VIT

Vellore Institute of Technology
(Deemed to be University under section 3 of UGC Act, 1956)
CHENNAI

Reg. Number:	

Continuous Assessment Test (CAT) – I AUGUST 2025

Programme	:	**B.Tech (ECE/ECM/CSE)**	Semester	:	**FS 2025-26**
Course Code & Course Title	:	**BCSE409L & Natural Language Processing**	Class Number	:	**CH2025260101740 CH2025260100534**
Faculty	:	**Dr.G.Sudhakaran Dr.R.Krithiga**	Slot	:	**D1+TD1**
Duration	:	**90 Mins**	Max. Mark		**50**

General Instructions: < Use this space to provide additional information such as graph sheet, data book etc.>

- Write only your registration number on the question paper in the box provided and do not write other information
- Use statistical tables supplied from the exam cell as necessary
- Use graph sheets supplied from the exam cell as necessary
- Only non-programmable calculator without storage is permitted

Answer all questions

Q. No	Sub Sec.	Description	Marks	CO	BT Level
1.		Imagine you are building a corpus to study the impact of social media on youth mental health in India using data from various online platforms. Explain in detail the factors you must keep in mind while creating this corpus.	10	1	2
2.		You are building a plagiarism detection system that measures document similarity using Minimum Edit Distance, which calculates the fewest insertions, deletions, or substitutions to transform one word into another; describe a dynamic programming approach that breaks the problem into overlapping subproblems with stored intermediate results, define the recurrence relation, explain initialization, and demonstrate by computing the minimum edit distance between **"teacher"** and **"cheater".**	15	1	3
3.		You are developing a morphological analyzer for an agglutinative language (like Tamil or Finnish), where a single word can encode a lot of grammatical information. To handle such complex word formations, you decide to use Finite State Transducers (FSTs). 1. Explain about FST and how it differs from a Finite State Automaton. (3 Marks) 2. Describe its usage in NLP for morphological analysis and generation. (2 Marks) 3. Enumerate the state transitions of the FST for the symbols of the following intermediate level inputs. (5 Marks) a) day^s# b) gas# c) gas^s# d) news^studio#	10	2	3

4.		A computational linguistics team is testing an HMM-based POS tagger (using Table 1) on the ambiguous sentence <S> Will can chase Mary <E> Apply the Viterbi algorithm to find the most probable POS tag sequence, showing transition and emission probabilities, step-by-step Viterbi table calculations, and the final POS-tagged output. Table. 1	15	2	3

<S>	Mary	Jane	can	see	Will	<E>
	N	N	M	V	N	
<S>	Chase	will	see	Mary	<E>	
	M	M	V	N		
<S>	Will	Jane	chase	Mary?	<E>	
	M	N	V	N		
<S>	Mary	will	walk	Chase	<E>	
	N	M	V	N		

**********All the best **********