CBCS SCHEME

USN | | | | | | | | | | |

18CS743

Seventh Semester B.E. Degree Examination, Dec.2023/Jan.2024
Natural Language Processing

Time: 3 hrs. Max. Marks: 100

Note: ***Answer any FIVE full questions, choosing ONE full question from each module.***

Module-1

1 a. What is NLP? Explain the challenges of NLP. **(08 Marks)**

b. Explain different levels of language processing. **(08 Marks)**

c. Differentiate between the rationalist and empiricist approaches to NLP. **(04 Marks)**

OR

2 a. Explain n-gram model. How to handle data sparseness problem in n-gram model. **(10 Marks)**

b. Represent the following sentence in GB (S-structure and d-structure) **(06 Marks)**

c. With example, explain the important features of Indian Languages. **(04 Marks)**

Module-2

3 a. What is Morphology? Explain two step morphological parser. **(06 Marks)**

b. Write and explain minimum edit distance algorithm. Compute minimum edit distance between tutor and tumour. **(10 Marks)**

c. What is POS (Part of Speech) tagging? List POS tagging methods. **(04 Marks)**

OR

4 a. Discuss the disadvantages of the logic top-down parser with the help of an appropriate example. **(10 Marks)**

b. Write a note on :

i) CFG (Context-Free Grammar) for natural language

ii) Lexicalization **(10 Marks)**

Module-3

5 a. Explain the shortest path hypothisis with example. **(10 Marks)**

b. Explain how to capture relation pattern with a string Kernel. **(10 Marks)**

OR

6 a. With diagram explain learning framework architecture. **(10 Marks)**

b. Explain the strategies used in active learning approach. **(10 Marks)**

Module-4

7 a. Explain with neat diagram evolutionary nodel for KDT (Knowledge Discovery from Text) **(10 Marks)**

b. Define :

i) Cohesion

ii) Interestingness

iii) Coherence

iv) Coverage

v) Plausibility of origin. **(10 Marks)**

OR

8 a. Explain SVM learning method in sequential model estimation. **(10 Marks)**

b. Write a note on various approaches to analyzing texts. **(10 Marks)**

Module-5

9 a. What are the benefits of eliminating stop words? Give examples in which stop word elimination may be harmful. **(06 Marks)**

b. What is IR (information system)? Explain design features of IR with neat diagram. **(10 Marks)**

c. State and explain Zipf's law. **(04 Marks)**

OR

10 Write a short note on :

i) POS tagger

ii) WORDNET

iii) FRAMENET

iv) STEMMERS **(20 Marks)**

* * * * *

Visvesvaraya Technological University

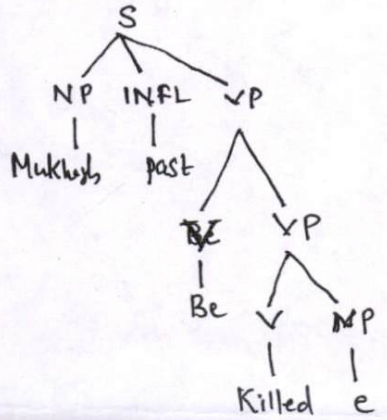
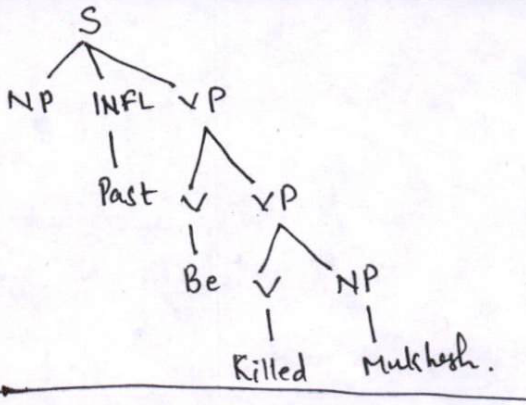
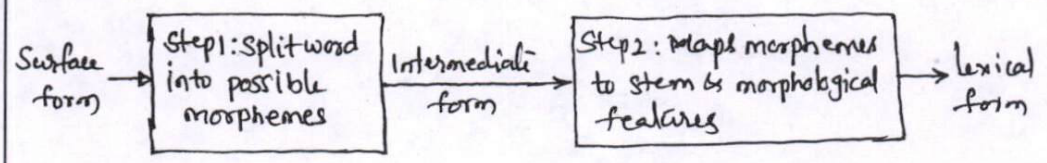
Belagavi, Karnataka - 590 018.

Scheme & Solutions Signature of Scrutinizer

Subject Title : Natural Language Processing **Subject Code :** 18CS743

Question Number	Solution	Marks Allocated
1 a.	Definition: Concerned with the development of computational models of aspects of human language processing challanges in * representation & interpretation. * identifying semantics. * frequency of words * Idioms, Metaphor & ellipses * Quantifier Scoping * Ambiguity .... } Explanation	2 Marks 6 Marks
1 b.	Levels of language processing: * lexical analysis. * Syntactic analysis. * Semantic analysis. * discourse analysis * pragmatic analysis. (with explanation)	8 Marks
1 c	Rationalist approach: which assumes the existance of some language faculty in the human brain. (2M) Empiricist: do not believe in existance of a language faculty Instead they believe in the existance of some general organization Principles. (2M)	4 Marks.
2 a.	Explanation of n-gram (Probability estimation) $P(w_i/h_i) = P(w_i/w_{i-n+1} \dots w_{i-1})$ $P(w_i/w_{i-n+1}, \dots w_{i-1}) = \frac{C(w_{i-n+1}, \dots w_{i-1}, w_i)}{C(w_{i-n+1}, \dots w_{i-1})}$ - Techniques to handle data Sparseness Problem * Add-one Smoothing * Good-Turing Smoothing } along with explanation. * Caching Technique	5 marks. 5 Marks

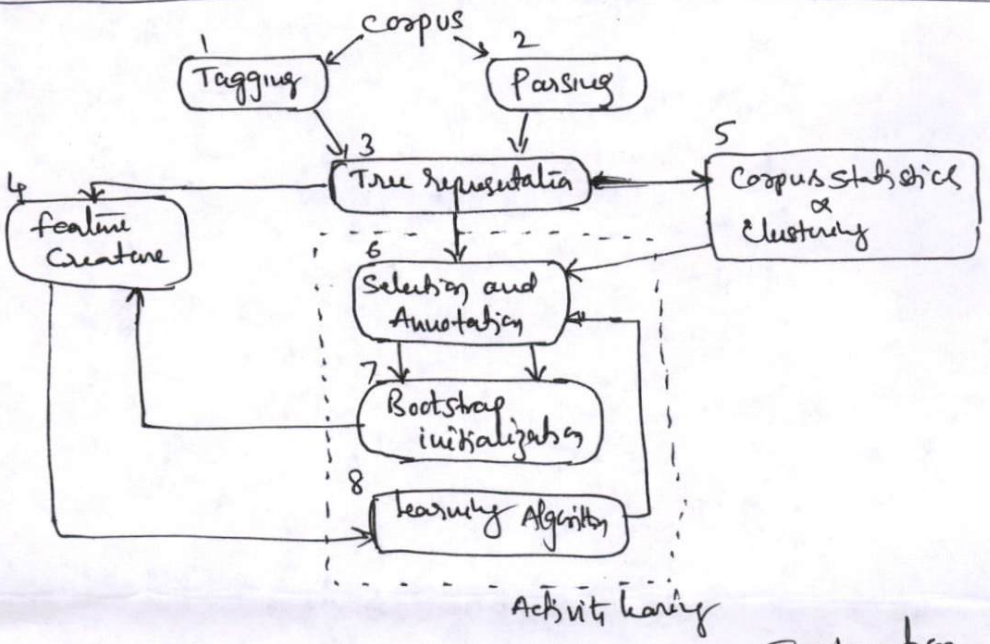
Subject Title : Natural Language Processing Subject Code : 18CS743.

Question Number	Solution	Marks Allocated
2 b	S- Structure representation :  D- Structure  	3 Marks 3 Marks
2C	* Morphologically rich languages. * flexibility allow word groups in any order * Nouns are followed by post positions * Verbs are formed differently in Indo-Aryan & Dravidian languages. Explanation with example	4 Marks
3 a	Morphology: It studies word structure & the formation of words from smaller units (morphemes).  Explanation	2 marks 2 marks 2 marks

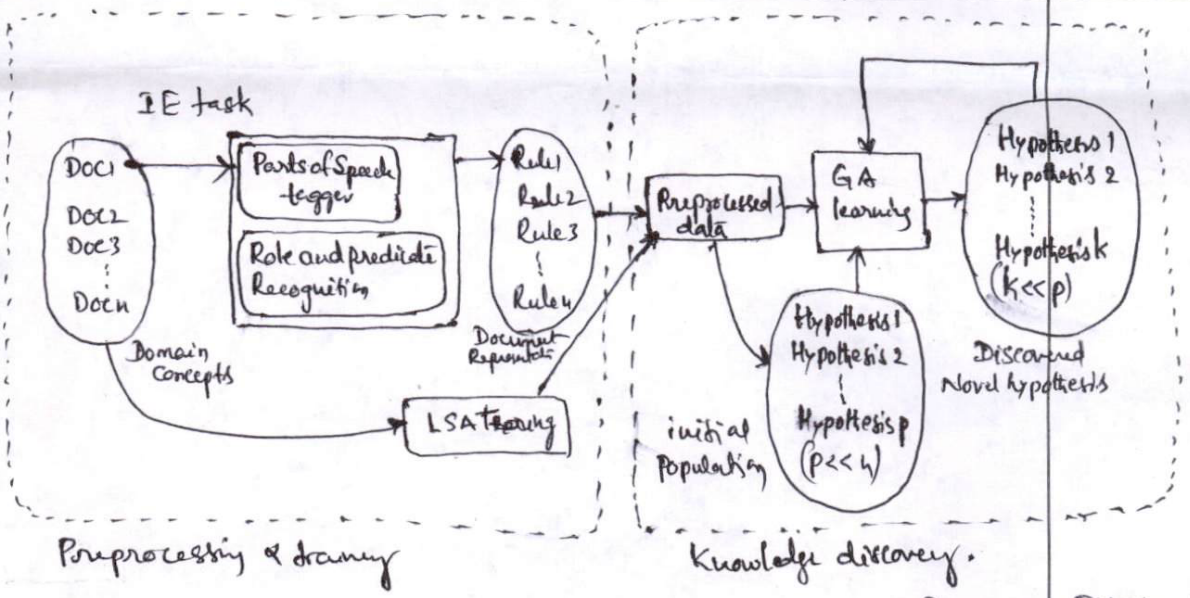
Subject Title : Natural Language Processing **Subject Code :** 18CS743

Question Number	Solution	Marks Allocated
3 b	Algorithm: Input : two strings X and Y Output : The minimum distance b/w X & Y m ← length(X) n ← length(Y) for i=0 to m do dist[i,0] ← i for j=0 to n do dist[0,j] ← j for i=0 to m do for j=0 to n do dist[i,j] = min{ dist[i-1,j] + insert_cost, dist[i-1,j-1] + subst_cost(x_i, y_j), dist[i,j-1] + delet_cost } Explanation ——	4 Marks 3 Marks.
	Computing Minimum distance.	3 Marks
3 c	POS tagging: Process of assigning a part of Speech such as a noun, verb, Pronoun, Preposition, adverb & adjective to each word in a sentence. POS tagging Methods: * Rule-based (Linguistic), * Stochastic (data driven) * Hybrid	2 Marks. 2 Marks.
4 a	Disadvantages: * left recursion form stuck in an infinite loop * Structural ambiguity. * Coordination ambiguity, * local ambiguity, * repeated parsing Explanation — 4 Marks Example — 3 Marks	3 Marks
4 b	Explanation (i) CFG. —— (ii) Lexicalization ——	5 Marks 5 Marks

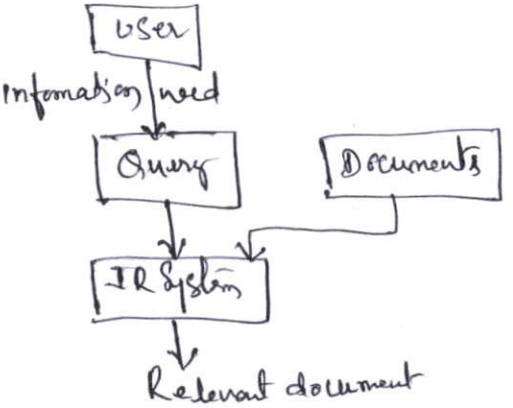
Subject Title : Natural Language Processing Subject Code : 18CS743

Question Number	Solution	Marks Allocated
5 a	If e_1 & e_2 are two entities mentioned in the same sentence such that they are observed to be in a relation R, then the contribution of the sentence dependency graph to establishing the relationship $R(e_1, e_2)$ is almost exclusively concentrated in the shortest path between e_1 & e_2 in the undirected version of the dependency graph Explanation of the above concept example ———	4 Marks 3 Marks 3 Marks
5 b.	Explanation of along with the patterns: * [FB] Focus-Before . * [B] Between * [BA] Between-After * [M] Modifier	6+4 Marks.
6 a	 Explanation	5 Marks 5 Marks.
6 b.	Strategies: (a) Divide the corpus in clusters of sentences with the same target verb. If a cluster has fewer sentences than a given threshold, group sentences with verbs evoking the same frame into the same cluster. (b) within each cluster, group the sentences with the same parse subtree together. (c) Select sentences from the largest groups of the largest clusters and present them to the user for annotation. (d) Bootstrap initialization: Apply the labels assigned by the users to groups of sentences with the same parse subtree.	

Subject Title : Natural Language Processing Subject Code : 18CS743

Question Number	Solution	Marks Allocated
	(e) Train all the classifiers of the committee on the labeled instances; apply each trained classifier to the Unlabelled sentences. (f) Get a pool of instances where the classifier of the committee disagree & present to the user the instances belonging to sentences from the next largest clusters not yet manually labelled. (g) Repeat steps d-f a few times until a desired accuracy of classification is achieved.	4 Marks
7 a	Explanation  Eg — 5 Marks Explanation — 5 Marks.	3 Marks. 10 Marks
7 b	(i) $\text{cohesion}(H) = \sum_{r_i, P_i \in H} \frac{\text{Prob}(P_i / r_i)}{|H|}$ $P_i \rightarrow$ predicate, $r_i \rightarrow$ rhetorical role (ii) Interestingness (H) = < Semantic dissimilarity b/w Antecedent and Consequent >. (iii) $\text{Coherence}(H) = \sum_{i=1}^{|H|-1} \frac{\text{SemSim}(P_i(A_i), P_{i+1}(A_{i+1}))}{(|H|-1)}$ SemSim → Semantic Similarity, $(|H|-1)$ → adjacent pair (iv) $\text{Coverage}(H) = \frac{|\text{Rules Covered}(H)|}{|\text{Rule Set}|}$ (v) $\text{plausibility}(H) = \begin{cases} S_p & \text{if H was created from a Swanson's Crossover} \\ 0 & \text{if H is the original population or is a result of another operation.} \end{cases}$	2 x 5 = 10 Marks

Subject Title : Natural Language Processing Subject Code : 18CS748.

Question Number	Solution	Marks Allocated
8 a	Explanation of how SVM (Support Vector machine) learning Method used in Sequence model estimation.	10 Marks
8 b	Explanation of various approaches to analysing texts * Bag-of-words-based approaches. — 5 Marks * Highlevel Representation approaches — 5 Marks.	
9 a	Benifits of eliminating stopwords — 2 Marks drawback of eliminating stop words with ex — 4 Marks	
9 b	Definition of IR: deals with the organization, storage, retrieval, and evaluation of information relevant to a user's query	2 Marks
		2 Marks
	Explanation of design features: indexing, Stopword elimination, Stemming	6 Marks.
9 c	Zipf's law statement and explanation 2+2	4 Marks.
10	(i) Pos tagger — (5 Marks) (ii) WORDNET — (5 Marks) (iii) FRAMENET — (5 Marks) (iv) STEMMER — (5 Marks)	20 Marks.