Course Code : CST 408-1 **CXDW/RW – 18 / 5585**

Eighth Semester B. E. (Computer Science and Engineering) Examination

NATURAL LANGUAGE PROCESSING

Time : 3 Hours] [Max. Marks : 60

Instructions to Candidates :—

Assume suitable data and illustrate answers with neat sketches wherever necessary.

1. (a) Compute likelihood estimation of following sentences :—

<s> I am going to green city <e>

<s> It has good weather <e>

<s> Green city Clean city <e>

Compute any three bigram probabilities. 4 (CO 1)

(b) Write a python code to read a text file and remove stop words from the file. Use standard stop words dataset. 4 (CO 1)

(c) Ambiguity handling is main task in processing natural language, justify. 2 (CO 1)

2. (a) From the given matrix, find out any three bi–grams with maximum probability.

"Cricket is interesting and popular outdoor game".

	Cricket	Is	Interesting	And	Popular	Outdoor	Game
Cricket	0	650	0	0	0	0	0
Is	69	0	250	190	40	0	55
Interesting	250	345	0	730	40	0	0
And	302	20	120	0	340	30	220
Popular	504	220	50	7	0	123	0
Outdoor	205	20	0	0	0	0	350
Game	0	230	190	0	310	34	44

The unigram values from corpus :

	Cricket	Is	Interesting	And	Popular	Outdoor	Game
	1240	900	894	1220	1502	1608	950

Find the probability of sentence : (s1)

Interesting popular outdoor game

Perform Add – 1 smoothing and find the probability of sentence (s1). Assume V = 1400. 8 (CO 2)

(b) Low entropy is better for language processing. 2 (CO 2)

3. (a) Implement CYK Algorithm using following grammar and design derivation matrix for given string.

Grammar :

S $\longrightarrow$ NP VP | VP

VP $\longrightarrow$ V NP | V

NP $\longrightarrow$ NP NP | NP PP

NP $\longrightarrow$ N

PP $\longrightarrow$ P NP

N $\longrightarrow$ students | study | Algorithm | Language

V $\longrightarrow$ study

Sentence : Language students study algorithms. 7 (CO 3)

(b) Explain how regression is used in classification process, use suitable example. 3 (CO 3)

4. (a) Illustrate any two applications of Viterbi algorithm in NLP. 2 (CO 3)

(b) For the given set of tag values design best possible tag sequence for the phrase :

	Computer	Is	Fast
NN	60	25	15
VBZ	20	45	10
JJ	20	30	75

	NN	VBZ	JJ
NN	10	45	25
VBZ	50	25	25
JJ	40	30	50

P(start) = 1

Assume probability values between 0 – 100. 6 (CO 3)

(c) Compute precision and recall for L – PCFG given :

Gold standard constitutes : 9

Parser output : 5

Correct output : 6

Derive the generalized formulation. 2 (CO 3, CO 4)

5. (a) Write any two roles of WSD in MT. Explain Bootstrapping approach of WSD with suitable example. 4 (CO 4)

OR

Discuss selection restrictions and selection preferences with suitable example. 4 (CO 4)

(b) In Naïve based algorithm, explain the process of constructing "feature" array. Assume suitable data set and demonstrate the process. 6 (CO 4)

6. (a) Draw architectural component diagram of Text Summarizer and discuss the functionalities of each block. 5 (CO 5)

(b) What is phrased based machine translation ? How phrase size affects the accuracy of translation process, demonstrate with suitable example ? 5 (CO 5)

OR

(c) Device "argmax" equation for statistical machine translation system. Write about components of the "argmax" equation. 5 (CO 5)