**NLP Assignment-2**

1. **What are Corpora?**

A corpus is a large and structured set of machine-readable texts that have been produced in a natural communicative setting. Its plural is corpora. They can be derived in different ways like text that was originally electronic, transcripts of spoken language and optical character recognition, etc.

1. **What are Tokens?**

Tokens are the building blocks of Natural Language. Tokenization is a way of separating a piece of text into smaller units called tokens.

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

A 1-gram (or unigram) is a one-word sequence. ... A 2-gram (or bigram) is a two-word sequence of words, like “I love”, “love reading”, or “Analytics Vidhya”. And a 3-gram (or trigram) is a three-word sequence of words like “I love reading”, “about data science”

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

Add the Extract N-Gram Features from Text module to your experiment and connect the dataset that has the text you want to process.

For Text column, choose a column of type string that contains the text you want to extract.

By default, the module selects all string columns. However, because the result are verbose, you might need to process a single column at a time.

For Vocabulary mode, select Create to indicate that you are creating a new list of n-gram features.

For information about how to update an existing set of n-gram features, see this section.

For N-Grams size, type a number that indicates the maximum size of the n-grams to extract and store.

For example, if you type 3, unigrams, bigrams, and trigrams will be created.

For K-Skip size, type the maximum number of characters that can be different when identifying variants of n-grams. If the value of k is set to 0, n-grams can be created only from a unique, contiguous sequence of characters.

The option, Weighting function, is required only if you merge or update vocabularies. It specifies how terms in the two vocabularies and their scores should be weighted against each other.

For Minimum word length, type the minimum word length of strings that can be analyzed.

For example, assume the minimum word length was set to 3 (the default value), and you had one input that had a single word, and another that had some short text like "nice place". Both rows would be ignored.

For Maximum word length, type the maximum number of letters that can be used in any single word in an n-gram.

By default, up to 25 characters per word or token are allowed. Words longer than that are removed, on the assumption that they are possibly sequences of arbitrary characters rather than actual lexical items.

For Minimum n-gram document absolute frequency, type a number that indicates the minimum occurrences required for any single word or token to be included in the n-gram dictionary.

For example, if you use the default value of 5, any n-gram or skip-gram must appear at least five times in the corpus to be included in the n-gram dictionary.

For Maximum n-gram document ratio, type a number that represents this ratio: the number of rows that contain a particular n-gram, over the number of rows in the overall corpus

Select the option Normalize n-gram feature vectors if you want to normalize the feature vectors. When you do this, each n-gram feature vector is divided by its L2 norm.

Normalization is used by default.

Set Use filter-based feature selection to True if you want to enable additional options for managing the size of your text feature vector.

Run the experiment.

1. **Explain Lemmatization**

Lemmatisation (or lemmatization) in linguistics is the process of grouping together the inflected forms of a word so they can be analysed as a single item, identified by the word's lemma, or dictionary form.

1. **Explain Stemming**

Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma.

1. **Explain Part-of-speech (POS) tagging**

It is a process of converting a sentence to forms – list of words, list of tuples (where each tuple is having a form (word, tag)). The tag in case of is a part-of-speech tag, and signifies whether the word is a noun, adjective, verb, and so on.

1. **Explain Chunking or shallow parsing**

Shallow parsing (also chunking or light parsing) is an analysis of a sentence which first identifies constituent parts of sentences (nouns, verbs, adjectives, etc.) and then links them to higher order units that have discrete grammatical meanings (noun groups or phrases, verb groups, etc.).

1. **Explain Noun Phrase (NP) chunking**

Text chunking is dividing sentences into non-overlapping phrases. Noun phrase chunking deals with extracting the noun phrases from a sentence.

1. **Explain Named Entity Recognition**

Named Entity Recognition is one of the key entity detection methods in NLP. 2. Named entity recognition is a natural language processing technique that can automatically scan entire articles and pull out some fundamental entities in a text and classify them into predefined categories.