



Reg No.: \_\_\_\_\_

Name: \_\_\_\_\_

**APJ ABDUL KALAM TECHNOLOGICAL UNIVERSITY**  
B.Tech Degree S7 (S, FE) Examinations May 2026 (2019 Scheme)

**Course Code: CST473**

**Course Name: NATURAL LANGUAGE PROCESSING**

**Max. Marks: 100**

**Duration: 3 Hours**

**PART A**

*Answer all questions, each carries 3 marks.*

Marks

- |    |                                                                                                                                         |     |
|----|-----------------------------------------------------------------------------------------------------------------------------------------|-----|
| 1  | Describe tokenization and lemmatization with examples.                                                                                  | (3) |
| 2  | Explain why heuristic-based NLP systems struggle with unseen data.                                                                      | (3) |
| 3  | With the help of a neat diagram explain model stacking and ensembling.                                                                  | (3) |
| 4  | Construct the Bag of Words representation for the following sentences:<br>S1: "NLP is interesting"<br>S2: "NLP applications are useful" | (3) |
| 5  | Explain the working of a Named Entity Recognition (NER) system.                                                                         | (3) |
| 6  | Outline the process of model evaluation for text classifiers.                                                                           | (3) |
| 7  | Differentiate Information Retrieval and Information Extraction.                                                                         | (3) |
| 8  | What is an inverted index in information retrieval? Explain how it is used in term weighting and document scoring.                      | (3) |
| 9  | Differentiate between direct and transfer-based Machine Translation approaches.                                                         | (3) |
| 10 | Define Factoid Question Answering with an example.                                                                                      | (3) |

**PART B**

*Answer any one full question from each module, each carries 14 marks.*

**Module I**

- |    |                                                                                                                    |      |
|----|--------------------------------------------------------------------------------------------------------------------|------|
| 11 | a) With the help of a diagram, explain different building blocks of a language and the applications of each block. | (10) |
|    | b) Explain Rule based approach to solve an NLP task.                                                               | (4)  |

**OR**

- |    |                                                                                                                                    |      |
|----|------------------------------------------------------------------------------------------------------------------------------------|------|
| 12 | a) How does Logistic Regression function as a classifier in NLP applications? Compare its use in binary and multi-class scenarios. | (10) |
|    | b) Summarize any four challenges involved in an NLP task with suitable examples.                                                   | (4)  |

**Module II**

- |    |                                                                                          |     |
|----|------------------------------------------------------------------------------------------|-----|
| 13 | a) Explain two different approaches practiced for feature engineering in NLP.            | (8) |
|    | b) With suitable diagram and example explain the SkipGram model for text representation. | (6) |

OR

- 14 a) Explain the concept of word embeddings with an example. How do they capture semantic meaning between words? (8)
- b) Describe the data acquisition and text clean-up stages in an NLP pipeline with suitable examples. (6)

**Module III**

- 15 a) You are designing a system to categorize news articles into sports, politics, and technology. Explain how the text classification pipeline applies here. Which model would you choose for high-dimensional feature space? (8)
- b) Explain the general pipeline for Information Extraction with an example. (6)

OR

- 16 a) Explain the TextRank and SGRank algorithms for keyphrase extraction. (6)
- b) A university is building a research paper database that tags each paper with relevant authors, affiliations, and institutions. Identify the NLP task. Suggest suitable evaluation techniques and challenges in implementation. (8)

**Module IV**

- 17 a) Explain lightly supervised approach for relation analysis. Also state its advantages and limitations. (10)
- b) Summarize how cosine similarity is used for document scoring in information retrieval with a suitable example. (4)

OR

- 18 a) With the help of a diagram explain the architecture of an information retrieval system. (8)
- b) Explain the techniques for evaluating relation detection systems. (6)

**Module V**

- 19 a) With a block diagram, explain the architecture of a Factoid Question Answering System. (10)
- b) Explain the concept of word alignment and its importance in Machine Translation. (4)

OR

- 20 a) Explain in detail the classical approaches to Machine Translation. (9)
- b) Outline the decoding process in phrase-based statistical MT. (5)

\*\*\*\*