# National Forensics Sciences University, Goa Campus
# TA-1 Examination

**Program Name – MTECH AIDS**          **Sem –      II**                    **Date- 11-02-2025**
**Subject Name- Natural Language Processing    Subject Code  -  CTMTAIDS SII P3**
**Time- 45 mins**                                                    **Max. Marks- 25**
**Instructions - 1) Answer all questions. 2) Assume suitable data.**

Q.1       **Multiple Choice Questions (1 mark each)**                              10 marks

1.   Which of the following regex patterns correctly extracts only email addresses
     from a text?
     A) [a-zA-Z0-9]+@[a-zA-Z]+\.[a-z]+
     B) [a-zA-Z0-9._%+-]+@[a-zA-Z0-9.-]+\.[a-zA-Z]{2,}
     C) .*@.*
     D) \S+@\S+

2.   In an N-gram model, what happens as N increases?
     A) Data sparsity increases
     B) Computational complexity decreases
     C) The model becomes less dependent on context
     D) The model assigns lower probabilities to sentences

3.   Which of the following is not true about Word2Vec embeddings?
     A) They capture semantic relationships between words.
     B) They require large corpora to generate meaningful embeddings.
     C) They work by assigning random vectors to words.
     D) They can be fine-tuned for specific domains.

4.   Which of the following statements is true about perplexity in language
     models?
     A) Higher perplexity indicates better performance.
     B) Perplexity is calculated as the inverse of log-likelihood.
     C) Perplexity measures how well a model predicts unseen text.
     D) Perplexity is independent of the training corpus size.

5.   What does Inverse Document Frequency (IDF) help achieve in TF-IDF?
     A) It gives more weight to commonly occurring words.
     B) It reduces the influence of frequently occurring words across documents.
     C) It normalizes term frequency across multiple documents.
     D) It assigns equal importance to all words.

6.   Which of the following is an example of derivational morphology?
     A) "Run" → "Running"
     B) "Act" → "Actor"

C) "Cats" → "Cat"
D) "Write" → "Written"

7. If a word appears frequently in a document but rarely in other documents,
   what would be its TF-IDF score?
   A) High
   B) Low
   C) Zero
   D) Unchanged

8. Which regular expression would correctly match all words that start with "un"
   and end with "ing" (e.g., "understanding", "unraveling")?
   A) un.*ing
   B) ^un.*ing$
   C) un.+ing$
   D) ^un.*ing

9. Which of the following is not true about Word2Vec embeddings?
   A) They capture semantic relationships between words.
   B) They require large corpora to generate meaningful embeddings.
   C) They work by assigning random vectors to words.
   D) They can be fine-tuned for specific domains.

10. In an HMM used for speech recognition, what would typically be the
    observed variable?
    A) The sequence of phonemes
    B) The sequence of words
    C) The acoustic signal (audio features)
    D) The sentence structure

| Q.2 | **Answer any 3 questions (3x5 marks each)** | 15 Marks |
|---|---|---|
| | A. Draw and explain NLP Pipeline with suitable example. | 5 marks |
| | B. Compute the Minimum Edit Distance between the words "INTENTION" and "EXECUTION" using the Levenshtein Algorithm (Edit Distance). | 5 marks |
| | C. Identify the purpose of the following regular expression symbols: ^, $, \d, *, and \b. Provide one example for each symbol in a regex pattern. | 5 marks |
| | D. Document: *"I know you like pizza. I know now your mouth watered, but concentrate on the exam".* | 5 marks |

Calculate the Term Frequency (TF) for each word in the given document.
Show the formula for TF and provide the final term frequencies for each
word in the document.

END