NATURAL LANGUAGE PROCESSING
(CSEN 4242)

Time Allotted: 2½ hrs **Full Marks: 60**

Figures out of the right margin indicate full marks.

Candidates are required to answer Group A and any4 (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) Methods of Dependency Parsing are
(a) Dynamic programming (like in the CKY algorithm) (b) Constraint Satisfaction
(c) (a), (b), Deterministic parsing (d) None of the above.

(ii) Suppose you are asked to write a regular expression for the pattern "either 'a' or '$'". Which of the following is the correct regular expression of the said pattern?
(a) /[a$]/ (b) /[a-$]/ (c) [a|$] (d) None of these

(iii) The process of removing and replacing suffixes to get to the root form of the word is
(a) Stemming (b) Tokenization
(c) Stop word removal (d) All of these

(iv) The morphological parsing of the word "geese" is
(a) goose+N+3SG (b) goose+V
(c) goose+N+PL (d) goose+N+SG

(v) Which of the following includes major tasks of NLP?
(a) Automatic Summarization (b) Discourse Analysis
(c) Machine Translation (d) All of the mentioned.

(vi) In Hidden Markov Model (HMM), observation likelihoods measure
(a) The likelihood of a POS tag given a word
(b) The likelihood of a POS tag given the preceding tag
(c) The likelihood of a word given a POS tag
(d) The likelihood of a POS tag given two preceding tags

(vii) Morphological parsing is
(a) The process of finding the constituent morphemes in a word
(b) The process of finding the constituent morphemes in a sentence
(c) The process of finding the constituent morphemes in a corpus
(d) All of the above.

(viii) Machine learning approaches to sense disambiguation make it possible
(a) to automatically create robust sense disambiguation
(b) to find ambiguity
(c) to apply bayes rule
(d) all of these.

(ix) Which of following doesn't require application of NLP?
(a) Spam emails classification
(b) Generating captions for images
(c) Sentiment analyser for tweets
(d) Image classification of scanned handwritten documents

(x) Word probability is calculated by
(a) likelihood probability (b) bayes rule
(c) joint probability (d) none of these.

Fill in the blanks with the correct word

(xi) Viterbi algorithm is used in_____.

(xii) Semantics is concerned with _____.

(xiii) _____ is an anchor that represents start of a line.

(xiv) _____ is a two – word sequence of words.

(xv) Preposition is an example of _____ class of Parts of Speech (POS).

Group - B

2. (a) Explain the Chomsky hierarchy of languages. *[(CSEN4242.1)(Explain/LOCQ)]*

(b) Write down the steps to convert CFG (Context Free Grammar) to CNF (Chomsky Normal Form).
Convert the following CFG into CNF

$S \rightarrow ASA \mid aB$
$A \rightarrow B \mid S$
$B \rightarrow b \mid \varepsilon$

[(CSEN4242.1)(Apply/IOCQ)]

4 + (4 + 4) = 12

3. (a) Design a CFG for the language $L = \{0^n1^n \mid n>=0\} \cup \{1^n0^n \mid n>=0\}$, where $\sum = \{0, 1\}$. *[(CSEN4242.1)(Apply/IOCQ)]*

(b) Compute the minimum edit distance (considering insertion cost 1, deletion cost 1, substitution cost 1) of "leda" to "deal". Show your work (using the edit distance grid). *[(CSEN4242.3)(Apply/IOCQ)]*

3 + 9 = 12

Group - C

4. (a) What do you mean by N – gram model? Give example. *[(CSEN4242.2)(Remember,Understand/LOCQ)]*

(b) Consider the following bigram and unigram probabilities of eight words in the Berkeley Restaurant Project corpus of 9332 sentences, where the size of the vocabulary is |V| = 1446:

	i	want	to	eat	chinese	food	lunch	spend
i	0.002	0.33	0	0.0036	0	0	0	0.00079
want	0.0022	0	0.66	0.0011	0.0065	0.0065	0.0054	0.0011
to	0.00083	0	0.0017	0.28	0.00083	0	0.0025	0.087
eat	0	0	0.0027	0	0.021	0.0027	0.056	0
chinese	0.0063	0	0	0	0	0.52	0.0063	0
food	0.014	0	0.014	0	0.00092	0.0037	0	0
lunch	0.0059	0	0	0	0	0.0029	0	0
spend	0.0036	0	0.0036	0	0	0	0	0

i	want	to	eat	chinese	food	lunch	spend
2533	927	2417	746	158	1093	341	278

Now, answer the following questions:

(i) Calculate the probability of the sentence **<s>i want chinese food</s>**, where P(i|<s>) = 0.25 and P(</s>|food) = 0.68.

(ii) Consider the add-1 smoothed bigram probabilities of the same eight words as follows:

	i	want	to	eat	chinese	food	lunch	spend
i	0.0015	0.21	0.00025	0.0025	0.00025	0.00025	0.00025	0.00075
want	0.0013	0.00042	0.26	0.00084	0.0029	0.0029	0.0025	0.00084
to	0.00078	0.00026	0.0013	0.18	0.00078	0.00026	0.0018	0.055
eat	0.00046	0.00046	0.0014	0.00046	0.0078	0.0014	0.02	0.00046
chinese	0.0012	0.00062	0.00062	0.00062	0.00062	0.052	0.0012	0.00062
food	0.0063	0.00039	0.0063	0.00039	0.00079	0.002	0.00039	0.00039
lunch	0.0017	0.00056	0.00056	0.00056	0.00056	0.0011	0.00056	0.00056
spend	0.0012	0.00058	0.0012	0.00058	0.00058	0.00058	0.00058	0.00058

Calculate the probability of the same sentence **<s>i want chinese food</s>**, where add – 1 smoothed probabilities are P(i|<s>) = 0.19 and P(</s>|food) = 0.40. *[(CSEN4242.2)(Apply/IOCQ)]*

(iii) Which of the two probabilities you computed in part (i) and (ii) is higher, unsmoothed or smoothed? Explain why. *[(CSEN4242.2)(Analyse/IOCQ)]*

(c) What is the role of smoothing in any language model? Define add – k smoothing method. *[(CSEN4242.3)(Understand/LOCQ)]*

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

5. (a) Consider the following corpus

<s> I am Sam </s>
<s> Sam I am </s>
<s> I am Sam </s>
<s> I do not like green eggs and Sam </s>

If we use linear interpolation smoothing between a maximum-likelihood bigram model and a maximum-likelihood unigram model with $\lambda1 = 0.5$ and $\lambda2 = 0.5$ (where $\lambda1$ and $\lambda2$ are the coefficient from each model), what is P(Sam|am)? Include <s> and </s> in your counts just like any other token. *[(CSEN4242.2)(Apply/IOCQ)]*

(b) Assume the following fragments comprise your Document Collection:
Doc1: banks/NNS to/TO raise/VB the/DT interest/NN rate/NN ./PUNCT
Doc2: jogging/VBG along/IN the/DT river/NN bank/NN ./PUNCT
Doc3: jogging/VBG to/TO the/DT bank/NN to/TO look/VB at/IN the/DT interest/NN rate/NN ./PUNCT
Predict Emission Probabilities and Transition Probabilities considering Bigram HMM? *[(CSEN4242.3)(Predict/HOCQ)]*

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

Group - D

6. (a) Define term frequency (TF) and inverse document frequency (IDF) of a word with the help of suitable example. *[(CSEN4242.4)(Understand/LOCQ)]*

(b) Consider the following three documents:

Document 1: It is going to rain today.
Document 2: Today I am not going outside.
Document 3: I am going to watch the season premiere.

Now, first compute TF-IDF matrix of all the words present in all the 3 documents mentioned above and then show that if a query like "It is rain outside" comes then Document 1 will be the most relevant for the query. *[(CSEN4242.4)(Apply/IOCQ)]*

(2 + 2) + (6 + 2) = 12

7. (a) What are the different metrics by which you can measure the similarity between two vectors? Explain any one of them with suitable example. *[(CSEN4242.3)(Understand/LOCQ)]*

(b) What is meant by 'word sense disambiguation'? *[(CSEN4242.4)(Understand/LOCQ)]*

(c) Explain the working principle of Skip-gram embeddings of Word2vec model. How does the learning take place in this model? *[(CSEN4242.3)(Apply/IOCQ)]*

3 + 2 + (4 + 3) = 12

Group - E

8. (a) Differentiate between Top-down and Bottom-up parsing. *[(CSEN4242.3)(Understand/LOCQ)]*

(b) Consider the following CFG and lexicon:

Grammar	Lexicon
S → NP VP .	*DT → the | that*
S → NP VP	*JJ → cold | empty | full*
NP → DT NN	*NN → sky | fire | light | flight | tomorrow*
NP → NN CC NN	*CC → and*
NP → DT JJ , JJ NN	*IN → of | at*
NP → NN	*CD → eleven*
VP → MD VP	*RB → a.m.*
VP → VBD ADJP	*VB → arrive*
VP → MD VP	*VBD → was | said*
VP → VB PP NP	*MD → should | would*
ADJP → JJ PP	
PP → IN NP	
PP → IN NP RB	

Draw the Top-down parse tree of the sentence "That cold, empty sky was full of fire and light." from the above grammar and lexicon. *[(CSEN4242.3)(Apply/IOCQ)]*

(c) Given the following short movie reviews, each labelled with a genre, either comedy or action marked in bold:

1. fun, couple, love, love **comedy**
2. fast, furious, shoot **action**
3. couple, fly, fast, fun, fun **comedy**
4. furious, shoot, shoot, fun **action**
5. fly, fast, shoot, love **action**

Compute the most likely class for a new document **D: fast, couple, shoot, fly** by using Naïve Bayes classifier and add-1 smoothing method. *[(CSEN4242.5)(Apply/IOCQ)]*

2 + 3 + 7 = 12

9. (a) Write the reasonable representation of the below phrase,

"A restaurant that serves Mexican food near HITK".

Also, explain the semantics of First Order Predicate Calculus of that sentence. *[(CSEN4242.4)(Explain/IOCQ)]*

(b) (i) What are the different types of lexical items?

(ii) Briefly outline selectional association-based word sense disambiguation algorithm. *[(CSEN4242.5)(Define/LOCQ)]*

6 + (3 + 3) = 12

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	26.04	67.71	6.25

Course Outcome (CO):

After the completion of the course students will be able to

CSEN4242.1. Learn various models, methods, and algorithms of Natural Language Processing (NLP) to build an automation tool to solve speech recognition, machine translation, spam filtering, text classification, spell checking etc.
CSEN4242.2. Understand and estimate the parameters of the probabilistic models.
CSEN4242.3. Identify problems solvable in language automation environments and also identify certain tasks that are challenging to carry out with traditionally existing statistical models.
CSEN4242.4. Understand the linguistic phenomena and will explore the linguistic features relevant to each NLP task.
CSEN4242.5. Identify the opportunities for research await and prepare to conduct research in NLP or related fields.

**LOCQ: Lower Order Cognitive Question; IOCQ: Intermediate Order Cognitive Question; HOCQ: Higher Order Cognitive Question.*