

Natural Language Processing (315329)

MSBTE Diploma · K-Scheme · Semester 5 (AI & ML – AN and allied) · Important Questions — unit-wise

Based on MSBTE papers W-2025 and S-2026 (the only two papers of this K-scheme subject available on msbte.ac.in) and the K-scheme syllabus. Practice guidance only — not a prediction of the exam.

Probability: High = asked in both papers (or 3+ times) · **Medium** = asked in one paper · **Low** = in syllabus but not asked yet. With only two papers, these are trends, not certainties — study the full syllabus.

Unit	Title	Syllabus marks	Marks seen in papers (W-25 + S-26)
1	Natural Language Basics	12	20 + 26
2	Text Normalization and Preprocessing	16	24 + 26
3	Text Syntax and Structure	16	24 + 14
4	Text Feature Extraction	14	18 + 18
5	NLP Application and Transformers	12	16 + 18

Paper marks include optional questions (e.g. 7 questions printed in Q.1, any 5 to attempt), so they are higher than the syllabus marks. Unit allotment of a question is our judgement from the syllabus.

Unit I — Natural Language Basics (syllabus: 12 marks)

Question / topic	Marks	Asked in	Count	Probability
Parse tree / syntax tree — structure, components, draw for a given sentence S-26 Q.2c: Discuss the structure and components of parse tree with example. S-26 Q.4c: Draw and explain the syntax tree for "The girl saw the boy with telescope." S-26 Q.5b: Working of parser; parse tree for "She believes he knows the answer."	4, 6	S-26	3	High
Define NLP / real-world applications of NLP W-25 Q.1a: Define Natural Language Processing (NLP). W-25 Q.3a: List and briefly explain any four real-world applications of NLP. S-26 Q.1g: Give any two applications of NLP.	2, 4	S-26, W-25	3	High
Semantic network (concept / with example) W-25 Q.1g: Explain the concept of a semantic network in NLP. S-26 Q.4a: Explain semantic network with suitable example.	2, 4	S-26, W-25	2	High
Text corpora / corpus annotation and utilities W-25 Q.3c: Explain Corpora Annotation and Utilities with example. S-26 Q.2a: Describe the text-corpora.	4	S-26, W-25	2	High
Lexical semantic relations — synonymy, antonymy, hyponymy, polysemy	4	W-25	1	Medium

W-25 Q.2a: Define synonymy, antonymy, hyponymy, polysemy with one example each.				
Word order typology (SVO, SOV, VSO ...) W-25 Q.4a: Explain different types of word order typologies with one example each.	4	W-25	1	Medium
Ambiguity in NLP S-26 Q.1a: Define ambiguity in natural language processing.	2	S-26	1	Medium
Need for NLP and areas of linguistics (phonology, morphology, syntax, semantics, pragmatics)	—	Not yet	0	Low
Words, phrases, clauses and grammar (phrase types)	—	Not yet	0	Low
Popular corpora — Brown, Penn Treebank, WordNet (NLTK)	—	Not yet	0	Low

Unit II — Text Normalization and Preprocessing (syllabus: 16 marks)

Question / topic	Marks	Asked in	Count	Probability
Regular expressions — definition, uses, write a pattern W-25 Q.1b: Define regular expressions. W-25 Q.1c: Write a regular expression to match a 10-digit phone number. S-26 Q.1e: List any two uses of Regular Expression in NLP.	2	S-26, W-25	3	High
N-grams — unigram, bigram, trigram / n-gram generation W-25 Q.6c: Concept of N-gram vectors with types; examples of unigram, bigram, trigram. S-26 Q.3d: Describe n-gram generation from a sentence.	4, 6	S-26, W-25	2	High
Stemming vs lemmatization (compare / differentiate) W-25 Q.2b: Differentiate between stemming and lemmatization. S-26 Q.2b: Compare Stemming and Lemmatization (any four points).	4	S-26, W-25	2	High
Byte-Pair Encoding (BPE) tokenization W-25 Q.4e: Explain Byte Pair Encoding (BPE) in tokenization. S-26 Q.3a: Explain Byte-pair encoding with example.	4	S-26, W-25	2	High
Tokenization — definition, rule-based tokenization using spaCy S-26 Q.1b: Define tokenization. S-26 Q.4b: Describe rule-based tokenization using spacy.	2, 4	S-26	2	Medium
Stemming and the Porter Stemmer (stem given words) W-25 Q.5b: Define stemming; working of Porter Stemmer; stem "running", "happily", "cats", "flying".	6	W-25	1	Medium
Steps of raw text preprocessing S-26 Q.5a: Describe all the steps of raw text preprocessing with example.	6	S-26	1	Medium
Regex operators — disjunction, grouping, precedence, character sets, counters	—	Not yet	0	Low

Removing special characters, expanding contractions, case conversion, stop-word removal	—	Not yet	0	Low
Snowball stemmer (compare with Porter)	—	Not yet	0	Low
n-gram probabilities (bigram model) — related to n-grams	—	Not yet	0	Low
Minimum edit distance (Levenshtein) — from the reference book chapter on text normalization; not named separately in the syllabus	—	Not yet	0	Low

Unit III — Text Syntax and Structure (syllabus: 16 marks)

Question / topic	Marks	Asked in	Count	Probability
Part-of-speech (POS) tagging — technique, algorithm, challenges, tag a sentence W-25 Q.2d: Perform POS tagging on 'Sachin Tendulkar played cricket for India'. W-25 Q.6a: Describe POS tagging; four major challenges with examples. S-26 Q.6a: Describe POS tagging with technique, algorithm and example.	4, 6	S-26, W-25	3	High
Chunking (partial parsing) — explain / implement with NLTK W-25 Q.4c: Explain the process of chunking with examples. S-26 Q.6b: Implement chunking in python using NLTK.	4, 6	S-26, W-25	2	High
Named Entity Recognition and BIO / IOB tagging W-25 Q.1f: Define Named Entity Recognition (NER). W-25 Q.3d: Define the BIO (IOB) tagging format used in NER with example.	2, 4	W-25	2	Medium
Dependency parsing W-25 Q.3b: Explain dependency parsing with example.	4	W-25	1	Medium
Top-down and bottom-up parsing S-26 Q.1c: State top-down and bottom-up parsing.	2	S-26	1	Medium
English word classes (open / closed class) and Penn Treebank tag set	—	Not yet	0	Low
HMM-based POS tagging (Viterbi) — stochastic technique	—	Not yet	0	Low

Unit IV — Text Feature Extraction (syllabus: 14 marks)

Question / topic	Marks	Asked in	Count	Probability
Cosine similarity (state / numerical on term vectors) W-25 Q.5c: Cosine similarity of $A=[3,0,2,5,1]$, $B=[1,2,0,3,4]$; interpret. S-26 Q.1f: State cosine similarity.	2, 6	S-26, W-25	2	High
Contextual embeddings — polysemy, use cases S-26 Q.3b: Describe contextual embeddings to solve problem of polysemy.	4, 6	S-26	2	Medium

S-26 Q.5c: Explain contextual embeddings also give use cases.				
TF-IDF (with example / disadvantages) S-26 Q.1d: Give any two disadvantages of TF-IDF. S-26 Q.4d: Discuss TF-IDF with example.	2, 4	S-26	2	Medium
Vector space model — words as vectors W-25 Q.6b: Explain vector space model with respect to words and vectors.	6	W-25	1	Medium
Bag-of-Words vs Word2Vec W-25 Q.4d: Differentiate between Bag-of-Words and Word2Vec.	4	W-25	1	Medium
One-hot encoding W-25 Q.1e: Define One-Hot Encoding.	2	W-25	1	Medium
Word2Vec — CBOW and Skip-gram architectures	—	Not yet	0	Low
Bag-of-Words model (build a BoW matrix)	—	Not yet	0	Low

Unit V — NLP Application and Transformers (syllabus: 12 marks)

Question / topic	Marks	Asked in	Count	Probability
Transformer architecture (encoder–decoder, diagram) and its challenges W-25 Q.1d: Mention any two challenges faced while using Transformers. W-25 Q.5a: Explain Transformer architecture with encoder-decoder diagram and working. S-26 Q.3c: Draw and explain the architecture of transformer.	2, 4, 6	S-26, W-25	3	High
Transfer learning in NLP (and its benefits) W-25 Q.4b: Explain transfer learning in NLP. Mention its benefits. S-26 Q.2d: Explain transfer learning in NLP.	4	S-26, W-25	2	High
Hugging Face — datasets, tokenizers, Trainer API, fine-tuning W-25 Q.2c: Hugging Face Transformers with custom datasets for fine-tuning a sentiment model; tokenizers and Trainer API. S-26 Q.4e: Describe Hugging Face dataset with example.	4	S-26, W-25	2	High
Sentiment / text classification application in Python S-26 Q.6c: Implement sentiment classification application in python.	6	S-26	1	Medium
Attention / self-attention mechanism	—	Not yet	0	Low
Encoder–decoder framework	—	Not yet	0	Low
NLP application pipeline; evaluating classifiers (accuracy, precision, recall, F1)	—	Not yet	0	Low
Hugging Face Hub, Tokenizers library, Accelerate	—	Not yet	0	Low

Paper pattern (same in W-2025 and S-2026)

Q.	Attempt	Marks each	Total
1	Any FIVE of 7	2	10
2	Any THREE of 4	4	12
3	Any THREE of 4	4	12
4	Any THREE of 5	4	12
5	Any TWO of 3	6	12
6	Any TWO of 3	6	12
	3 Hours		70

Quick tips from the two papers

- Every paper had at least one **sentence-based question** (POS-tag a sentence, draw a parse/syntax tree) — practise tagging and drawing trees.
- **Transformer architecture** appeared for 4 M and 6 M — learn to draw the encoder–decoder diagram neatly.
- Programs asked so far are short **NLTK / Python** programs (chunking, sentiment classification) — write imports, the main steps and sample output.
- One **numerical** (cosine similarity) was asked for 6 M — practise TF-IDF and cosine similarity calculations step by step.

Note: This is an **important-questions list 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