Roll No											

PRESIDENCY UNIVERSITY
BENGALURU

SET A

SCHOOL OF ENGINEERING
END TERM EXAMINATION - JAN 2024

Semester : Semester VII - 2020

Course Code : CSE3014

Course Name : Fundamentals of Natural Language Processing

Program : B.Tech.

Date : 04-JAN-2024

Time : 9:30AM - 12:30 PM

Max Marks : 100

Weightage : 50%

Instructions:

(i) Read all questions carefully and answer accordingly.
(ii) Question paper consists of 3 parts.
(iii) Scientific and non-programmable calculator are permitted.
(iv) Do not write any information on the question paper other than Roll Number.

PART A

ANSWER ALL THE QUESTIONS **5 X 2M = 10M**

1. Mention the name of the scientist who devised the Imitation Game.

(CO1) [Knowledge]

2. Mention one morphologically rich language and one morphologically poor language.

(CO1) [Knowledge]

3. Mention the university where an experiment in machine translation was conducted by IBM in 1954.

(CO1) [Knowledge]

4. State true or false. The probabilty of a parse tree is the product of the derivations of the parse tree.

(CO1) [Knowledge]

5. Given an observation matrix O of size n*n, write the expression to calculate the value of the ith row and jth column (i.e. E[i][j]) in the expectation matrix.

(CO1) [Knowledge]

PART B

ANSWER ALL THE QUESTIONS **5 X 10M = 50M**

6. Consider the problem of title casing. Title casing is where we capitalize the first letter of ***some parts of speech***, while other words start with lower case. Explain, using a HMM, how we will perform title casing, given that we have a very large list of titles, but no part-of-speech tagger, or a part-of-speech tagged corpus. You will need to explain the states, as well as how you calculate the initial, transition and emission probabilities.

(CO3) [Comprehension]

7. Calculate the BLEU score between the following pair of sentences. Consider that we use only unigrams and bigrams, with weights of 0.5 each (i.e. no trigrams or 4-grams).
Candidate: All I want for Christmas is your baby
Reference: All I want for Christmas is you baby

(CO4) [Comprehension]

8. Consider the following probabilistic context free grammar:

- $S \rightarrow NN\ VP\ (0.50)$
- $VP \rightarrow VB\ NP\ (0.20)$
- $NP \rightarrow NN\ PB\ (0.40)$
- $PB \rightarrow PP\ NN\ (0.30)$
- $NN \rightarrow children\ (0.15) \mid songs\ (0.12) \mid friends\ (0.20)$
- $VB \rightarrow hear\ (0.30)$
- $PP \rightarrow with\ (0.10)$

Starting from the non-terminal symbol S, derive a parse tree, **such that the probability of the parse tree** is 1.296×10^{-6} for the sentence "**children hear songs with friends**". Draw the parse tree from the derivations.

(CO3) [Comprehension]

9. **BIOSE** is another variant of the **BIO** tags, where we have the following expansions:

- **B** = Beginning of the Named Entity span.
- **I** = Inside the Named Entity span.
- **O** = Outside the Named Entity span.
- **S** = Single word Named Entity
- **E** = Ending of the Named Entiity span

For example, the named entity "United Arab Emirates" will be tagged as "B-LOCATION I-LOCATION E-LOCATION" (NOTE: B and E take precedence over I). Similarly, the name "Mausam" will be tagged as "S-PERSON". Now, consider the following sentences:

1. European authorities fined [**Google**] a record $5.1 billion on Wednesday for abusing its power in the mobile phone market and ordered the company to alter its practices.
2. [**Barry Schwartz**] entered the classroom and asked questions to the students about human nature and thinking skills.
3. [**Barcelona**] is the capital of [**Catalunya**] in [**Spain**].
4. [**Narendra Modi**] is the Prime Minister of [**India**].
5. [**Jimmy Doolittle**] led a famous raid on [**Tokyo**] during World War II.

For each of words in the [**spans**], assign the **appropriate BIOSE tag**. Assume that the **only NER classes** are PERSON, LOCATION, and ORGANIZATION.

(CO3) [Comprehension]

10. Consider a situation where we classify instances into a set of **K** classes. However, we normalize the weights in the weight matrix using the following equation:

$W_N[i][j] = \frac{W[i][j]}{|K-1|^c}$,

where

$W_N[i][j]$ is the value of the cell (i, j) of the **normalized** weight matrix,

$W[i][j]$ is the value of the cell (i,j) of the **unnormalized** weight matrix,

K is the number of classes, and

c is the power of the weights (0 for unweighted Kappa, 1 for linear weighted Kappa, 2 for quadratic weighted Kappa, 3 for cubic weighted Kappa, etc.)

For a given observation / confusion matrix, **show that** the value of the Kappa agreement is the same whether we use normalized or unnormalized weights.

To help you out, write down the different equations for Kappa using the **normalized** weight matrix (κ_n) and the **unnormalized** weight matrix (κ). Also note that, for a given value of c and K, the value of $|K-1|^c$ is a constant.

(CO2) [Comprehension]

PART C

ANSWER ALL THE QUESTIONS **2 X 20M = 40M**

11. Two annotators - A1 and A2 - use the following part-of-speech tags:

- NN = Noun
- VB = Verb
- JJ = Adjective
- RB = Adverb
- FW = Function Word (all other words)
- PM = Punctuation mark

The following is their annotations for the tokenized text "You are a secret agent man , who is after the secret plans . How do you act when they do not know you are a spy ?"

- A1 = FW VB FW JJ NN NN PM FW VB RB FW JJ NN PM FW VB FW VB FW FW VB FW VB FW VB FW NN PM
- A2 = FW VB FW JJ NN NN PM FW VB FW FW JJ NN PM FW VB FW VB FW FW VB RB VB FW VB FW NN PM

Construct the observation, expectation and weight matrices. Based on that, calculate the agreement between the two annotators using the **appropriate** Kappa(s), as well as the **percentage agreement** between the annotators.

(CO4) [Application]

12. Tag the following text: "**the fans watch the races**" using the Viterbi algorithm. Assume that you have only **3 tags** - DT, VB, and NN. You can use the following tables:

Emission Probability:

Emission	the	fans	watch	races
DT	0.5	0	0	0
NN	0	0.1	0.3	0.1
VB	0	0.2	0.15	0.3

Transition Probability:

Transition	DT	NN	VB
$(START)	0.66	0.33	0.01
DT	0	0.8	0.2
NN	0	0.5	0.5
VB	0.5	0.5	0

Draw the trellis. For each **non-zero emission probability** node, calculate the Viterbi probabilities as well as the back pointers. Then, you should tag the sentence correctly.

(CO3) [Application]