Roll No											

PRESIDENCY UNIVERSITY
BENGALURU

SCHOOL OF ENGINEERING
MID TERM EXAMINATION - APR 2023

Semester : Semester VI - 2020

Course Code : CSE3014

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

Program : CAI,CST

Date : 15-APR-2023

Time : 09:30AM - 11AM

Max Marks : 60

Weightage : 30%

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. Mention the name of the group who published a report in 1966 about the lack of growth in NLP.

(CO1) [Knowledge]

2. Morphological segmentation involves splitting a word into individual units. Mention the name of those units.

(CO1) [Knowledge]

3. Stopwords are words which are very frequently used in NLP. Consider a situation where weigh the counts of words by their tf-idf values. Mention the value of a the weighted count (weighted by the product of the tf and the idf) of a stop word, that is present in all the documents of a corpus.

(CO2) [Knowledge]

4. Mention any 2 multilingual pre-trained language models for Indian languages

(CO2) [Knowledge]

5. List any **two** activation functions, their formulae and the range of values that they take.

(CO2) [Knowledge]

PART B

ANSWER ALL THE QUESTIONS **(4 X 5 = 20M)**

6. Consider a sentiment analysis classifer that classifies texts into **3 classes** - positive, negative, and neutral. The results of the classification are as follows in the given confusion matrix.

Confusion Matrix for **300 documents**, of which 100 documents are positive, 100 documents are neutral and 100 documents are negative.

	Positive	**Neutral**	**Negative**
Positive	50	30	20
Neutral	40	50	10
Negative	10	30	60

Assuming that each class actually has **100 documents**, calculate the **accuracy of the classifier**, as well as the **precision, recall, and F1-scores** of **all 3 classes**.

(CO1) [Comprehension]

7. Compute the **edit distance** for the given pair of words and substitution cost which you are allotted based on the **last digit of your roll number**. Assume an insertion cost of +1 and a deletion cost of +1.

Allotments of word1, word2, and substitution costs, based on roll number.

Roll No. Ending	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
Substitution Cost	1	1	1	1	1	2	2	2	2	2
word1	sitting	donkey	grain	table	hello	kitten	money	grail	stall	helm
word2	kitten	money	grail	stall	helm	sitting	donkey	grain	table	hello

(CO1) [Comprehension]

8. Assume that we are using a small, **26-dimension** vector to represent our words, such that each dimension represents the **count of the character** (from a to z) of our words. Eg. "sandeep" = [1, 0, 0, 1, 2, 0, 0, 0, 0, 0, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]. For each word pair, compute the **dot product** and **cosine similarity**.

- word1 = sitting, word2 = kitten
- word1 = donkey, word2 = money
- word1 = grain, word2 = grail
- word1 = table, word2 = stall
- word1 = hello, word2 = helm

(CO2) [Comprehension]

9. Consider the following documents (Yes, each bullet point is a **document**):

- Principles of Artificial Intelligence
- Artificial Intelligence for Gaming
- Artificial Intelligence and Machine Learning
- Artificial Intelligence for Game Development

Assume only the following **terms**:

- Principles
- Artificial
- Intelligence
- Gaming
- Machine
- Learning
- Game
- Development

Write down the **raw counts matrix**, and generate the **TF-IDF matrix**, whose elements are weighted by the product of the **TF** and the **IDF**. Consider that the logarithm we are using is in **base 10**.

(CO2) [Comprehension]

PART C

ANSWER ALL THE QUESTIONS **(2 X 15 = 30M)**

10. A Naive Bayes classifier is used to classify a number of reviews. The following table displays the annotated labels:

Sentence	Label
I will always cherish the original misconception I had of you	NEG
I find it rather easy to portray a businessman	POS
Being bland, rather cruel and incompetent comes naturally to me	POS
It is like an all-star salute to Disney's cheesy commercialism	NEG
Detecting sarcasm is very easy ;)	POS

Predict the class of the reviews using the following table of counts with add-1 smoothing to calculate the scores of each sentence for each class. Assume a prior probability of 0.5 for both the positive and negative classes.

word	count(+)	count(-)	word	count(+)	count(-)
all-star	3	0	I	5	5
bland	1	3	incompetent	1	4
businessman	2	1	misconception	1	3
cheesy	2	3	naturally	3	1
cherish	5	0	original	3	1
commercialism	2	2	rather	2	2
cruel	0	3	salute	1	0
detecting	2	1	sarcasm	2	4
easy	4	0	very	3	1
find	3	2	;)	5	0

Construct the **confusion matrix** and **calculate** the **accuracy of the classifier**, as well as the **precision, recall and F1-score** for **BOTH** the positive and negative classes.

(CO2) [Application]

11. Consider the following movie review: "When I need an **amusing** diversion, nothing helps quite like watching one of those *dreadful* 50's sci-fi flicks. Ed Wood's *infamous* film is a good choice too. I can forgive it for some of its, let us say ... *imperfections*: anthropomorphic aliens who speak English; women aliens who wear lipstick; the *hammy*, *sophomoric* acting; the *dime-store* special effects ... But there's really no excuse for a mickey mouse script. You get the feeling that the film was put together by a *quarrelsome* committee of third graders, and aimed at an audience of chimpanzees. And yet, specifically because of its technical *crudeness*, the film is **fun** to watch. We may not want to admit it, but the film gives us viewers a chance to feel **superior** to Ed Wood; we get to conjecture that even we could make a film that has more **credibility** than that."
To help you out, words in the positive lexicon are in **boldface** and those in the negative lexicon are in *italics*. Assume that we have the following features with their weights:

Features and their weights. NOTE: **bias** is given a value of **0.1**.

FeatureID	Feature	Weight
x1	Count of words in the positive lexicon of the document	2
x2	Count of words in the negative lexicon of the document	-4
x3	Count of "!" in the document	1
x4	Count of "?" in the document	0.5
x5	Count of sentences in the document	1.5
x6	Natural Logarithm of the Count of words in the document	1.25
bias	Classifier bias	1

Using the above learnt weights, **find out** whether the film is positive (y = 1) or negative (y = 0).

(CO2) [Application]