

**Time Allotted : 3 hrs****Full Marks : 70***Figures out of the right margin indicate full marks.*

Candidates are required to answer Group A and any 5 (five) from Group B to E, taking at least one from each group.

Candidates are required to give answer in their own words as far as practicable.

Group – A
(Multiple Choice Type Questions)

1. Choose the correct alternative for the following: **10 × 1 = 10**
 - (i) One of the main challenges of Natural Language Processing is:
 - (a) ambiguity of sentences
 - (b) usage of non-standard English
 - (c) handling tokenization
 - (d) new senses of a word.
 - (ii) If first corpus has $TTR_1 = 0.013$ and second corpus has $TTR_2 = 0.13$ [where, TTR_1 and TTR_2 represent type/ token ratio in first and second corpus, respectively], then
 - (a) first corpus has more tendency to use the different words
 - (b) second corpus has more tendency to use the different words
 - (c) both (a) and (b)
 - (d) none of the above.
 - (iii) The grammar that consists rules with a single symbol on the left hand side of the rewrite rules is
 - (a) CFG
 - (b) CNG
 - (c) any context
 - (d) HMM.
 - (iv) Classifiers used for deciding whether a word is spelt correctly are
 - (a) decision tree
 - (b) support vector machine
 - (c) logistic regression
 - (d) all of these.
 - (v) To set up part of speech problem as a sequence labelling task we use
 - (a) argmax computation
 - (b) applying Bayes theorem
 - (c) Markov assumption
 - (d) all of these.
 - (vi) Viterbi algorithm is used in
 - (a) speech processing
 - (b) language processing
 - (c) speech and language processing
 - (d) none of these.
 - (vii) Consider a Markov chain $\{x_n; n=0, 1, \dots\}$, specified by the transition diagram shown in figure below.

Given that the chain starts with $x_0=1$, find the probability $x_2=2$.

- (a) 0.44 (b) 0.33 (c) 0.23 (d) 0.42.

- (viii) Word probability is calculated by
 - (a) likelihood probability
 - (b) prior probability
 - (c) Bayes rule
 - (d) none of these.
- (ix) In the sentence “In Delhi I took my hat off. But I can’t put it back on”, total number of word tokens and word types are
 - (a) 14, 13
 - (b) 13, 14
 - (c) 15, 14
 - (d) 14, 15.
- (x) Machine learning approaches to sense disambiguation make it possible
 - (a) to automatically create robust sense disambiguation systems
 - (b) to find ambiguity
 - (c) to apply Bays rule
 - (d) none of these.

Group – B

2. (a) Explain the stages of natural language processing.
- (b) Describe briefly the role of Finite State Transducers (FST) with suitable example.

6 + 6 = 12

3. (a) What is regular expression? Construct regular expressions for the following language (L) over the alphabet $\Sigma = \{0,1\}$
 $L = \{w \in \Sigma^* \mid \text{such that the second or third position from the end is a 1}\}$
- (b) A bank uses handwriting recognition software to read the amounts written as text on cheques and convert it to numerical amounts.
Example I: Rupees Forty Eight and paise Thirty Three Only $\rightarrow$ 48.33
Example II: Rupees Nine Only $\rightarrow$ 9.00
 Design a Finite State Transducer (FST) which does the above [Assume that (a) the FST gets a stream of tokens representing words/ punctuation. (b) the smallest amount on a cheque can be one paise and the largest amount can be ninety nine rupees and ninety nine paise].
 Carefully define your input and output alphabets and draw the FST as a state transition diagram labelling the arcs suitably.

(1 + 2) + 9 = 12

Group – C

4. (a) A Hidden Markov Model (HMM) includes states, observations, transition probabilities, observation and likelihoods. Describe what each of these would correspond to when using HMM for part of speech tagging.
- (b) Why is smoothing technique required in Natural Language Processing? How does Witten-Bell smoothing differ from add-one smoothing?

$$5 + (2 + 5) = 12$$

5. (a) Explain different models of computational phonology that use finite automata in various ways to realize phonological rules.
- (b) >>> saying
["Ravi", 'is', 'my', 'friend']
What is the output of the following Python code?
>>> [w for w in saying if len(w) < 4].

$$7 + 5 = 12$$

Group – D

6. (a) Give one example of each an English sentence with semantic ambiguity, syntactic ambiguity and anaphoric ambiguity.
- (b) In what way is the sentence *every man loves some woman* ambiguous? Explain using illustrative examples. Give logical formula in first order logic which clearly distinguishes between the two meanings.
- (c) Distinguish between homonymy and polysemy with suitable examples.

$$3 + (2 + 3) + (2 + 2) = 12$$

7. (a) Draw the parse tree of the sentence "Show me the meal on Flight 768 from New-York to California".
Consider the rules of the sentence as follows for top-down parsing:
S → VP
NP → DET N | NP PP
VP → V NP
PP → P NP
Explain top down parsing.
- (b) Write down the equation for trigram probability estimation.

$$(5 + 2) + 5 = 12$$

Group – E

8. (a) What is the need of discourse analysis in natural language processing?

- (b) Rhetorical Structure Theory (RST) is determined by two terminologies: Nucleus and Satellite. Explain these two terminologies with suitable example.
- (c) The table shows the actual probability distribution of a random variable X and two approximate distributions m_1 and m_2 .

X	x_1	x_2	x_3	x_4	x_5
p	0.3	0.2	0.1	0.2	0.2
m_1	0.2	0.2	0.2	0.2	0.2
m_2	0.3	0.1	0.1	0.1	0.4

Calculate the entropy of X and the cross-entropy for m_1 and m_2 . Determine which one of the two approximate distributions is better at approximating the true distribution p.

$$2 + 4 + 6 = 12$$

9. (a) Define predetermines, cardinal numbers, ordinal numbers and quantifiers with suitable examples.
- (b) Explain add-one smoothing or Witten-Bell smoothing.

$$6 + 6 = 12$$