RV College of Engineering
Mysore Road, RV Vidyaniketan Post,
Bengaluru - 560059, Karnataka, India

Department of Artificial Intelligence and Machine Learning

Date: 29.04.2025	Test - 1	Max. Marks: 50+10
Semester: VI	UG	Duration: $1\frac{1}{2}$ Hrs + ½ Hr
Course Title: Natural Language Processing and Transformers		Course Code:AI363IA

Note: Answer all the Questions

Q. No		Questions	M	BT	CO
1	a)	Define "tokenization" in NLP	2	1	1
	b)	Write a code snippet to calculate similarity between two WordNet Synsets.	2	2	3
	c)	Write the purpose of (\w) and \2 in the regular expression (\w*)(\w)\2(\w*).	2	1	2
	d)	What does the regular expression \s+ match in a text?	2	1	1
	e)	List any two data augmentation techniques used in NLP	2	1	1

Note: Answer all the Questions

Q. No		Questions	M	BT	CO
1	a)	Differentiate between Heuristic-based, Machine Learning-based, and Deep Learning-based approaches in NLP.	5	1	4
	b)	Explain the role of Unicode normalization and system-specific error correction in text pre-processing	5	2	2
2	a)	Write a python script to Train a custom sentence tokenizer using NLTK's PunktSentenceTokenizer. Include the training step in your code	5	2	3
	b)	Write a code that retrieves all the synonyms and lemmas for the word "happy" using WordNet.	5	2	3
3	a)	Write a custom function that takes a raw HTML web page as input and returns cleaned, plain text using Beautiful Soup and regex. Ensure you handle Unicode normalization and remove HTML escape characters	5	2	4
4	b)	What are collocations? Write a Python program using NLTK to extract significant collocations (bigrams and trigrams) from a given corpus.	5	2	3
	a)	Briefly describe the steps involved in a typical NLP pipeline from data acquisition to advanced processing.	5	1	2
	b)	What is a Synset? Write a Python program using NLTK to retrieve only the lemmas for the word "computer" from its Synsets.	5	1	2
5	a)	Explain the different ways to tokenize using wordnet with example	5	3	2
	b)	You are developing a chatbot that analyzes live IPL commentary data fetched from Cricbuzz. Often, fans type team names or expressions with exaggerated repeated letters to show excitement, such as typing "RRRRCCCCBB" instead of "RCB". These stretched-out versions can affect the chatbot's ability to understand and respond correctly. **Write a Python script that cleans such inputs by removing excessive consecutive repeating characters and retaining only a single occurrence of each. The script should define a class RepeatReplacer with a method replace(word) that uses regular expressions and recursion to normalize any given word—for example, converting "RRRRCCCCBB" to "RCB"**	5	3	5