Roll No											

GAIN MORE KNOWLEDGE
REACH GREATER HEIGHTS

PRESIDENCY UNIVERSITY
BENGALURU

SCHOOL OF ENGINEERING

END TERM EXAMINATION - JUN 2023

Semester : Semester VI - 2020

Course Code : CSE3014

Course Name : Sem VI - CSE3014 - Fundamentals of Natural Language Processing

Program : CAI&CST

Date : 14-JUN-2023

Time : 9.30AM -12.30PM

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 2 = 10M)**

1. Recall the HISK normalization function (nHISK(x,y)) given by the following equation:

$$nHISK(x,y) = \frac{HISK(x,y)}{\sqrt{HISK(x,x) \times HISK(y,y)}}$$

Write down the range of the HISK normalization function.

(CO1) [Knowledge]

2. List any 2 examples of static dense word representations.

(CO2) [Knowledge]

3. Expand GPE in the context of named entity recognition.

(CO3) [Knowledge]

4. Mention the technique which we use to penalize candidate sentences which are **shorter** than the reference sentences.

(CO4) [Knowledge]

5. Out of the following evaluation metrics, select the six that **CAN** be used for part-of-speech tagging: (a) Accuracy, (b) BLEU, (c) Cohen's Weighted Kappa, (d) Cohen's Unweighted Kappa, (e) Classification Error, (f) F1-Score, (g) Precision, and (h) Recall. NOTE: If you write any wrong evaluation metric, you will be getting 0 marks for this question.

(CO3) [Knowledge]

PART B

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

6. German is a language where, in addition to the regular words of English which are capitalized, all *nouns* (not just proper nouns) also begin with a capital letter, even when inside a sentence. Let us say that converting an English sentence to a German format (which all the common nouns are also capitalized) is called *Germanization* and the text which has this property is called *Germanized* text. For example, the sentence "Long years ago, we made a tryst with destiny" becomes "Long Years ago, we made a Tryst with Destiny". While we can easily do this with a part-of-speech tagger, let us consider the situation where we do not have either a part-of-speech tagger, or even a part-of-speech tagged corpus! However, we do have a corpus of Germanized text. Explain how we will use that corpus to perform Germanization.

(CO3) [Comprehension]

7. **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:

- 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.
- [**Barry Schwartz**] entered the classroom and asked questions to the students about human nature and thinking skills.
- [**Barcelona**] is the capital of [**Catalunya**] in [**Spain**].
- [**Narendra Modi**] is the Prime Minister of [**India**].
- [**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.

(CO1) [Comprehension]

8. For the following sentence pairs, calculate their BLEU scores:

- Candidate: "And a heartbreaking new year". Reference: "And a happy new year".
- Candidate: "One day when my dreams". Reference: "Some day when my dreams come true"
- Candidate: "The desire to implement a dream in your heart about you is lying". Reference: "A dream is a wish your heart makes when you are fast asleep"
- Candidate: "Stuff my heart agricultural to know he wants things to keep in mind". Reference: "Things my heart used to know things it yearns to remember".
- Candidate: "And he went to his eunuchs". Reference: "And away to his castle we will go"

Consider using only modified unigram and modified bigram precisions (as well as Brevity Penalty whereever applicable). Show the necessary steps of working (calculating modified precisions, brevity penalties, etc.) as well for full credit.

(CO4) [Comprehension]

9. 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]

10. Professor SAM wants to create a document corpus. So, he takes 20 *unlabeled* documents, and asks 2 annotators - PCM & PGM - to label them as either COMEDY or TRAGEDY. Here is the result of the classifications by the 2 annotators:

Document	PCM Label	PGM Label
D01	Tragedy	Comedy
D02	Comedy	Comedy
D03	Tragedy	Comedy
D04	Comedy	Comedy
D05	Comedy	Comedy
D06	Comedy	Comedy
D07	Tragedy	Comedy
D08	Tragedy	Comedy
D09	Tragedy	Tragedy
D10	Tragedy	Tragedy
D11	Tragedy	Comedy
D12	Comedy	Comedy
D13	Comedy	Comedy
D14	Tragedy	Comedy
D15	Comedy	Comedy
D16	Tragedy	Tragedy
D17	Tragedy	Tragedy
D18	Tragedy	Tragedy
D19	Tragedy	Tragedy
D20	Tragedy	Comedy

Construct the observation, expectation and weight matrices and help him find the agreement between the two annotators using all 3 versions (unweighted, linear weighted, and quadratic weighted) of the Cohen's Kappa metric.

(CO3) [Comprehension]

PART C

ANSWER ALL THE QUESTIONS **(2 X 20 = 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 wearing your squeaky shoes , and right there taking a snooze , is a tiger , so how do you walk on by ?"

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

Calculate the agreement between the two annotators using the **appropriate** Kappa.

(CO1) [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	0.2	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)	0.8	0.2	0
DT	0	0.9	0.1
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 probability. Then, you should tag the sentence.

(CO3) [Application]