U.S.N.										

B.M.S. College of Engineering, Bengaluru-560019

Autonomous Institute Affiliated to VTU

June 2025 Semester End Main Examinations

Programme: B.E. **Semester: VI**

Branch: Artificial Intelligence and Machine Learning **Duration: 3 hrs.**

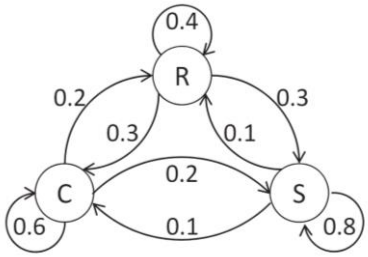
Course Code: 24AM6PENLP **Max Marks: 100**

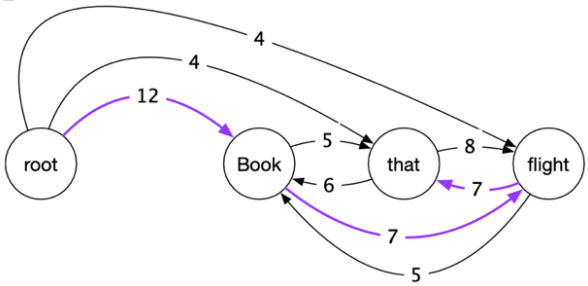
Course: Natural Language Processing

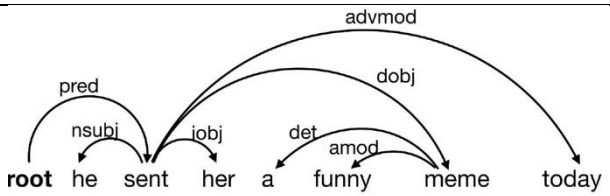
Instructions: 1. Answer any FIVE full questions, choosing one full question from each unit.
2. Missing data, if any, may be suitably assumed.

Important Note: Completing your answers, compulsorily draw diagonal cross lines on the remaining blank pages. Revealing of identification, appeal to evaluator will be treated as malpractice.

			UNIT - I	*CO*	*PO*	Marks
1	a)		Outline different phases of Natural Language Processing (NLP) with a neat diagram by giving suitable example at each phase.	*CO1*	*PO1*	**8**
	b)		Justify the role of regular expressions in NLP. Provide regular expressions for the following: i. Email address ii. Date in the format DD/MM/YYYY with a year between 1900 and 2099	*CO2*	*PO2*	**6**
	c)		Identify the type of ambiguity and all possible meanings arise from each of the sentences given. i. Time flies like an arrow. ii. Visiting relatives can be boring. iii. Are you still watching? iv. The teacher asked the student to bring his book to class. v. She observed the painting with a sharp eye.	*CO2*	*PO1*	**6**
2			**OR**			
	a)		Apply Byte Pair Encoding for the following words with their frequencies. Write algorithm for the same. ("hug", 10), ("pug", 5), ("pun", 12), ("bun", 4), ("hugs", 5).	*CO2*	*PO1*	**10**
	b)		Analyze the difference between tokenization, lemmatization and stemming with an example for each.	*CO2*	*PO2*	**5**
	c)		Alice is tasked with developing NLP system for analyzing customer reviews. Describe the challenges he might encounter in this project, providing an example for each.	*CO1*	*PO1*	**5**
			UNIT - II			
3	a)		i. A statistical language model is a probability distribution P(s) over all possible word sequences. Justify the statement using a language model. ii. Consider the following Corpus of three sentences: <s>There is a big garden. </s>	*CO2*	*PO2*	**10**

			<s> Children play in a garden. </s> <s> They play inside beautiful garden. </s> Assume a bi-gram language model and calculate the probability for the sentence "<s>They play in a big Garden</s>".			
		b)	Once a day (e.g. at noon), the weather is observed as one of states 1: rainy(R) 2: cloudy(C) 3: sunny (S). Answer the following question based on the given state transition diagram:  Given that the weather on day 1 (t = 1) is sunny (state 3), what is the probability that the weather for the next 7 days will be "sun- sun – rain – rain – sun – cloudy - sun"?	*CO2*	*PO2*	**5**
		c)	Assume a part-of-speech tagging system using the Viterbi algorithm for analyzing medical records. Review the possible methods to handle an unseen medical term by the training data while testing, to make sure non-zero probabilities.	*CO2*	*PO1*	**5**
			OR			
	4	a)	Consider a Hidden Markov Model (HMM) with two states (Rainy, Sunny) and three observations (Walk, Shop, Clean). The following matrices define the model: • **States**: S={Rainy,Sunny} • **Observations**: O={Walk,Shop,Clean} • **Initial Probabilities (π): π= (0.6, 0.4)** • **Transition Probabilities (A)**: $\begin{pmatrix} 0.7 & 0.3 \\ 0.4 & 0.6 \end{pmatrix}$ • **Emission Probabilities (B)**: $\begin{pmatrix} 0.1 & 0.4 & 0.5 \\ 0.6 & 0.3 & 0.1 \end{pmatrix}$ Given the observation sequence (Walk, Shop, Clean), compute the most probable state sequence.	*CO2*	*PO1*	**6**
		b)	Apply the knowledge of Conditional Random Fields (CRF) to answer the following questions: i. linear chain CRF with suitable mathematical equations. ii. Inference and Training for CRFs	*CO1*	*PO1*	**8**

		c)	Given a corpus with the following sentences: I like NLP I like machine learning Machine learning is fun i. Calculate the bigram probabilities. ii. Given the previous word "like", predict the next word using the bigram probabilities.	*CO2*	*PO2*	**6**
			UNIT - III			
	5	a)	Parse the below sentences using dependency parsing: 1. I saw the queen England's hat. 2. The angry squirrel stole the mixed nuts from wallmart. 3. The cat eats tasty fish 4. Turn on the office lights.	*CO2*	*PO2*	**8**
		b)	Parse the sentence "Bart watched a squirrel with binoculars. "and check whether the given grammar is ambiguous. S-> NP VP VP-> VP PP VP-> V NP NP-> DET N NP-> NP PP PP -> P NP NP-> N N-> Bart | Squirrel |Binoculars V-> Watched DET-> a P-> with	*CO2*	*PO3*	**5**
		c)	Apply Chu-Liu Edmonds algorithm for finding a maximum spanning tree to the given weighted directed graph. Write the algorithm for the same. 	*CO2*	*PO3*	**7**
			OR			
	6	a)	Illustrate the working of shift reduce parser with a neat diagram. Apply shift reduce parsing for the below dependency graph.	*CO2*	*PO3*	**10**

						
		b)	Check whether the below context-free grammar is in Chomsky Normal Form. S → NP VP NP → DT NN NP → NN NP → NN NNS VP → VBP NP VP → VBP VP → VP PP PP → IN NP DT → a | an NN → time | fruit | arrow | banana NNS → flies VBP → flies | like IN → like Parse the sentence "Fruit flies like a banana" to predict whether the given grammar is ambiguous using dynamic programming approach.	*CO2*	*PO3*	**10**
			UNIT - IV			
	7	a)	i. Illustrate the working and applications of TF-IDF with suitable mathematical equations. ii. Calculate TF-IDF scores for all the words in the given documents: **Doc 1:** Ben studies about computers in Computer Lab. **Doc 2:** Steve teaches at Brown University. **Doc 3:** Data Scientists work on large datasets.	*CO1* *CO2*	*PO3*	**10**
		b)	Justify the need of Word Sense Disambiguation. Interpret the role of contextual embeddings in Word sense and similarity with suitable diagram and equations.	*CO1*	*PO1*	**10**
			OR			
	8	a)	Derive Word2vec skip gram embeddings with suitable mathematical formulas.	*CO1*	*PO2*	**10**
		b)	Calculate the Pointwise Mutual Information (PMI) of the words "**cat**" and "**dog**" using given documents. Document 1: "The cat sat on the mat." Document 2 : "The dog sat on the mat." Document 3: "The cat chased the dog."	*CO2*	*PO2*	**5**

<table>
<tr><td></td><td></td><td>c)</td><td>Refer the table to answer following questions:

<table><tr><th></th><th>skills</th><th>Years</th><th>Salary</th></tr><tr><td>Engineer</td><td>60</td><td>40</td><td>70</td></tr><tr><td>Doctor</td><td>80</td><td>50</td><td>90</td></tr><tr><td>Profession</td><td>70</td><td>45</td><td>85</td></tr></table>
i. Calculate the cosine similarity between "engineer" and "profession."
ii. Calculate the cosine similarity between "doctor" and "profession."
iii. Determine which word ("engineer" or "doctor") is more closely related to "profession".</td><td>CO2</td><td>PO2</td><td>5</td></tr>
<tr><td></td><td></td><td></td><td>UNIT - V</td><td></td><td></td><td></td></tr>
<tr><td></td><td>9</td><td>a)</td><td>Summarize the architectures for Coreference and entity linking Algorithms.</td><td>CO1</td><td>PO1</td><td>10</td></tr>
<tr><td></td><td></td><td>b)</td><td>Compare the architectures of BERT, RoBERTa, and ALBERT with a diagram for each.
i. Highlight the key advantages of each model.
ii. Recommend the best model for a large-scale, high-accuracy, and resource-efficient application, and justify your choice.</td><td>CO1</td><td>PO1</td><td>10</td></tr>
<tr><td></td><td></td><td></td><td>OR</td><td></td><td></td><td></td></tr>
<tr><td></td><td>10</td><td>a)</td><td>Outline the key terms in Coreference resolution with suitable diagram.</td><td>CO1</td><td>PO1</td><td>5</td></tr>
<tr><td></td><td></td><td>b)</td><td>Define Entity Linking, Illustrate how does it differ from traditional Coreference Resolution.</td><td>CO1</td><td>PO1</td><td>5</td></tr>
<tr><td></td><td></td><td>c)</td><td>Describe the architecture and working of XLNet. Highlight the key innovations introduced in XLNet, and how does it overcome BERT’s limitations.</td><td>CO1</td><td>PO1</td><td>10</td></tr>
</table>

****