain using illustrative examples **[2marks]**

Every boys love some girls

Solution: The ambiguity is whether every man loves the same woman as in ‘Every boys loves Priyanka Chopra’ or every boys has a different woman whom he loves as ‘Every boy loves his wife’.

1. For a given test set T=w1, w2,…wn and language models M1 and M2 .How do you decide which language model is better and why. **[1mark]**

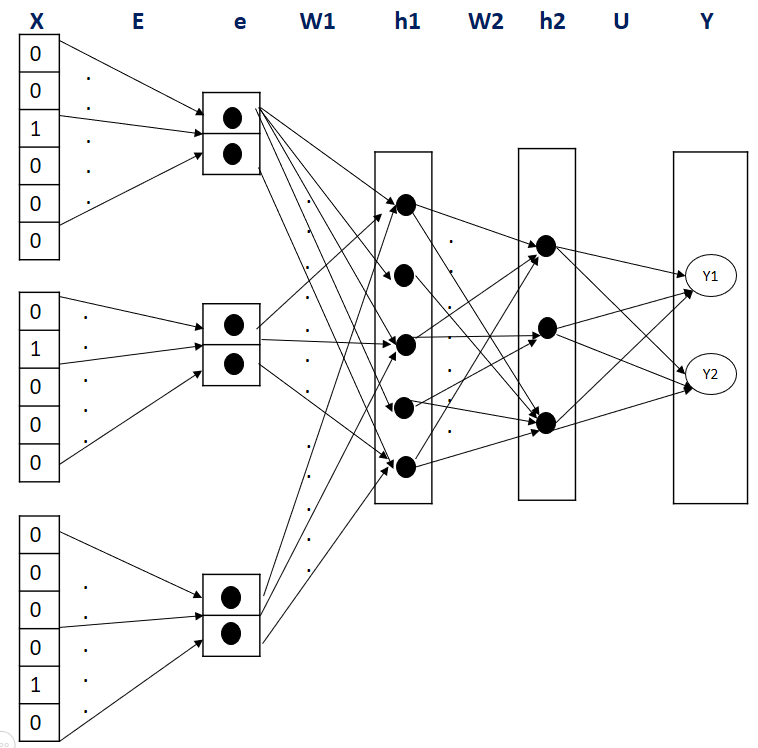
Solution:

Language models on a test set T = w1w2 . . . wn can be compared by calculating the perplexity (PP) each language model assigns to T. PP (T) = P (w1 . . . wn) − 1. the lower the perplexity the better the model. So, between M1 and M2 the model with lower P (T) is the better model.

**Question 2.**

1. **Neural language models [4 Marks]**

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 with sigmoid activation function and “Y” is the output layer emitting continuous valued output, identify no more than 2 issues/error in the architecture for each of the below two scenarios: If there are no corrections required then mention “No Error” explicitly. Assuming no bias is included in the network design, find the dimension (matrix size) w.r.t to section/layer “W1” and “U”

*Use Case: Given a training corpus with below vocabulary each vectorized with three dimensions, and following test sentence phrase, the neural network, should have predictive ability to classify given phrase of two tokens into positive, negative or neutral sentiment.*

*Vocabulary: {today, tomorrow, sunny, rainy, day, like, season, dislike, ice-cream, chocolate, is, was, both, I}*

*Test Sentence: “I dislike ice-cream”*

**Solution & Marking Scheme:**

1 Mark: Each for below error & suggestions: (any four is accepted)

* No.of.one-hot-encoding vectors must be 2 but is three in number
* Output needs 3 no.of.neuron for Y-Pred but is two
* One-hot-encoding size is not consistent 🡪 Change to size of vocabulary = 14
* Embedding size ‘E’ should be 3 in dimension but its two
* Activation function in output layer must be softmax but in above design statement, its said to emit continuous values output

1. Given below term document matrix.

Identify similar words and similar documents. Justify your answer. **[2 marks]**

Use TF-IDF to identify informative word for the doc 3. **[1 mark]**

|  |  |  |  |
| --- | --- | --- | --- |
|  | Doc 1 | Doc 2 | Doc 3 |
| AIML | 0 | 20 | 0 |
| BITS | 20 | 0 | 30 |
| PILANI | 30 | 20 | 50 |
| MTECH | 20 | 30 | 30 |
| WILP | 10 | 10 | 20 |

**Solution**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | D1 | N | df | idf |
| AIML | 0 | 3 | 1 | 0.477 |
| tf | 0 |  |  |  |
| tf-idf | 0 |  |  |  |
| BITS | 20 | 3 | 2 | 0.176 |
| tf | 1.32 |  |  |  |
| tf-idf | 0.232 |  |  |  |
| PILANI | 30 | 3 | 3 | 0 |
| tf | 1.49 |  |  |  |
| tf-idf | 0 |  |  |  |
| MTECH | 20 | 3 | 3 | 0 |
| tf | 1.32 |  |  |  |
| tf-idf | 0 |  |  |  |
| WILP | 10 | 3 | 3 | 0 |
| tf | 1.04 |  |  |  |
| tf-idf | 0 |  |  |  |

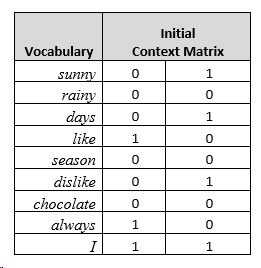
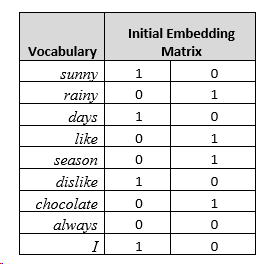
Similar docs-doc 1 and doc 2, similar terms- MTECH and WILP

BITS is most informative word in doc 3

**Question 3. Word embedding [1+3+2=6 Marks]**

Given a training corpus: “*I always like rainy days*”, use the skip-gram negative sampling method and answer the following: The initial embedding matrix and initial context matrix has dimensions |v| x 2 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. Use the gradient descent update equation as is. No need to show derivations of equation. Round all the calculations to exactly two decimal places.*



1. Generate the training dataset for an input target word “*like*” and context window of one previousword and hyper parameter value k=3 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. Using the result in part b), update the context weight of only the word “*always*” with learning rate=1.

Solution & Marking Scheme:

1. Training data: 1 mark

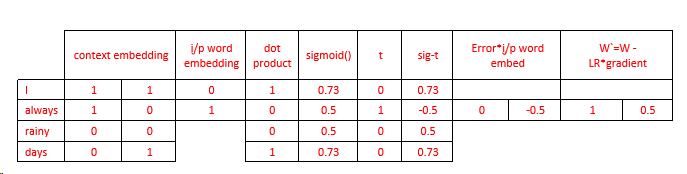


1. Error:

Dot product = 1 mark

Sigmoid = 1 mark

Error = 1 mark



1. Weight update for positive context word: Last two columns in above table

Gradient = 1 mark

New weight = 1 mark

**Q4 – 6 Marks**

**Module 5: POS Tagging (Problem) – 6 Marks**

1. **List the possible sequences for the sentence, “He will fly”. ( 2 Marks)**

|  |  |  |
| --- | --- | --- |
| **1** | **He** | **PRP** |
| **2** | **will** | **NN, MD** |
| **3** | **fly** | **NN, VB** |

1. **Find the best sequence of POS tags , if the fly is VB. ( 4 Marks)**

|  |  |  |
| --- | --- | --- |
| **Word to Tag combination** | | |
| **1** | **p(fly/VB)** | **0.0041** |
| **2** | **p(fly/NN)** | **0.0083** |
| **3** | **p(will/NN)** | **0.0056** |
| **4** | **p(will/MD** | **0.0072** |

|  |  |  |
| --- | --- | --- |
| **Tag to Tag combination** | | |
| **1** | **p(VB/MD)** | **0.51** |
| **2** | **p(VB/NN)** | **0.09** |
| **3** | **p(MD/PRP)** | **0.78** |
| **4** | **p(NN/PRP)** | **0.14** |

**Solution:**

1. The possible sequence of tags are: **( 2 Marks)**
   1. He/PRP will/MD fly/NN
   2. He/PRP will/NN fly /NN
   3. He/PRP will/MD fly /VB
   4. He/PRP will/NN fly /VB

**B) The best sequence of tags ( 4 Marks)**

**p(MD/PRP)\* p(VB/MD)\* p(will/MD) =** 0.00286

**p(NN/PRP) \* p(VB/NN) \*p(will/NN) =** 0.00007056

**The most appropriate POS tag for will is MD.**

**The best sequence of tags is He/PRP will/MD fly/VB**

**Q5 – 4 Marks**

**Module 6: Hidden Markov Models (Problem) – 4 Marks**

1. **How forward algorithm is different from Viterbi algorithm in HMM for computing the probability of an observed sequence given a model. ( 1 Mark)**

**Solution:**

* **Nearly identical to Viterbi; replace the MAX with a SUM**

****

1. Frame the Transition matrix for the given transition of states: **( 3 Marks)**

A diagram of a network

Description automatically generated

**A white grid with red text

Description automatically generated**