IIITV

Mid-Semester Examination (Autumn 2022)

Degree Program : B.Tech (CS) (VII Sem)	**Student's name** :
Course Title : Natural Language Processing	**Roll No**:
Course Code : IT 437	**Time Duration** : 2:00
Date of Examination : 6 Sept, 2022.	**Total Mark** : 60

Instructions:

1. This exam consists of 5 questions. Before answering, ensure that all the questions are present.

Q.1 Consider the following vocabulary: **{BOS, EOS, here, David, are, you, the}** where BOS is the dummy token indicating the beginning of a sentence and EOS indicates end of a sentence. Note that we need never compute the (conditional) probability of BOS and so we should not include unigram or bigram counts of the BOS token. Consider the following training data:

BOS here you are EOS — ①
BOS here you are David EOS
~~BOS are you~~ here EOS
BOS you are here EOS
BOS you are here EOS
BOS ~~David~~ you ~~are here~~ EOS
BOS you are EOS 2

(a) Compute all n-gram counts up to $n = 2$. **2 marks**

(b) Calculate the following probabilities:

(i) p(you)

(ii) p(you | are) =
~~true~~ [illegible]

4 marks

(c) Using unigram and bigram language models, compute the probabilities of the following sentences

(i) BOS here you/are EOS

(ii) BOS are you EOS
What do you observe?

4 marks

(d) Apply Laplace smoothing with $\lambda = 1$ to the bigram model and compute the probabilites of the sentences from part (c) again. **10 marks**

Q.2 Consider the following word-context matrix with the three words "orange", "banana" and "car" and the three context words "juice", "the" and "drive".

	juice	the	drive
orange	10	20	0
banana	8	20	0
car	1	20	10

(a) Compute the MLEs using frequencies for the probabilities $P(w), P(c)$ and $P(w,c)$ for each word w and each context word c. **4 marks**

(b) Based on these, compute the PPMI values for the cells in the matrix. **4 marks**

(c) Now compute the cosine similarity values of the PPMI vectors for "orange" and "banana" and for "orange" and "car". **2 marks**

Q.3 Let $< X_1, X_2, X_3, X_4 > \; = \; < b, d, b, a >$ be the feature vector of an object to be classified. We use a multivariate Bernoulli Naive Bayes classifier, with three classes ($C = 0, C = 1, C = 3$), and the training data is given in the following table.

$\mathbf{X}_1$	$\mathbf{X}_2$	$\mathbf{X}_3$	$\mathbf{X}_4$	C
a	b	b	a	1
b	a	a	b	1
b	b	a	b	1
a	a	b	b	2
a	b	b	b	2
b	a	b	a	2
c	d	d	c	3
d	ç	c	d	3
d	d	c	d	3

$\frac{20}{60}$

What will be the decision of the classifier? Use Laplace estimates for $P(X_i \mid C)$. Each feature X_i has four possible values: a, b, c, d. **10 marks**

Q.4 Explain shortly what the n-gram assumption is. What are some advantages and disadvantages of having bigger or smaller n? **5 marks**

Q.5 Write short notes on the following **15 marks**

(a) Kneser-Ney Smoothing

(b) Perplexity

(c) bag of words and tf-idf

$\frac{8}{19+28}$

$\frac{8}{47}$ $\frac{20}{28}$