1. What are Corpora?
Corpora (singular: corpus) are large and structured sets of texts used for training or testing natural language processing systems. They can be collections of news articles, books, tweets, or any other type of written or spoken text.

---

2. What are Tokens?
Tokens are the individual units or pieces resulting from breaking down text. In most cases, tokens are words, but they can also be punctuation marks or subwords, depending on the tokenizer.

Example:
Text: "I love NLP." → Tokens: ["I", "love", "NLP", "."]

---

3. What are Unigrams, Bigrams, Trigrams?
These are types of **n-grams**, sequences of 'n' tokens from text.

- **Unigram**: 1 word → "I", "love", "NLP"
- **Bigram**: 2 words → "I love", "love NLP"
- **Trigram**: 3 words → "I love NLP"

---

4. How to Generate N-Grams from Text?
To generate n-grams:

1. Tokenize the text.
2. Slide a window of size *n* over the tokens to form groups.

Example:
Text: "I love NLP"
Bigrams: [("I", "love"), ("love", "NLP")]

Many libraries like NLTK or scikit-learn have built-in functions for this.

---

5. Explain Lemmatization
Lemmatization reduces a word to its base or dictionary form (called a lemma), using vocabulary and grammar.
Example:

"running" → "run"
"better" → "good"
It considers the context and part of speech.

---

6. Explain Stemming

Stemming cuts off prefixes or suffixes to reduce words to a base form, often by crude rules, without understanding grammar.
Example:
"running" → "run"
"flies" → "fli"
It's faster but less accurate than lemmatization.

---

7. Explain Part-of-Speech (POS) Tagging

POS tagging assigns a grammatical category (like noun, verb, adjective) to each word in a sentence.
Example:
"I love NLP" → [("I", pronoun), ("love", verb), ("NLP", noun)]

---

8. Explain Chunking or Shallow Parsing

Chunking groups words into meaningful phrases (chunks), like noun or verb phrases, without doing full syntactic parsing. It uses POS tags to form these chunks.

Example:
"The quick brown fox" → noun phrase (NP)

---

9. Explain Noun Phrase (NP) Chunking

NP chunking specifically targets groups of words that function as nouns. It identifies base noun phrases using POS patterns.
Example:
In "The red car", the chunk "The red car" is a noun phrase.

---

10. Explain Named Entity Recognition (NER)

NER identifies and classifies named entities in text into predefined categories like person names, organizations, locations, dates, etc.
Example:
"Barack Obama was born in Hawaii."
NER tags: "Barack Obama" → PERSON, "Hawaii" → LOCATION