1. What are Corpora?
2. What are Tokens?
3. What are Unigrams, Bigrams, Trigrams?
4. How to generate n-grams from text?
5. Explain Lemmatization
6. Explain Stemming
7. Explain Part-of-speech (POS) tagging
8. Explain Chunking or shallow parsing
9. Explain Noun Phrase (NP) chunking
10. Explain Named Entity Recognition

Answer:

1. Corpora: Corpora are large collections of written or spoken texts that are used as a basis for linguistic analysis. Corpora can include texts from a variety of sources, such as books, articles, websites, speeches, and conversations, and can be annotated with linguistic information like part-of-speech tags, named entities, and syntactic structures.
2. Tokens: Tokens are the basic units of text that are used for natural language processing. A token is typically a word or a sequence of characters that is separated by whitespace or punctuation marks. Tokens are important for tasks like text classification, sentiment analysis, and machine translation.
3. Unigrams, Bigrams, Trigrams: Unigrams, Bigrams, and Trigrams are types of n-grams that represent sequences of one, two, and three words, respectively. N-grams are used to capture the local context of words and can be used for tasks like language modeling, spell checking, and keyword extraction.
4. Generating n-grams from text: N-grams can be generated from text by splitting the text into individual tokens and then grouping adjacent tokens into n-grams. This can be done using libraries like NLTK or Scikit-learn in Python.
5. Lemmatization: Lemmatization is the process of reducing words to their base or dictionary form, called a lemma. The goal of lemmatization is to reduce the inflectional forms of words to a common base form, which can help in tasks like information retrieval and text mining. Lemmatization is often used in conjunction with part-of-speech tagging to disambiguate the meaning of words.
6. Stemming: Stemming is the process of reducing words to their root or base form, called a stem. The goal of stemming is to reduce the number of different forms of a word that occur in text, which can help in tasks like information retrieval and text mining. Stemming is a simpler and faster process than lemmatization, but it can result in less accurate results.
7. Part-of-speech (POS) tagging: Part-of-speech tagging is the process of assigning grammatical tags to words in a sentence, based on their syntactic role. The tags can include information like noun, verb, adjective, adverb, and others, which can be used to disambiguate the meaning of words and to identify patterns in the text. POS tagging is a fundamental task in natural language processing and is used in applications like text classification, machine translation, and information extraction.
8. Chunking or shallow parsing: Chunking or shallow parsing is the process of identifying and grouping together syntactically related words in a sentence, based on their part-of-speech tags. The resulting groups are called chunks, and they can include noun phrases, verb phrases, prepositional phrases, and others. Chunking is a useful technique for identifying patterns in text and for extracting information from unstructured data.
9. Noun Phrase (NP) chunking: Noun Phrase (NP) chunking is a type of chunking that specifically identifies and groups together noun phrases in a sentence. A noun phrase is a group of words that functions as a noun in a sentence, and it typically includes a noun and any accompanying modifiers like adjectives, determiners, and prepositional phrases. NP chunking is a useful technique for identifying important concepts in text and for extracting structured data from unstructured data.
10. Named Entity Recognition: Named Entity Recognition (NER) is the process of identifying and classifying named entities in text, such as people, organizations, locations, and others. NER is typically done using machine learning algorithms, and it can be used to extract structured data from unstructured text, such as in news articles, social media posts