

Natural Language Processing (315329)

MSBTE Diploma · K-Scheme · Model Question Paper — Set C

Time: 3 Hours**Max. Marks: 70****Instructions:**

1. All questions are compulsory.
2. Answer each next main question on a new page.
3. Illustrate your answers with neat sketches wherever necessary.
4. Figures to the right indicate full marks.
5. Assume suitable data, if necessary.
6. Mobile phone, pager and any other electronic communication devices are not permissible in the examination hall.

Q.1 Attempt any FIVE of the following:**10**

- a) Define NLP. Name any four levels (areas) of linguistic analysis used in NLP.
- b) Write a regular expression to extract all hashtags from the tweet: “Loving the #IPL2026 final! #CSK #Cricket_Fever”.
- c) Define lemmatization. Give two examples.
- d) Give any two disadvantages of the Bag-of-Words / TF-IDF representation.
- e) State top-down and bottom-up parsing.
- f) What is a Hugging Face dataset? Write the statement to load a dataset.
- g) What are stop words? Give any four examples.

Q.2 Attempt any THREE of the following:**12**

- a) List and explain any four real-world applications of NLP.
- b) Explain ambiguity in NLP. Describe lexical, syntactic, semantic and pragmatic ambiguity with examples.
- c) Explain the following regular expression concepts with examples: disjunction, grouping, precedence, character sets and counters.
- d) Explain semantic network. Draw it for: “Rohan is a student. A student is a person. Rohan studies in a polytechnic. A polytechnic has a library. A person has a name.”

Q.3 Attempt any THREE of the following:**12**

- a) Describe rule-based tokenization using spaCy with example.
- b) Explain Byte-Pair Encoding with example. Apply 5 merges on: hope (8), hoped (4), hoping (6), rope (7).
- c) Differentiate between Bag-of-Words and Word2Vec (any four points).

- d) Explain transfer learning in NLP. How is a pre-trained model like BERT reused for a new task?

Q.4 Attempt any THREE of the following:

12

- Explain n-gram vectors and their types. Generate unigrams, bigrams and trigrams for “Pune is the cultural capital of Maharashtra”.
- Find the minimum edit distance between “algorithm” and “logarithm” (insertion, deletion and substitution cost 1 each). Show the table.
- Draw the syntax tree(s) for the sentence “The students watched a movie on the laptop.” Is the sentence ambiguous? Explain.
- Explain one-hot encoding. Represent the sentence “I like NLP and I like AI” using one-hot vectors. State its limitations.
- Explain the Hugging Face ecosystem — Transformers, Hub, Tokenizers, Datasets and Accelerate — with an example of pipeline().

Q.5 Attempt any TWO of the following:

12

- Define stemming. Compare the Porter, Snowball and Lancaster stemmers with suitable examples.
- Explain TF-IDF with formula. For D1 “AI is the future”, D2 “AI and ML are the future”, D3 “ML needs data”, compute the TF-IDF matrix and the cosine similarity between D1 and D2 using TF-IDF vectors.
- Explain Named Entity Recognition. Write a Python program using spaCy to extract named entities and their BIO tags.

Q.6 Attempt any TWO of the following:

12

- Explain the Transformer architecture with a neat diagram. Explain multi-head attention and positional encoding.
- Describe POS tagging using a Hidden Markov Model. Using the given probabilities, find the best tag sequence for “fish can swim” using the Viterbi algorithm.

	P(start)	→ N	→ V
N	0.7	0.3	0.7
V	0.3	0.6	0.4

Emission	fish	can	swim
N	0.6	0.2	0.2
V	0.1	0.5	0.4

- Implement chunking in Python using NLTK to extract noun phrases, verb phrases and adjective phrases from “The very tall boy quickly climbed the old mango tree”.

Note: This is a **model question paper prepared by Ananda Technologies** only for practice and exam preparation, based on the pattern of previous MSBTE question papers and the K-scheme syllabus. It is **not an official MSBTE paper**, and it does **not** mean that these questions will come in the exam. Answers follow standard textbook points; marks shown are indicative. Free for students — anandatechnologies.in