Name: **Enrolment No:**	UPES UNIVERSITY WITH A PURPOSE

UNIVERSITY OF PETROLEUM AND ENERGY STUDIES
End Semester Examination, December 2020

Course: Computational Linguistics and Natural Language Processing **Semester: 5th**
Program: B.Tech CSE (AI & ML) **Time 03 hrs.**
Course Code: CSEG3024 **Max. Marks: 100**

Instructions: All questions are compulsory

SECTION A

S. No.		Marks	CO
Q 1	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.	**5**	**CO1**
Q 2	What is the range of the sigmoid function S(X)? If we simulate the 'OR' function using a basic neural network without weights. What should be the threshold?	**5**	**CO2**
Q 3	Mention some areas where NLP is applied.	**5**	**CO1**
Q 4	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.	**5**	**CO1**
Q 5	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?	**5**	**CO1**
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"?	**5**	**CO2**

SECTION B

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." Construct the Huffman tree. Compute in ratio how much text was compressed using the technique.	**10**	**CO1**

<table>
<tr>
<td>Q 8</td>
<td>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.</td>
<td>10</td>
<td>CO1</td>
</tr>
<tr>
<td>Q 9</td>
<td>Given the following table where the first column shows the rank ordering of documents in response to some query, and the second column indicates whether the document is relevant to the query. It is assumed that there are 6 relevant documents in the entire collection and those 10 documents have been retrieved and displayed.

Show the recall and precision values at each recall level.
<table>
<tr><td>Rank ordering</td><td>Relevant?</td><td>Recall</td><td>Precision</td></tr>
<tr><td>1</td><td>-</td><td></td><td></td></tr>
<tr><td>2</td><td>R</td><td></td><td></td></tr>
<tr><td>3</td><td>R</td><td></td><td></td></tr>
<tr><td>4</td><td>-</td><td></td><td></td></tr>
<tr><td>5</td><td>R</td><td></td><td></td></tr>
<tr><td>6</td><td>-</td><td></td><td></td></tr>
<tr><td>7</td><td>-</td><td></td><td></td></tr>
<tr><td>8</td><td>R</td><td></td><td></td></tr>
<tr><td>9</td><td>-</td><td></td><td></td></tr>
<tr><td>10</td><td>R</td><td></td><td></td></tr>
</table>
</td>
<td>10</td>
<td>CO2</td>
</tr>
<tr>
<td>Q 10</td>
<td>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</td>
<td>10</td>
<td>CO2</td>
</tr>
</table>

<table>
<tr><td></td><td>$$A = \{a_{ij}\} = \begin{bmatrix} 0.4 & 0.3 & 0.3 \\ 0.2 & 0.6 & 0.2 \\ 0.1 & 0.1 & 0.8 \end{bmatrix}$$
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”?</td><td></td><td></td></tr>
<tr><td>Q 11</td><td>Given the input context:
<table><tr><td></td><td>a</td><td>b</td><td>c</td><td>d</td></tr><tr><td>A</td><td>1</td><td>0</td><td>1</td><td>0</td></tr><tr><td>B</td><td>1</td><td>1</td><td>1</td><td>0</td></tr><tr><td>C</td><td>1</td><td>1</td><td>0</td><td>1</td></tr><tr><td>D</td><td>1</td><td>1</td><td>1</td><td>1</td></tr></table>Where {A,B,C,D} is the set of objects and {a,b,c,d} is the set of attributes.
Formulate some concepts and create an ontology of the concepts in a lattice diagram.</td><td>10</td><td>CO2</td></tr>
<tr><td colspan="4">SECTION-C</td></tr>
<tr><td>Q 12</td><td>1) Consider the following sentences, along with the particular class they belong to:
• Language independent system data driven dependency parsing- dependency parsing.
• Algorithm deterministic incremental dependency parsing- dependency parsing
• Transition based techniques projective dependency parsing- dependency parsing
• Structured models sentiment analysis- sentiment analysis

a) Build a text classifier using Naïve Bayes algorithm and using the classifier, classify the following sentence, “dependency sentiment analysis”
b) Suppose you run your algorithm for 10 documents and the actual and predicted classes by the classifier is shown in the table below. Construct the confusion matrix for your classifier where “SA” denotes sentiment analysis and “DP” denotes dependency parsing.</td><td>20</td><td>CO2</td></tr>
</table>

Doc	Actual	Predicted
1	DP	SA
2	DP	DP
3	SA	DP
4	DP	DP
5	SA	SA
6	SA	SA
7	DP	SA
8	SA	DP
9	DP	DP
10	SA	DP

OR

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).

a) Which class has highest posterior probability?
b) What is the posterior probability that the document is about sports?