**NLP Assignment 2 Solution**

**1. What are Corpora?**

**Ans:** Corpora in natural language processing (NLP) are collections of written texts and spoken words that are used to train algorithms so they can better understand and interpret human language. They are made up of large bodies of text, audio recordings, and other types of data that are used to train algorithms to recognize patterns and features in language. Corpora are used to develop and improve models that can generate language and interpret it accurately. They can also be used to create resources like dictionaries, spell checkers, and machine translation systems.

**2. What are Tokens?**

**Ans:** A token is a single unit of a language like a word, number, punctuation mark, or any other symbol that is used to form a sentence. In Natural Language Processing (NLP), tokens are the basic building blocks of a document, and they are used to identify and analyze the structure of a piece of text. Tokens can be words, phrases, numbers, punctuation marks, symbols, and even emoji. They are used to identify the different components of a sentence, phrase, or document and to determine how they are related to each other.

**3. What are Unigrams, Bigrams, Trigrams?**

**Ans:** Unigrams are single words that are used to represent the meaning of a sentence. Bigrams are two successive words in a sentence that are used together to represent the meaning of a sentence. Trigrams are three successive words in a sentence that are used together to represent the meaning of a sentence.

**4. How to generate n-grams from text?**

**Ans:** N-grams are a type of text analysis technique used to analyze the occurrence and frequency of sequences of words in a text. To generate n-grams from text, you can use a variety of Natural Language Processing (NLP) methods.

1. Tokenization: Tokenization is the process of breaking a sentence into individual words or phrases. This is a key step in generating n-grams from a text.
2. N-gram extraction: After tokenization, you can use an NLP library to extract n-grams from the text. An n-gram is a sequence of n tokens (words or phrases) from a sentence. For example, a bigram (2-gram) is a sequence of two words from a sentence, such as “air pollution”.
3. Frequency analysis: After extracting the n-grams, you can use frequency analysis to determine the most commonly occurring n-grams in a text. This can help you identify key phrases and topics in the text.
4. Visualization: Finally, you can use data visualizations to represent the data. Common visualizations used with n-grams are word clouds, bar charts, and line graphs.

**5. Explain Lemmatization?**

**Ans:** Lemmatization is a process of text normalization in Natural Language Processing (NLP) which reduces words to their base form or root. It is very similar to stemming, but lemmatization uses an actual language dictionary to identify the root form of words. This process helps to reduce the number of unique words in the corpus, as well as helping to reduce the noise in the text. The output of lemmatization is usually more accurate than stemming and helps to improve the accuracy of the results of the NLP model.

**6. Explain Stemming?**

**Ans:** Stemming is a process in Natural Language Processing (NLP) used to reduce a word to its base form or root. This is done by removing suffixes and prefixes from a word, such as "-ing", "-ed", or "-ly". Stemming is commonly used in search engines, as a way to reduce multiple words with the same meaning to a single keyword. This allows a search engine to find all words with the same meaning and return them in the search results.

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

**Ans:** Part-of-speech (POS) tagging is a process of assigning a part of speech to each word in a sentence. This is done by a computer program, using algorithms and statistical models. The most common parts of speech are nouns, verbs, adjectives, adverbs, and pronouns. POS tagging can help to disambiguate words that have multiple meanings, as well as to assign correct grammar to a sentence. For example, the word "run" can be a noun or a verb, depending on the context. POS tagging can help to distinguish between the two meanings. Additionally, POS tagging can be used to create a more accurate machine translation of a text.

**8. Explain Chunking or shallow parsing?**

**Ans:** Chunking or shallow parsing is a process of parsing a sentence into small chunks of information, such as individual words or phrases. This process is often used in natural language processing (NLP) to quickly identify key words or phrases in a sentence. The goal of chunking is to simplify the processing of a sentence by breaking it down into smaller, more manageable pieces of information. This allows NLP algorithms to more accurately and quickly identify the meaning of a sentence.

**9. Explain Noun Phrase (NP) chunking?**

**Ans** Noun phrase (NP) chunking is a process of identifying and segmenting individual phrases in a sentence into their respective noun phrase components. This is an important part of natural language processing (NLP) because it helps to identify the main topics or concepts of a sentence and can be used for summarization and question answering. The goal of NP chunking is to automatically identify noun phrases in a sentence by using a set of heuristics or rules that define the syntax of noun phrases. This can be done using rule-based methods or machine learning algorithms.

**10. Explain Named Entity Recognition?**

**Ans:** Named Entity Recognition (NER) is a process in Natural Language Processing (NLP) of locating and classifying named entities in text. NER is used to identify the names of people, places, organizations and other entities in text, and classify them into pre-defined categories. For example, a sentence like "John went to the mall" could be labeled with the categories "Person" (John) and "Location" (the mall). NER is an important step in many NLP tasks such as question answering, topic segmentation, and document summarization.