Vellore Institute of Technology
CHENNAI

Reg. Number:	

Continuous Assessment Test (CAT) – I (January 2025)

Programme	:	B.Tech. Computer and Science Engineering	Semester	:	Winter-2025
Course Code & Course Title	:	BCSE409L & Natural Language Processing	Class Number	:	CH2024250501961 CH2024250501963
Faculty	:	**Dr. Sankar.P** **Dr.Krithiga.R**	Slot	:	D1 + TD1
Duration	:	1½ Hours	Max. Mark		50

General Instructions:

- 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
1		Analyze the lexical relations present in the following corpora and provide a brief explanation for each case. Highlight how these lexical relations impact the intended NLP task. **Corpora**: **Corpus 1:** You are building a search engine for an online bookstore. A user searches for the term ***Python***. **Corpus 2:** An NLP chatbot is parsing customer reviews. A review states the ***dish*** at this restaurant was delicious. **Corpus 3:** A virtual assistant interprets commands for a smart system. The user says the ***service*** at the cafe was excellent. **Corpus 4:** You are developing an email summarization tool. An email as received a ***quick*** response from the support team	8
		Explain Zipf's first law and second law in the context of the sentence ***"Can you can a can as a canner can can a can?"*** Analyze the frequency of the word 'can' and its variations, and describe how these laws apply to the distribution and length of words in the sentence.	4

3			Mention the major text normalization operations (except stemming) in NLP. For each of the below paragraphs, perform all text normalization operations (except stemming) and list the resulting tokens: **Paragraph 1**: With the passing of the legendary tabla maestro Zakir Hussain Sahab, we've lost a legend. His groundbreaking contributions aren't measurable and they made classical traditions feel both timeless and new. Yeah! We'll always remember his iconic 'Wah Taj'. "Zakir Bhai! He left too soon. It's an irreparable loss...!" wrote the actor-filmmaker Kamal Hassan. **Paragraph 2**: Neeraj Chopra sought the blessings from the world on his marriage with tennis player and coach Himani Mor. Most of his over 10 million followers across social media platforms wonder, "Wow! just how'd he keep such a thing private?". "Yes. He informed us about his marriage. He'd've wanted to do it in private with very close ones," a top Athletics Federation of India official said.	6
4			Ravi is working on a spelling correction tool where the minimum edit distance algorithm is one of many algorithms used in it. For the source token '***evinbare***', compute the Levenshtein distance and alignments for the target tokens '***entrance***' and '***vibrance***'. Identify which target is more similar to the source token and explain the result.	12
5			You are developing a sentiment analysis model for customer reviews in an online product feedback system. The reviews are in English and include terms like: "***Allied, Pride, Ride, Guide, Side, Tried, Wide, Reality, Vitality, Past.***" Using Byte Pair Encoding (BPE), perform sub word tokenization and demonstrate each step of merging sub word units for both short and long reviews.	10
6			A B.Tech third-year student was reading the below set of sentences: "***May flower can sing jazz. She dropped a can. She may give a drive. Tina can drive. Both can food.***" He concluded that 'can' in the last sentence refers to the action of canning food. Using these sentences as a training set, evaluate the POS tags for each word in the test sentence: "Jazz can give a drive" using Viterbi decoding for an HMM. Describe the steps involved in computing the most likely sequence of tags.	10

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