ASSIGNMENT - 2

1. What are Corpora?

Ans: Corpora (plural of corpus) are large collections of text data. These can be books, articles, websites, social media conversations, or any other type of written material. Corpora are essential for training NLP models and analyzing language patterns.

2. What are Tokens?

Ans: Tokens are the smallest units of meaning in a text that are considered individually. They can be words, numbers, punctuation marks, or even emojis. Tokenization is the process of breaking down text into its constituent tokens.

3. What are Unigrams, Bigrams, Trigrams?

Ans: N-grams are sequences of n tokens that appear together in a text. Common types include:

* Unigrams: Single words ("the", "cat", "jumped")
* Bigrams: Two consecutive words ("the cat", "cat jumped")
* Trigrams: Three consecutive words ("the cat jumped", "cat jumped over") N-grams can be used to capture word co-occurrence patterns and analyze language structure.

4. How to generate n-grams from text?

Ans; Most NLP libraries offer functions to generate n-grams from text data. Here's a general process:

* Tokenize the text.
* Use a loop to iterate through the tokens.
* For unigrams, simply add each token to the n-gram list.
* For bigrams and trigrams, create sequences of consecutive tokens within a defined window size (e.g., bigrams look at the current and next token).
* Add the n-gram sequence to the list.

5. Explain Lemmatization

Ans: Lemmatization aims to reduce words to their dictionary base form (lemma). This helps handle variations like singular/plural ("runs" -> "run") or present/past tense ("walked" -> "walk"). Lemmatization uses morphological analysis to identify the lemma based on the word's structure and context.

6. Explain Stemming

Ans: Stemming is a simpler approach that chops off suffixes to reduce words to their base stem. However, stemming can sometimes create incorrect base forms ("running" -> "run" might lose the present participle meaning). Lemmatization generally leads to more accurate base forms.

7. Explain Part-of-speech (POS) tagging

Ans; Part-of-speech (POS) tagging assigns a grammatical category (e.g., noun, verb, adjective) to each token in a sentence. POS taggers analyze the word's context and morphology to assign the most likely part of speech. This information is crucial for tasks like syntactic analysis and dependency parsing.

8. Explain Chunking or shallow parsing

Ans; Chunking (shallow parsing) identifies groups of related words that form grammatical units in a sentence. It's less complex than full parsing but helps identify chunks like noun phrases (NPs), verb phrases (VPs), and prepositional phrases (PPs). Chunking typically relies on rules or statistical models based on POS tags.

9. Explain Noun Phrase (NP) chunking

Ans: Noun Phrase (NP) chunking specifically focuses on identifying noun phrases in text. These phrases typically contain a noun (head) and its modifiers (articles, adjectives). NP chunking is useful for information extraction tasks and sentiment analysis where understanding the subjects and objects of sentences is important.

10. Explain Named Entity Recognition

Ans: Named Entity Recognition (NER) identifies and classifies named entities in text, such as people, organizations, locations, dates, monetary values, etc. NER systems use rules, machine learning models, or a combination of both to recognize these entities and categorize them into predefined types. NER is valuable for tasks like information retrieval and question answering.