Name: **Enrolment No:**	UPES UNIVERSITY OF TOMORROW

School of Business
UPES
End Semester Examination, May 2024

Program: MBA BA KPMG | **Semester: 2**
Subject/Course: Natural Language Processing | **Max. Marks: 100**
Course Code: DSBA7019_P | **Duration: 3 hours**

SECTION A
10Qx2M=20Marks

S. No.		Marks	CO
Q 1	Attempt all questions		
a.	Word segmentation is mostly used when: a) Hyphens are present b) Multiple alphabets are intermingled c) Long sentences d) No space between words	**2**	**CO1**
b.	Given in a corpusC2, the Maximum Likelihood Estimation for the bigram "dried berries" is 0.3 and the count of occurrence of the word "dried" is 580. For the same corpus C2, the likelihood of "dried berries" after applying add-one smoothening is 0.04. What is the vocabulary size of C2? a) 3585 b) 3795 c) 4955 d) 3995	**2**	**CO1**
c.	Which one of these is not an example of Neologism? a) Cryptocurrency b) Blogging c) Friendship d) Googling	**2**	**CO1**
d.	Ambiguity can occur in which of the following steps?	**2**	**CO1**

	a) Tokenization b) Language understanding c) Sentence Segmentation d) All of these		
e.	If first corpus has TTR=0.013 and second corpus has a TTR=0.13 then a) First corpus has more tendency to use different words b) Second corpus has more tendency to use different words c) Both a and b d) None of these	**2**	**CO2**
f.	Which of the following is/are true for English language? a) The output of Lemmatization and stemming for the same word might be different b) Output of lemmatization are always real words c) Output of stemming are always real words	**2**	**CO2**
g.	Which of the following are instances of stemming? a) Are -> be b) Plays -> play c) Saw -> s d) University -> univers	**2**	**CO2**
h.	As per the zipf's law, the correct statement about a corpus is: a) 10^{th} most common word will occur with 10 times the frequency of the 100^{th} most common word b) 100^{th} most common word will occur with 10 times the frequency of 10^{th} most common word c) Frequency of a word is directly proportional word to its position in the ranked list d) None of the above	**2**	**CO2**
i.	Which of the following does not require application of NLP algorithms? a) Classifying spam emails from good ones b) Classifying images of scanned documents as "hand written" or "printed" c) Automatically generating captions for the emails d) Building a sentiment analyzer for tweets on Twitter	**2**	**CO2**

j.	For the string 'mash', identify which of the following set of strings have a Levenshtein distance of 1. a. Smash, mas, lash, mushy, hash b. Bash, stash, lush, flash, dash c. Smash, mas, lash, mush, ash d. None of these	**2**	**CO1**
SECTION B **4Qx5M= 20 Marks**			
Q2.	Elaborate on Heap's law.	**5**	**CO1**
Q3.	Give five applications that can use NLP. What are the challenges faced during developing a NLP application?	**5**	**CO3**
Q4.	What is zipf's law? Why is it useful?	**5**	**CO2**
Q5.	What do you understand by "Tokenization", "Lemmatization and TTR?	**5**	**CO3**
SECTION-C **3Qx10M=30 Marks**			
Q6.	What are n-gram language models? Explain Maximum Likelihood Estimation for a bigram model.	**10**	**CO3**
Q7.	Explain Maximum Likelihood Estimation	**10**	**CO3**
Q8.	How do you evaluate different language models? What is perplexity?	**10**	**CO4**
SECTION-D **Attempt any two Questions** **2Qx15M= 30 Marks**			
Q9.	Calculate the edit distance between the words "Intention" and Execution" using Levenshtein distance algorithm	**15**	**CO4**
Q10.	Explain the Noisy Channel Model in detail	**15**	**CO4**
Q.11	Explain Shannon Visualization.	**15**	**CO4**