NLP assingments-2

1. Corpora are large collections of texts, often used for linguistic research. They can be either general, containing texts from a wide range of sources, or specialized, containing texts from a particular field or genre.
2. Tokens are the basic units that a natural language processing system uses to represent text. They can be words, punctuation marks, or other symbols, and are often used as the basic building blocks for processing and analyzing text.
3. Unigrams are individual tokens, bigrams are pairs of tokens, and trigrams are sets of three tokens. These terms are often used in natural language processing to refer to the length of the sequences of tokens being considered.
4. To generate n-grams from text, you can use a natural language processing library or toolkit that provides this functionality. Many such libraries and toolkits provide functions or methods that allow you to specify the length of the n-grams you want to generate and the text you want to process, and they will output a list of the resulting n-grams.
5. Lemmatization is the process of reducing words to their base forms. This is typically done by removing inflections and derivational affixes, so that words are represented in their dictionary forms. For example, the lemma of the word "jumping" might be "jump," and the lemma of the word "jumps" might also be "jump."
6. Stemming is a process that reduces words to their base forms by removing inflections and derivational affixes, but it is generally less accurate and more aggressive than lemmatization. Stemming algorithms often use heuristics and rules of thumb to identify and remove affixes, which can lead to the creation of non-existent words or the modification of existing words in ways that change their meanings.
7. Part-of-speech (POS) tagging is the process of marking each word in a text with its corresponding part of speech. This is often done as a preprocessing step in natural language processing tasks, as the part of speech of a word can provide important clues about its meaning and function in a sentence.
8. Chunking, also known as shallow parsing, is the process of extracting partial syntactic structure from a sentence. It involves dividing a sentence into shorter chunks or "constituents," each of which corresponds to a single grammatical unit, such as a verb phrase or noun phrase.
9. Noun phrase (NP) chunking is a specific type of chunking that involves extracting noun phrases from a sentence. Noun phrases are sequences of words that function as a single noun, and they often contain a noun or pronoun as their head word.
10. Named entity recognition is the process of identifying and classifying named entities in a text. Named entities are proper nouns that refer to specific persons, organizations, locations, or other entities, and they are often marked with special tags or labels in natural language processing tasks.