

**NATURAL LANGUAGE PROCESSING
(CSEN 4231)**

Time Allotted : 2½ hrs

Full Marks : 60

Figures out of the right margin indicate full marks.

*Candidates are required to answer Group A and
any 4 (four) from Group B to E, taking one from each group.*

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

Group – A

1. Answer any twelve:

12 × 1 = 12

Choose the correct alternative for the following

- (i) Which of the following algorithms is primarily used for stemming?
(a) EM (b) CKY
(c) Porter (d) Chu-Liu-Edmonds
- (ii) Consider a corpus with N= 4000 word tokens. Let the number of word types be V =1000. Suppose the number of novel word types with zero occurrences is N₀ = 10. Then using Witten Bell Smoothing, what is the probability of each novel word after smoothing?
(a) 0.02 (b) 0.04
(c) 0.08 (d) 0.10
- (iii) The “m”-value measure in Porter’s algorithm for “Porter” is
(a) 0 (b) 1
(c) 2 (d) 3
- (iv) In HMM taggers extended to trigrams, we assume that
(a) The probability of a tag depends on two previous tags
(b) The probability of a tag depends on three previous tags
(c) The probability of a tag depends on two previous words
(d) The probability of a tag depends on three previous words
- (v) Let the number of edges in the shortest path in the thesaurus graph between sense nodes c1 and c2 be 4. Then pathlength based similarity $\text{simpath}(c1, c2)$ is
(a) 0.2 (b) 0.25
(c) 1/16 (d) 1/3
- (vi) In the HMM model, which of these probabilities add upto 1
(a) Emission probabilities for a single state
(b) Transition probabilities out of a state
(c) Transition probabilities into a single state
(d) Both (a) and (b)

- (vii) The complexity of the original Chu-Liu-Edmonds algorithm was:
 - (a) $O(n)$ (b) $O(n^2)$
 - (c) $O(n^3)$ (d) $O(n \log n)$
- (viii) Semantic Parsing helps in understanding which structure of a sentence?
 - (a) Syntactic (b) Semantic
 - (c) Pragmatic (d) Morphological
- (ix) ROUGE is a metric for evaluating
 - (a) document summaries (b) answers to questions
 - (c) query retrieval relevance (d) classification accuracy
- (x) Which of the following incorporates subword information?
 - (a) GloVe (b) Skip-Gram
 - (c) FastText (d) Word2Vec

Fill in the blanks with the correct word

- (xi) Labelling a word as noun/pronoun/verb is known as _____.
- (xii) For N tokens and T tags, the time complexity of the Viterbi algorithm is _____.
- (xiii) The stemming algorithm which operates by computing a measure “m” for a word is _____ algorithm.
- (xiv) The _____ architecture enables models to process and generate text using the attention mechanism.
- (xv) An example of a dense word embedding would be _____.

Group - B

2. (a) In Porter’s Algorithm, a word may be represented in a canonical form with a measure or “m” value based on consonant vowel sequences. Explain how this value is computed. [[C01](Remember/LOCQ)]
 (b) What is the value of m for the following words: RHYTHM, NATURAL, LANGUAGE, PROCESSING? [[C01](Apply/IOCQ)]
 (c) Explain the usage of Porter’s algorithm in NLP. [[C06](Remember/LOCQ)]

4 + 4 + 4 = 12
3. (a) In n-gram modelling, differentiate between backoff and interpolation based estimations. [[C01](Understand/LOCQ)]
 (b) In Spelling Correction using the noisy channel model of Kernighan et al., if the word $w = \text{actress}$ is misspelt as $x = \text{acress}$ and the count available from the confusion matrix $\text{del}[\]$ is $\text{del}[c,t] = 40$, $\text{count}("ct") = 100$, $\text{count}("c") = 200$, what is $P(x|w)$ estimated as? [[C03](Apply/IOCQ)]
 (c) Suppose we train a trigram language model with add-one smoothing on a given corpus. The corpus contains V word types. Express a formula for estimating $P(w_3|w_1, w_2)$, where w_3 is a word which follows the bigram (w_1, w_2) , in terms of various N-gram counts and V . Use the notation $c(w_1, w_2, w_3)$ to denote the

number of times that trigram (w_1, w_2, w_3) occurs in the corpus, and so on for bigrams and unigrams.

[[C03](Apply/IOCQ)]

$$4 + 4 + 4 = 12$$

Group - C

4. (a) In a text classification application to classify documents into positive (pos) and negative (neg) classes, if $P(\text{pos}) = 0.5$, $P(s | \text{pos}) = 0.9$, $P(s | \text{neg}) = 0.3$ calculate $P(\text{pos} | s)$ using Bayes' Theorem. [[C03](Apply/IOCQ)]
- (b) What are the two major simplifying assumptions made by Naïve Bayes classifiers and why are these assumptions made? [[C01](Remember/LOCQ)]
- (c) Consider applying Naïve Bayes to the problem of Word Sense Disambiguation.

Training	Doc	Words	Class
	d1	fish smoked fish	Fish
	d2	fish bass	Fish
	d3	fish haul smoked	Fish
	d4	guitar jazz bass	Music
Test	d5	bass guitar jazz jazz	?

Argue how the test document should be classified and hence determine the sense of the ambiguous word bass.

[[C03](Apply/IOCQ)]

$$4 + 4 + 4 = 12$$

5. (a) Consider a variant of the HMM model wherein the emission probabilities are dependent on the "from" state only as follows:

Output-> From State	a	b	c
A	0.6	0.1	0.3
B	0.1	0.7	0.2

The Transition probabilities are $P(A,A) = 0.7$, $P(A,B) = 0.3$, $P(B,A) = P(B,B) = 0.5$.

- (i) If the HMM always starts in state B, compute the probability of the output sequence (a, b). Show key calculations.
- (ii) Repeat calculations for the case wherein the HMM always starts in state A.

[[C03](Apply/IOCQ)]

- (b) Give examples of 4 POS tags for the word "back".

[[C04](Remember/LOCQ)]

$$(4 + 4) + 4 = 12$$

Group - D

6. (a) Consider the sentence: "Ram called Vinay from Delhi". Also consider the following PCFG:

$S \rightarrow NP VP (p_1)$

$NP \rightarrow NP PP (p_2)$

$VP \rightarrow V NP (p_3)$

$VP \rightarrow VP PP (p_4)$

$PP \rightarrow P NP (p_5)$

$NP \rightarrow \text{Ram} (p_6)$

$NP \rightarrow \text{Vinay} (p_7)$

NP → Delhi (p8)

V → called (p9)

P → from (p10)

(i) Draw two parse trees for the given sentence.

(ii) Derive the simplest expression in terms of the probabilities π_i 's to indicate which parse tree will be chosen and under what condition. [[C05](Apply/IOCQ)]

(b) What are some problems associated with PCFGs? [[C05](Remember/LOCQ)]

(4 + 4) + 4 = 12

7. (a) Explain with an example how cycles are eliminated in the Chu-Liu-Edmonds algorithm for dependency parsing. [[C04](Analyse/HOCQ)]

(b) Give the characterization of projectivity in a dependency tree with an example. [[C04](Remember/LOCQ)]

6 + 6 = 12

Group - E

8. (a) Explain briefly functions of 3 stages of a traditional factoid question answering system:

(i) question processing (ii) passage retrieval (iii) answer processing.

[[C01](Remember/LOCQ)]

(b) Explain briefly how the following steps are implemented in the question processing module of a question answering system?

(i) answer type recognition (ii) query reformulation.

[[C01](Remember/LOCQ)]

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

9. (a) Briefly describe the frequency-based approaches to word embeddings.

[[C01](Remember/LOCQ)]

(b) Explain how the GloVe word representation technique addresses the limitations of frequency based approaches and the Word2Vec technique. [[C01](Remember/LOCQ)]

(c) Enumerate the benefits of the FastText over other embedding techniques.

[[C01](Remember/LOCQ)]

4 + 4 + 4 = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	52.08	41.67	6.25