

Mid-Semester Examination (Autumn 2022)

Degree Program : B.Tech (CS) (VII Sem)

Course Title: Natural Language Processing

Course Code: IT 437

Date of Examination: 6 Sept, 2022.

Student's name:

Roll No:

Time Duration: 2:00

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:

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

2 marks

(b) Calculate the following probabilities:

(i) p(you)

(ii) p(you | are)

4 marks

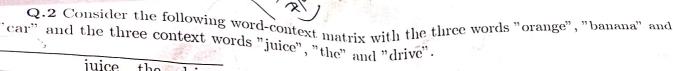
(c) Using unigram and bigram language models, compute the probabilities of the following sentences

(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 probabilities of the sentences from part (c) again. 10 marks



			35 309
	juice	the	drive
orange	10	20	0
banana	8	20	0
car	1.25	20	10

- (a) Compute the MLEs using frequencies for the probabilities P(w), P(c) and P(w,c) for each 4 marks word w and each context word c.
- 4 marks (b) Based on these, compute the PPMI values for the cells in the matrix.
- (c) Now compute the cosine similarity values of the PPMI vectors for "orange" and "banana" and for "orange" and "car".

Q.3 Let $\langle X_1, X_2, X_3, X_4 \rangle = \langle b, d, b, a \rangle$ 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.

X_1	X_2	X_3	\mathbf{X}_4	\mathbf{C}
a	b	b	a	1
b	1	-2	_b_	of C. et al
b	b	a	b	1
a	a	b	b	2
a	b	b	b	2
b	a	b	a	2
	$\frac{d}{d}$	d	С	3
C	$\frac{a}{c}$	$\frac{1}{c}$	d	3
d	$\frac{1}{d}$	c	d	3
d	u	ــــــــــــــــــــــــــــــــــــــ		

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.

- $\mathbf{Q.4}$ Explain shortly what the n-gram assumption is. What are some advantages and disadvantages of having bigger or smaller n?
 - Q.5 Write short notes on the following
- (a) Kneser-Ney Smoothing
- (b) Perploxity
- (c) bag of words and tf-idf

15 marks