

**NATURAL LANGUAGE PROCESSING
(CSEN 4242)**

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) The grammar that consists rules with a single symbol on the left-hand side of the rewrite rules is
(a) CFG (b) CSG (c) Naive Bayes (d) HMM.
- (ii) One of the main challenge/s of NLP Is
(a) Handling Ambiguity of Sentences (b) Handling Tokenization
(c) Handling POS-Tagging (d) all of (a), (b) & (c).
- (iii) In linguistic morphology, _____ is the process for reducing inflected words to their root form.
(a) Rooting (b) Stemming (c) Text-Proofing (d) Both (a) & (b)
- (iv) Phonology is study of organizing
(a) sound systematically (b) signal processing
(c) speech and language processing (d) all of the above.
- (v) For a document d and a class c, Bayes' Rule Applied to Documents and Classes is given by
(a) $p(c|d) = \frac{p(d|c)p(c)}{p(d)}$ (b) $p(c|d) = \frac{p(d|c)p(d)}{p(d)}$
(c) $p(c|d) = \frac{p(d|c)p(c)}{p(d)p(c)}$ (d) $p(c|d) = p(d|c)p(d)$
- (vi) Viterbi algorithm is used in
(a) speech processing (b) language processing
(c) speech and language processing (d) none of (a), (b) & (c).
- (vii) Many words have more than one meaning. We have to select the meaning which makes the most sense in context. This can be resolved by
(a) Fuzzy Logic (b) Word Sense Disambiguation
(c) Shallow Symantic Analysis (d) all of (a), (b) & (c).

- (viii) Semantics is concerned with _____.
(a) measure the information (b) meaning of words
(c) correct the information (d) handle the noise
- (ix) Word probability is calculated by
(a) likelihood probability (b) bayes rule
(c) joint probability (d) none of (a), (b) & (c).
- (x) Machine learning approaches to sense disambiguation make it possible
(a) to automatically create robust sense disambiguation systems
(b) find ambiguity
(c) apply Baye's Rule
(d) none of (a), (b) & (c).

Group - B

2. (a) Explain the Chomsky hierarchy of languages. Write down the steps to convert CFG (Context Free Grammar) to CNF (Chomsky Normal Form).
[[CSEN3111.1](Understand/LOCQ)]
- (b) What is Regular Expression?
Construct the parse tree from the following grammar,
S → NP VP
PP → P NP
NP → Det N | Det N PP | 'I'
VP → V NP | VP PP
Det → 'a' | 'my'
N → 'tiger' | 'garden'
V → 'shot'
P → 'in'
[[CSEN3111.1](Understand/IOCQ)]
(3 + 3) + (1 + 5) = 12
3. (a) Describe the class of strings matched by the following regular expression:
(i) [a-zA-Z]⁺
(ii) [A-Z][a-z]^{*}.
[[CSEN3111.2](Understand/LOCQ)]
- (b) Convert the following CFG into CNF
S → ASA | aB
A → B | S
B → b | ε
[[CSEN3111.1](Understand/IOCQ)]
- (c) What is Finite State Transducer? Explain probability language modelling with examples.
[[CO4](Remeber/LOCQ)]
(2 + 2) + 3 + (2 + 3) = 12

Group - C

4. (a) Explain the use of **smoothing** in language models using an example. Continuing with the previous example illustrate the difference between **Laplace Smoothing** and **Add-One Smoothing**.
[[CO2](Explain/LOCQ)](Contrast/IOCQ)]
- (b) 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>

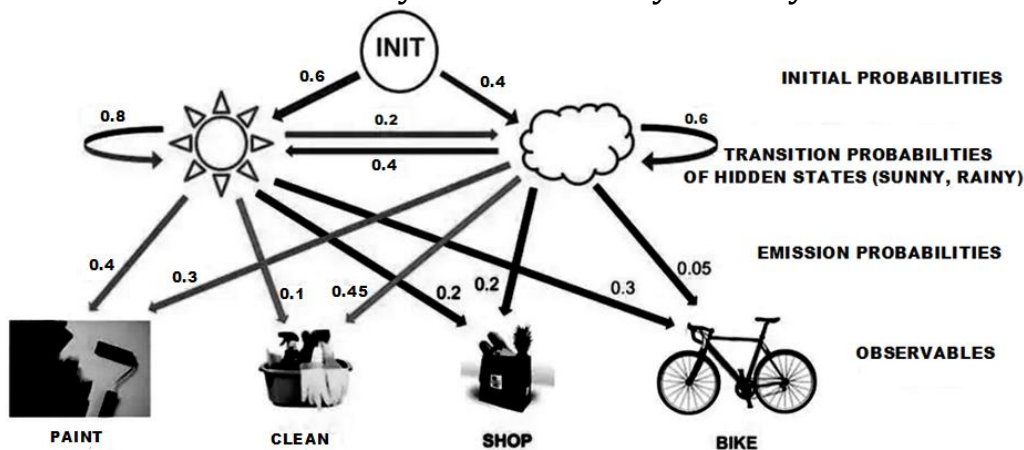
If we use *linear interpolation smoothing* between a *maximum-likelihood bigram model* and a *maximum-likelihood unigram model* with $\lambda_1 = 0.5$ and $\lambda_2 = 0.5$, what is $P(\text{Sam}|\text{am})$? Include <s> and </s> in your counts just like any other token.

$$[[CO2](Solve/IOCC)] \\ (3 + 3) + 6 = 12$$

5. (a) Define *Hidden Markov Model (HMM)* with the help of a suitable example. How will you compute the accuracy of that model? $[[CO2](Define/LOCQ)](Compute/IOCC)]$
- (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*. $[[CO2](Predict/HOCQ)]$
- $(3 + 3) + (3 + 3) = 12$

Group - D

6. (a) Differentiate between *Synonymy*, *Homonymy*, and *Polysemy*. Give examples of each. $[[CO4](Recall/LOCQ)]$
- (b) Consider the sentence: **The fan waved at the team with a flag.**
 Provide **POS** tags to the above sentence.
 Now, construct *two different parse trees* to justify why the sentence is structurally ambiguous. $[[CO4](Recall/LOCQ)](Construct/HOCQ)]$
- (c) What is a *gloss*? State at least two glosses for each of the following words: *Bank*, and *Right*. $[[CO4](State/LOCQ)]$
- $4 + (2 + 4) + 2 = 12$
7. (a) Lea is our imaginary friend and let's suppose that in four days she does the following: painting, cleaning, shopping, biking. By observing Lea's activity, we want to know whether the day has been sunny or rainy.



In this case the likelihood that it is Sunny on the first day is 0.6, while the likelihood that it is Rainy is 0.4.

$$\Pi = [0.6, 0.4]$$

Identify the following problems from the above diagrams,

- (i) Emission Matrix
- (ii) Observable Matrix
- (iii) Identify the output (weather condition) from the observable sequence (Paint, Clean, Shop and Bike) using HMM. [[CSEN3111.5](Analyse/HOCQ)]
- (b) Sketch 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:
 $S \rightarrow VP$
 $NP \rightarrow DET\ N|NP\ PP$
 $VP \rightarrow V\ NP$
 $PP \rightarrow P\ NP$

$$[[CO4](Sketch/IOCQ)]$$

$$(1 + 1 + 6) + 4 = 12$$

Group - E

- 8. (a) Random sequence of letters with probabilities are $P=1/8$, $t=1/4$, $k=1/8$, $a=1/4$, $i=1/8$, $u=1/8$ Calculate the entropy of each letter. [[CSEN3111.4](Understand/IOCQ)]
- (b) What is the need of sentiment analysis? Briefly give the steps of sentiment analysis. [[CSEN3111.2](Remember/LOCQ)]
- (c) Write down the importance of WordNet. Briefly explain hypernyms and hyponyms. [[CSEN3111.2](Remember/LOCQ)]
 $3 + 5 + 4 = 12$
- 9. (a) Analyze the usage of **Text Coherence** and **Discourse Structure** with the help of examples. [[CO4](Analyze IOCQ)]
- (b) **Rhetorical Structure Theory (RST)** is determined by two terminologies: **Nucleus** and **Satellite**.
Outline and contrast these two terminologies with suitable examples. [[CO4](Outline/Contrast/IOCQ)]
- (c) Illustrate the usage of the **Pronoun Resolution Algorithm**. [[CO4](Illustrate/IOCQ)]
 $4 + 4 + 4 = 12$

Cognition Level	LOCQ	IOCQ	HOCQ
Percentage distribution	39.58	41.67	18.75

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 like, 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.