Affiliated to CBSE, New Delhi (Affiliation No. 1030642)

GNPS

GOMTI NANDAN PUBLIC SCHOOL

Gandhi Ward, Bina (M.P.) 470113. | 9165 609 333. | info@mygnps.com | www.mygnps.com

CLASS X · ARTIFICIAL INTELLIGENCE (417) · SKILL SUBJECT · CBSE BOARD EXAMINATION 2027

PART B · UNIT 6

Natural Language Processing

Unit Booklet · Short Notes · Board-Pattern Questions · Model Answers

Part B · 8 of 40 theory marks | 29 questions with model answers | Sonakshi Ma'am | Taught: Sat 28 Nov

INSIDE THIS BOOKLET

1	**Unit Notes** 9 note blocks, key terms, exam tips	2
A	**Objective Questions** 17 questions · 1 mark	6
B	**Short Answer Questions** 6 questions · 2 marks	7
C	**Long Answer Questions** 4 questions · 4 marks	7
D	**Case-Based Questions** 2 cases · 4 marks	8
E	**Answers** Key + model answers, marking-scheme style	9

STUDENT DETAILS

NAME _____

CLASS & SECTION _____

ROLL NO. _____

UNIT FINISHED ON _____

OBJECTIVE SELF-CHECK _____ / 17

1

READ THIS FIRST · SHORT NOTES

Unit Notes

What makes human language hard for a computer, where NLP is used, the five stages of NLP, script bots and smart bots, text normalisation (segmentation, tokens, stopwords, stemming, lemmatization), the Bag of Words model and TF-IDF with worked sums.

THIS UNIT IN THE BOARD PAPER

8 / 40

Unit 6 carries **8 theory marks** (the sentiment-analysis walkthrough in Orange is tested in the practical exam). In the CBSE sample paper it gave **one-mark items** such as naming an NLP stage and counting tokens, and a **4-mark Bag of Words** question.

SYLLABUS 2026–27 · WHAT CBSE LISTS FOR THIS UNIT

- Introduction: features of natural language; need for NLP
- Applications: voice assistants, auto captions, translation, sentiment analysis, text classification, keyword extraction
- Stages of NLP
- Chatbots: script bot and smart bot
- Text processing: text normalisation, Bag of Words, TF-IDF
- Practical part: NLP use case walkthrough – sentiment analysis with Orange

HOW THE CBSE SKILL-SUBJECT PAPER IS SET (THEORY 50 MARKS · 2 HOURS)

1 mark
Objective: MCQ, assertion–reason, fill in the blank, true/false, match. Q1 is Part A; Q2–Q5 (answer 5 of 6 in each) are Part B.

2 marks
Short answer in **20–30 words**: two clear points. Part B: answer any 4 of 6 (Q11–Q16).

4 marks
Long answer in **50–80 words**: four points, or 2 + 2. Answer any 3 of 5 (Q17–Q21); often case-based.

Part B = 40
Objective 20 + short 8 + long 12. Part A (Employability Skills) = 10. Practical = 50.

Write in points, start each point with the key word, and stay inside the word limit. Objective answers: write the option letter **and** the answer.

PRACTICAL WORK FOR THIS UNIT

Keyword extraction from a news paragraph · chat with a script bot and a smart bot and compare · normalise a short text by hand · build a Bag of Words table for three sentences · sentiment analysis of reviews in Orange (no-code).

1 What is NLP?

> **Natural Language Processing (NLP):** the domain of AI that lets computers **read, understand, process and reply in human language** – spoken or written.

Features of natural language (why it is hard for a computer):

- It has **rules** of grammar and word order, and every language has its own.
- One word can have **many meanings**: 'bat' (cricket bat or the animal); the meaning comes from **context**.
- Different words can mean the **same** thing: 'big' and 'large'.
- A sentence can be correct in grammar but have **no sense**; people also use slang, short forms and emojis.
- Language keeps **changing**: new words are added every year.

Need for NLP: huge amounts of text and speech are created every day (messages, reviews, calls). NLP lets machines **talk with people** in their own language and **automate** work such as translation, customer help and sorting e-mails.

2 Applications of NLP

Application	What it does	Example
Voice assistants	Turn speech into text, understand it and reply	Alexa, Siri, Google Assistant
Auto-generated captions	Write on-screen text for what is spoken	Captions on videos and online classes
Language translation	Convert text from one language to another	Hindi notice → English
Sentiment analysis	Find the **feeling** in a text: positive, negative or neutral	Sorting product reviews
Text classification	Put a text into a **category**	Spam / not spam; sports or business news
Keyword extraction	Pick out the most important words of a text	Tags for a blog post

3 Stages of NLP

Stage	What happens	Example
1. Lexical analysis	Text is broken into paragraphs, sentences and words; each word's meaning is looked up	'Riya reads.' → Riya | reads | .
2. Syntactic analysis (parsing)	Checks **grammar** and word order	'Sat the mat on cat' is rejected
3. Semantic analysis	Checks that the sentence **makes sense**	'The pen ate an apple' – grammar fine, meaning wrong
4. Discourse integration	Meaning depends on the **sentences before**	'Riya bought a book. She liked it.' – she = Riya, it = book
5. Pragmatic analysis	Finds the **intended** meaning in a real situation	'Can you close the door?' is a request, not a yes/no question

Memory line: **L**ook – **S**tructure – **S**ense – **D**iscourse – **P**urpose.

4 Chatbots: script bot and smart bot

> **Chatbot:** a program that chats with people through text or voice. It is one of the most common uses of NLP.

	Script bot	Smart bot
Works on	A fixed **script** of rules and options	AI and NLP with a large database
Language processing	Little or none	Understands free-form questions
Learning	Does not learn	**Learns** and improves with more data
Making it	Easy, needs little coding	Needs much data, coding and training
Example	Bank or delivery helpline bot with menu buttons	Alexa, Siri, Google Assistant, AI chat assistants

5 Text normalisation

Text normalisation: cleaning text into a simple, standard form so that a computer can work on it. The whole text taken from all documents is called the **corpus**.

1. **Sentence segmentation** – split the corpus into sentences.
2. **Tokenisation** – split each sentence into **tokens**. Every word, number and punctuation mark is a separate token.
3. **Remove stopwords, special characters and numbers** – stopwords are very common words (a, an, the, is, and, to) that add little meaning.
4. **Convert to the same case** – usually lowercase, so 'Cricket' and 'cricket' are one word.
5. **Stemming or lemmatization** – reduce each word to its root.

COUNTING TOKENS

'Hello! Riya, are you coming today?' → tokens: Hello | ! | Riya | , | are | you | coming | today | ? → **9 tokens**.

6 Stemming and lemmatization

	Stemming	Lemmatization
How	Cuts off the ending (affix) by rule	Finds the dictionary root word, called the **lemma**
Result	May **not** be a real word	Always a **meaningful** word
'studies'	studi	study
'caring'	car	care
Speed	Faster	Slower but more accurate

7 Bag of Words (BoW)

Bag of Words: an NLP model that turns text into numbers by **counting how often each word occurs**. Word order is ignored – the words are thrown into a 'bag'.

1. **Normalise** the text.
2. Make the **dictionary** – every unique word of the corpus, written once.
3. Make a **document vector** – for one document, count each dictionary word.
4. Repeat for **all documents**.

Worked example. D1: 'Aman likes cricket.' D2: 'Riya likes football.' D3: 'Aman and Riya play cricket.'

After normalisation (lowercase, full stops and the stopword 'and' removed) the dictionary is: **aman, likes, cricket, riya, football, play** (6 words).

	aman	likes	cricket	riya	football	play
D1	1	1	1	0	0	0
D2	0	1	0	1	1	0
D3	1	0	1	1	0	1

8 TF-IDF

- **TF (term frequency):** how many times the word occurs in **this** document (the BoW count).
- **DF (document frequency):** in how many documents the word occurs. **N** = total number of documents.
- **IDF = log(N ÷ DF)** (log to base 10). **TF-IDF = TF × log(N ÷ DF)**.
- A word found in **every** document gets log(1) = **0** – it tells nothing special (like stopwords). A word that is **frequent here but rare elsewhere** gets a **high** value – it is a keyword.

Word (example above)	DF	N ÷ DF	IDF
aman, likes, cricket, riya	2	1.5	log 1.5 = 0.176
football, play	1	3	log 3 = 0.477

READING THE VALUES

In D3, aman, cricket and riya each score 1 × 0.176 = 0.176, but **play** scores 1 × 0.477 = **0.477**. So 'play' is the most important word of D3.

9 For the lab: sentiment analysis in Orange

- NLP can be done with **code** (Python libraries) or with **no-code** tools such as Orange.
- In Orange: load the reviews (Corpus) → Preprocess Text (tokens, lowercase, stopwords) → Sentiment Analysis → view the scores in a Data Table or chart.

This part is tested in the practical exam and viva, not in the theory paper.

Key terms

17 terms

NLP: AI domain for human language

Corpus: all the text from all documents

Document: one piece of text in the corpus

Token: a word, number or punctuation mark

Stopword: very common word with little meaning

Sentence segmentation: splitting text into sentences

Tokenisation: splitting sentences into tokens

Stemming: cutting a word to its stem

Lemma: meaningful root word

Bag of Words: word counts, order ignored

Dictionary: list of unique words in the corpus

TF: count of a word in one document

DF: number of documents having the word

IDF: log(N ÷ DF)

Sentiment analysis: positive, negative or neutral feeling

Script bot: rule-based chatbot

Smart bot: AI chatbot that learns

Common mistakes and exam tips

AVOID THESE

- Count **punctuation marks** as tokens.
- Stemming may give a non-word (studi); lemmatization always gives a real word (study).
- Stage order: Lexical → Syntactic → Semantic → Discourse → Pragmatic.
- In BoW keep the dictionary in the **same order** for every document vector.
- IDF is **log(N ÷ DF)**, not log(DF ÷ N). A word in every document has IDF 0.

A

PART A · 1 MARK EACH

Objective Questions

17 questions: MCQs, assertion–reason, fill in the blanks and true/false, like Q2–Q5 of the board paper. Answers in Part E.

ASSERTION–REASON: HOW TO ANSWER

(a) Both A and R are true, and R is the correct explanation of A.
(b) Both A and R are true, but R is not the correct explanation of A.
(c) A is true, but R is false.
(d) A is false, but R is true.

A1. Which AI domain helps computers understand and reply in human language? [1]

(a) Computer Vision
(b) Data Science
(c) Natural Language Processing
(d) Robotics

A2. Finding out whether a movie review is positive or negative is called: [1]

(a) Keyword extraction
(b) Sentiment analysis
(c) Translation
(d) Auto captioning

A3. The number of tokens in 'AI is fun!' is: [1]

(a) 3 **(b)** 4 **(c)** 5 **(d)** 2

A4. All the text taken from all the documents together is called the: [1]

(a) Corpus
(b) Dictionary
(c) Token
(d) Lemma

A5. Which stage of NLP checks whether a sentence follows the rules of grammar? [1]

(a) Lexical analysis
(b) Syntactic analysis
(c) Semantic analysis
(d) Pragmatic analysis

A6. Which of these is a stopword? [1]

(a) cricket
(b) the
(c) doctor
(d) football

A7. Stemming the word 'studies' gives: [1]

(a) study
(b) studi
(c) studies
(d) stud

A8. In TF-IDF, a word that occurs in every document has an IDF of: [1]

(a) 1 **(b)** 0 **(c)** N **(d)** 0.5

A9. Voice assistants such as Alexa and Siri are examples of: [1]

(a) Script bots
(b) Smart bots
(c) Search engines
(d) Spreadsheets

A10. ASSERTION–REASON [1]

Assertion (A): Lemmatization is slower than stemming.

Reason (R): Lemmatization checks that the root it gives is a meaningful dictionary word.

(a) Both A and R are true, and R is the correct explanation of A.
(b) Both A and R are true, but R is not the correct explanation of A.
(c) A is true, but R is false.
(d) A is false, but R is true.

A11. ASSERTION–REASON [1]

Assertion (A): Bag of Words keeps the order of words in a sentence.

Reason (R): Bag of Words counts how many times each word occurs.

(a) Both A and R are true, and R is the correct explanation of A.
(b) Both A and R are true, but R is not the correct explanation of A.
(c) A is true, but R is false.
(d) A is false, but R is true.

A12. FILL IN THE BLANK [1]

Breaking a paragraph into separate sentences is called _____.

A13. FILL IN THE BLANK [1]

Working out that 'she' in the second sentence means Riya happens in _____ integration.

A14. FILL IN THE BLANK [1]

TF-IDF = TF × log(_____ ÷ DF).

A15. TRUE OR FALSE [1]

A script bot learns from every chat and improves over time.

A16. TRUE OR FALSE [1]

Punctuation marks are counted as tokens.

A17. TRUE OR FALSE [1]

Understanding that 'Can you close the door?' is a request happens in pragmatic analysis.

B — Short Answer Questions

PART B · 2 MARKS EACH

Answer in 20–30 words: two clear points. Model answers in Part E.

B1. What is a token? Count the tokens in: 'Hello! Riya, are you coming today?' MUST DO [2]

B2. Differentiate between stemming and lemmatization. MUST DO [2]

B3. What are stopwords? Why are they removed? [2]

B4. Give any two applications of NLP. [2]

B5. Write the formula for TF-IDF. What does a high TF-IDF value show? [2]

B6. Differentiate between syntactic analysis and semantic analysis. [2]

C — Long Answer Questions

PART C · 4 MARKS EACH

Answer in 50–80 words: four points, or two points with an example each. Model answers in Part E.

C1. Explain any four stages of Natural Language Processing. MUST DO [4]

C2. Create a Bag of Words for: D1 'Aman likes cricket.' D2 'Riya likes football.' D3 'Aman and Riya play cricket.' (Remove the stopword 'and'.) MUST DO [4]

C3. Differentiate between a script bot and a smart bot (four points). [4]

C4. Explain the steps of text normalisation. [4]

D

PART D · CASE AND COMPETENCY BASED

Case-Based Questions

Read the situation, then answer from it and from the unit. These are the 4-mark board questions that test application.

D1. Case-Based · The canteen feedback bot [4]

Our school canteen collects feedback on a tablet. A bot greets students and offers five buttons: 'Menu', 'Price', 'Timings', 'Complaint' and 'Exit'. It cannot answer anything else. Written comments such as 'The samosa was great!' and 'Food was cold and late.' are sorted by a program into happy and unhappy feedback.

(i) Which NLP application sorts the comments into happy and unhappy? [1]

(ii) Is the canteen bot a script bot or a smart bot? Give one reason. [1]

(iii) Count the tokens in 'Food was cold and late.' Which of them are stopwords? [2]

D2. Competency-Based · Finding the keyword [4]

After normalisation, three short notes on the environment read:
D1: rain water save **D2:** save tree **D3:** tree plant rain
Use log 3 = 0.477 and log 1.5 = 0.176.

(i) Write the dictionary for this corpus. [1]

(ii) Find the document frequency (DF) of 'save' and of 'plant'. [1]

(iii) Find the TF-IDF of 'tree' and of 'plant' in D3. Which is the keyword of D3? [2]

E

PART E · ANSWERS IN CBSE MARKING-SCHEME STYLE

Answers

Objective key with a reason, then model answers: numbered value points, key word first, the mark for each point in orange, and the word count under each answer. Any other correct, relevant point also earns the mark.

Objective questions – answer key

Part A

Write the option letter **and** the answer, as the marking scheme does.

Q	Answer	Why (one line)
A1	**(c) Natural Language Processing**	NLP works with text and speech.
A2	**(b) Sentiment analysis**	It reads the feeling in the text.
A3	**(b) 4**	AI | is | fun | ! – the '!' is a token.
A4	**(a) Corpus**	Corpus = whole body of text.
A5	**(b) Syntactic analysis**	Syntax = grammar and word order.
A6	**(b) the**	Very common, little meaning.
A7	**(b) studi**	Stemming cuts the ending; the result need not be a real word.
A8	**(b) 0**	log(N ÷ N) = log 1 = 0.
A9	**(b) Smart bots**	They use AI and learn from data.
A10	**(a) Both true; R explains A**	The extra check takes time, so R explains A.
A11	**(d) A false, R true**	A is false (order is ignored); R is true.
A12	**sentence segmentation**	First step of text normalisation.
A13	**discourse**	Meaning depends on the earlier sentence.
A14	**N (total number of documents)**	IDF = log(N ÷ DF).
A15	**False**	Correct: a script bot follows a fixed script and does not learn; smart bots learn.
A16	**True**	Each one is a separate token.
A17	**True**	Pragmatic analysis finds the intended meaning.

Short answers

Part B · 2 marks

B1. What is a token? Count the tokens in: 'Hello! Riya, are you coming today?' [2]

1. **Token:** each word, number or punctuation mark after tokenisation. 1
2. **Count:** Hello | ! | Riya | , | are | you | coming | today | ? = 9 tokens. 1

30 words · CBSE asks 20–30 words | Marking: 1 + 1 = 2 | 1 + 1 | MUST DO

B2. Differentiate between stemming and lemmatization. [2]

1. **Stemming:** cuts the ending; result may not be a real word: studies → studi. 1
2. **Lemmatization:** gives a meaningful root (lemma): studies → study; slower but accurate. 1

26 words · CBSE asks 20–30 words | Marking: 1 + 1 = 2 | 1 + 1 | MUST DO

B3. What are stopwords? Why are they removed? [2]

1. **Stopwords:** very common words such as a, an, the, is, and. 1
2. **Reason:** they add little meaning, so removing them reduces the data and noise. 1

24 words · CBSE asks 20–30 words | Marking: 1 + 1 = 2 | 1 + 1

B4. Give any two applications of NLP. [2]

1. **Sentiment analysis:** finds whether a review is positive, negative or neutral. 1
2. **Translation:** converts text from one language to another, e.g. Hindi to English. 1

23 words · CBSE asks 20–30 words | Marking: 1 + 1 = 2 | 1 + 1

B5. Write the formula for TF-IDF. What does a high TF-IDF value show? [2]

1. **Formula:** TF-IDF = TF × log(N ÷ DF). 1
2. **High value:** the word is frequent in this document but rare in others, so it is important. 1

25 words · CBSE asks 20–30 words | Marking: 1 + 1 = 2 | 1 + 1

B6. Differentiate between syntactic analysis and semantic analysis. [2]

1. **Syntactic:** checks grammar and word order: 'Sat the mat on cat' fails. 1
2. **Semantic:** checks meaning: 'The pen ate an apple' is grammatical but meaningless. 1

24 words · CBSE asks 20–30 words | Marking: 1 + 1 = 2 | 1 + 1

Long answers

Part C · 4 marks

C1. Explain any four stages of Natural Language Processing. [4]

1. **Lexical analysis:** text is split into sentences and words; word meanings are looked up. 1
2. **Syntactic analysis:** checks grammar and word order. 1
3. **Semantic analysis:** checks that the sentence makes sense. 1
4. **Discourse integration:** meaning depends on earlier sentences: 'She liked it' refers back. 1

41 words · CBSE asks 50–80 words | Marking: 1 + 1 + 1 + 1 = 4 | 1 mark each | MUST DO

C2. Create a Bag of Words for: D1 'Aman likes cricket.' D2 'Riya likes football.' D3 'Aman and Riya play cricket.' (Remove the stopword 'and'.) [4]

1. **Normalise:** lowercase; remove full stops and 'and'. 1
2. **Dictionary:** aman, likes, cricket, riya, football, play. 1
3. **D1 and D2:** [1, 1, 1, 0, 0, 0] and [0, 1, 0, 1, 1, 0]. 1
4. **D3:** [1, 0, 1, 1, 0, 1]. 1

37 words · CBSE asks 50–80 words | Marking: 1 + 1 + 1 + 1 = 4 | 1 mark each | MUST DO

C3. Differentiate between a script bot and a smart bot (four points). **[4]**

1. **Working:** script bot follows fixed rules; smart bot uses AI and NLP. 1
2. **Learning:** script bot does not learn; smart bot improves with data. 1
3. **Making:** script bot is easy to build; smart bot needs data and training. 1
4. **Example:** helpline menu bot; Alexa or Siri. 1

43 words · CBSE asks 50–80 words | Marking: 1 + 1 + 1 + 1 = 4 | 1 mark each

C4. Explain the steps of text normalisation. **[4]**

1. **Sentence segmentation:** split the corpus into sentences. 1
2. **Tokenisation:** split sentences into words, numbers and punctuation tokens. 1
3. **Stopwords and case:** remove stopwords, special characters and numbers; change all text to lowercase. 1
4. **Stemming or lemmatization:** reduce each word to its root: studies → studi or study. 1

44 words · CBSE asks 50–80 words | Marking: 1 + 1 + 1 + 1 = 4 | 1 mark each

Case-based answers

Part D

D1. Case-Based · The canteen feedback bot (passage printed in Part D) **[4]**

(i) Which NLP application sorts the comments into happy and unhappy? **[1]**

Sentiment analysis (positive / negative).

(ii) Is the canteen bot a script bot or a smart bot? Give one reason. **[1]**

Script bot – it works only from five fixed buttons and cannot understand other questions or learn.

(iii) Count the tokens in 'Food was cold and late.' Which of them are stopwords? **[2]**

Food | was | cold | and | late | . = **6 tokens**. Stopwords: **was, and**.

Marking: 1 + 1 + 2 = 4 | 1 + 1 + (1 + 1)

D2. Competency-Based · Finding the keyword (passage printed in Part D) **[4]**

(i) Write the dictionary for this corpus. **[1]**

rain, water, save, tree, plant (**5 words**).

(ii) Find the document frequency (DF) of 'save' and of 'plant'. **[1]**

save is in D1 and D2: **DF = 2**. plant is only in D3: **DF = 1**.

(iii) Find the TF-IDF of 'tree' and of 'plant' in D3. Which is the keyword of D3? **[2]**

tree: 1 × log(3 ÷ 2) = **0.176**. plant: 1 × log(3 ÷ 1) = **0.477**. '**plant**' is the keyword – it is found only in D3.

Marking: 1 + 1 + 2 = 4 | 1 + 1 + (1 + 1)