Name: Enrolment No:	UPES

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES
End Semester Examination, December 2018

Course: Natural Language Processing (CSEG-415) **Semester: 7th**
Programme: B.Tech.
Time: 03 hrs. **Max. Marks: 100**

SECTION A
All questions in SECTION A are compulsory

S. No.		Marks	CO
Q 1	What do you mean by Natural Language Processing? Mention some areas where NLP is applied.	**4**	**CO1**
Q 2	What are the challenges of NLP? What do you mean by supervised learning?	**4**	**CO1**
Q 3	What do you mean by function and content words? Give examples. If we read Tom Sawyer, who dominated as the most frequent words?	**4**	**CO3**
Q 4	If first corpus has TTR1 = 0.013 and second corpus has TTR2 = 0.13, where TTR1 and TTR2 represents type/token ratio in first and second corpus respectively; then what can you say about both of the corpus? Explain your suggestion.	**4**	**CO3**
Q 5	In the sentence, “In Dehradun I took my hat off. But I can’t put it back on.”; compute the total number of word tokens and word types. Bigram models are what ordered Markov Models?	**4**	**CO3**
SECTION B **(Q 6, 7, 8 are compulsory. Attempt Q9A or Q9B)**			
Q 6	Given the following sentences: "I want to eat. I want to sing. I eat Chinese." If you are following the bigram model; what is the probability of the following sentence: "I want to eat Chinese"? Also compute the probability of the following sentence: "I want to sing and eat "?	**10**	**CO3**
Q 7	In Vector Space Model, suppose we have two sentences bear the words; S1:<man, eat, eat>; S2:<man, eat, chicken, chicken>; S3:<man, eat, chicken>. Find the cosine and Jaccard similarity between S1 and S3.	**10**	**CO3**
Q 8	How is the sigmoid model related to probability? What is the range of the sigmoid function S(X)? Simulate the 'OR' function using a basic neural network without weights. What should be the threshold?	**10**	**CO2**
Q 9A	“I made her duck”. What are the possible interpretations that you can make out from the statement? If some indices are inserted in a max-heap. What is the complexity of finding the minimum element? Explain the Hidden Markov Model related to NLP with examples.	**10**	**CO3**
Q 9B	For text compression in NLP we use the Huffman coding technique. Given the following sentences: "I want to eat. I want to sing. I eat Chinese. He too want to eat Chinese. I want to sing and eat." Give the Huffman tree. Compute in ratio how much text was compressed using the technique.	**10**	**CO3**

SECTION-C **(Q 10 is compulsory. Attempt Q11A or Q11B)**			
Q 10	Consider the following productions: S ⟶ NP VP NP ⟶ NP PP NP ⟶ sushi NP ⟶ I NP ⟶ chopsticks NP ⟶ you VP ⟶ VP PP VP ⟶ Verb NP Verb ⟶ eat PP ⟶ Prep NP Prep ⟶ with Where; NP – noun phrase VP –verb phrase PP -preposition phrase. a) Use the CYK parsing algorithm to find if the sentence "I eat sushi with chopsticks with you" belongs to the above grammar. b) Explain the CYK algorithm.	**20**	**CO3**
Q 11A	Consider a simple three-state Markov model of the weather. Any given day, the weather can be described as being • State 1: precipitation (rain or snow) • State 2: cloudy • State 3: sunny Transitions between states are described by the transition matrix $A = \{a_{ij}\} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$ a) Draw the state transition graph. b) Given that the weather on day t=1 is sunny, what is the probability that the weather for the next 7 days will be "sun, sun, rain, rain, sun, clouds, sun"? c) What is the probability that the weather stays in the same known state Si for exactly T consecutive days?	**20**	**CO4**
Q 11B	We seek to classify documents as being about sports or not. Each document is associated with a pair (x, y), where x is a feature vector of word counts of the document and y is the label for whether it is about sports (y = 1 if yes, y = 0 if false). The vocabulary is size 3, so feature vectors look like (0, 1, 5), (1, 1, 1), etc. Consider a naive Bayes model with the following conditional probability table:	**20**	**CO4**

word type	1	2	2
$P(w \mid y = 1)$	$1/10$	$2/10$	$7/10$
$P(w \mid y = 0)$	$5/10$	$2/10$	$3/10$

and the following prior probabilities over classes:

$P(y = 1)$	$P(y = 0)$
$4/10$	$6/10$

Consider the document with counts x = (1, 0, 1).

a) Which class has highest posterior probability?

b) What is the posterior probability that the document is about sports?

CONFIDENTIAL

Name of Examination (Please tick, symbol is given)	:	**MID**		**END**	☒	**SUPPLE**	
Name of the School (Please tick, symbol is given)	:	**SOE**		**SOCS**	☒	**SOP**	
Programme	:	**B.Tech.**					
Semester	:	**7th semester**					
Name of the Course	:	**Natural Language Processing**					
Course Code	:	**CSEG-415**					
Name of Question Paper Setter	:	**Bikram Pratim Bhuyan**					
Employee Code	:	**40001825**					
Mobile & Extension	:	**9854350562**					
Note: Please mention additional Stationery to be provided, during examination such as Table/Graph Sheet etc. else mention "NOT APPLICABLE": NOT APPLICABLE							

FOR SRE DEPARTMENT		
Date of Examination	:	
Time of Examination	:	
No. of Copies (for Print)	:	

SET: 2

Name: **Enrolment No:**	UPES

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES
End Semester Examination, December 2018

Course: Natural Language Processing (CSEG-415) **Semester: 7th**
Programme: B.Tech.
Time: 03 hrs. **Max. Marks: 100**

SECTION A
All questions in SECTION A are compulsory

S. No.		Marks	CO
Q 1	Mention some areas where NLP is applied. What are the challenges of NLP?	**4**	**CO1**
Q 2	I made her duck. What are the possible interpretations that you can make out from the statement? If some indices are inserted in a max-heap. What is the complexity of finding the minimum element?	**4**	**CO1**
Q 3	What do you mean by function and content words? Give examples. If we read Tom Sawyer, who dominated as the most frequent words?	**4**	**CO3**
Q 4	If first corpus has TTR1 = 0.059 and second corpus has TTR2 = 0.59, where TTR1 and TTR2 represents type/token ratio in first and second corpus respectively; then what can you say about both of the corpus? Explain your suggestion.	**4**	**CO3**
Q 5	In the sentence, “India is my homeland. I will not give up on it.” compute the total number of word tokens and word types. Bigram models are what ordered Markov Models?	**4**	**CO3**

SECTION B
(Q 6, 7, 8 are compulsory. Attempt Q9A or Q9B)

Q 6	Given the following sentences: "I want to eat. I want to sing. I eat Chinese." If you are following the bigram model; what is the probability of the following sentence: "I want to eat Chinese"? Also compute the probability of the following sentence: "I want to sing and eat "?	**10**	**CO3**
Q 7	For text compression in NLP we use the Huffman coding technique. Given the following sentences: "I want to eat. I want to sing. I eat Chinese. He too want to eat Chinese. I want to sing and eat." Give the Huffman tree. Compute in ratio how much text was compressed using the technique.	**10**	**CO3**
Q 8	How is the sigmoid model related to probability? What is the range of the sigmoid function S(X)? Simulate the 'OR' function using a basic neural network without weights. What should be the threshold?	**10**	**CO2**
Q 9A	Explain Naive Bayes and Hidden Markov Models related to NLP.	**10**	**CO2**
Q 9B	What are the different types of learning? Explain each type with examples.	**10**	**CO2**

SECTION-C
(Q 10 is compulsory. Attempt Q11A or Q11B)

Q 10	Consider the following productions: S ⟶ NP VP	**20**	**CO3**

	NP ⟶ NP PP NP ⟶ sushi NP ⟶ I NP ⟶ chopsticks NP ⟶ you VP ⟶ VP PP VP ⟶ Verb NP Verb ⟶ eat PP ⟶ Prep NP Prep ⟶ with Where; NP – noun phrase VP –verb phrase PP -preposition phrase. c) Use the CYK parsing algorithm to find if the sentence "I eat sushi with chopsticks with you" belongs to the above grammar. d) Explain the CYK algorithm.		
Q 11A	Consider a simple three-state Markov model of the weather. Any given day, the weather can be described as being • State 1: precipitation (rain or snow) • State 2: cloudy • State 3: sunny Transitions between states are described by the transition matrix $A = \{a_{ij}\} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$ d) Draw the state transition graph. e) Given that the weather on day t=1 is sunny, what is the probability that the weather for the next 7 days will be "sun, sun, rain, rain, sun, clouds, sun"? f) What is the probability that the weather stays in the same known state Si for exactly T consecutive days?	**20**	**CO4**
Q 11B	We seek to classify documents as being about sports or not. Each document is associated with a pair (x, y), where x is a feature vector of word counts of the document and y is the label for whether it is about sports (y = 1 if yes, y = 0 if false). The vocabulary is size 3, so feature vectors look like (0, 1, 5), (1, 1, 1), etc. Consider a naive Bayes model with the following conditional probability table: word type | 1 | 2 | 2 $P(w \mid y = 1)$ | 1/10 | 2/10 | 7/10 $P(w \mid y = 0)$ | 5/10 | 2/10 | 3/10 and the following prior probabilities over classes: $P(y = 1)$ | $P(y = 0)$ 4/10 | 6/10 Consider the document with counts x = (1, 0, 1).	**20**	**CO4**

	c) Which class has highest posterior probability? d) What is the posterior probability that the document is about sports?		