Roll No											

PRESIDENCY UNIVERSITY
BENGALURU

SCHOOL OF ENGINEERING
MID TERM EXAMINATION - OCT 2023

Semester : Semester VII - 2020

Course Code : CSE3014

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

Program : B.TECH

Date : 31-OCT-2023

Time : 9:30AM -11:00AM

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. State whether true or false. Antonymy is a word relationship in which a pair of words have low similarity because they are opposite in meaning.

(CO1) [Knowledge]

2. Mention the term which describes the number of documents in a corpus that a particular token is present in.

(CO1) [Knowledge]

3. Mention the activation function used for:

1. Binary logistic regression
2. Multinomial logistic regression

(CO1) [Knowledge]

4. State true or false. Accuracy for a classifier is evaluated on the testing dataset for the classifier.

(CO1) [Knowledge]

5. Recall that the formula for normalization of HISK is given by $\frac{HISK(x,y)}{\sqrt{Hisk(x,x) \times HISK(y,y)}}$.
Mention the range of values that the **normalized HISK** can take.

(CO1) [Knowledge]

PART B

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

6. 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) [Comprehension]

7. Match the entities in column A with those of Columns B and C

A Index	Column A	B Index	Column B	C Index	Column C
A	Sentiment Analysis	F	Syntactic Grammars	K	1954
B	Part-of-Speech Tagging	G	Document Classification	L	Colourless Green Ideas Sleep Furiously
C	Noam Chomsky	H	Machine Translation	M	Can Machines Think?
D	Alan Turing	I	Word Classification	N	Penn Treebank
E	Georgetown Experiment	J	Imitation Game	O	Polarity

NOTE: For your answers, you **ONLY NEED TO WRITE** the letters (Eg. AFK). No need to write all 3 entities of the group.

(CO2) [Comprehension]

PART C

ANSWER THE FOLLOWING QUESTION **(1 X 20 = 20M)**

8. 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 ;)	NEG

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]