**EXPERIMENT NO - 4**

**AIM:** Write a python script to implement N-Grams

**THEORY:**

**N-GRAMS:**

N-grams are contiguous sequences of n items (words, characters, or other units) extracted from a larger sequence of text. The value of n is typically 1, 2, or 3, although it can be any positive integer.

For example, in the sentence "The quick brown fox jumps over the lazy dog," some possible 2-grams are "The quick," "quick brown," "brown fox," "fox jumps," "jumps over," "over the," "the lazy," and "lazy dog."

N-grams are commonly used in natural language processing (NLP) tasks such as language modeling, text classification, and machine translation. They can provide useful information about the structure and context of text and can be used to generate probabilistic models of language.

**PROGRAM:**

**Text

Description automatically generated**

**CONCLUSION:**

The N-gram model is a statistical language model that has various applications in natural language processing such as spell correction etc. The bi-gram model specifically has been studied in detail and has been implemented.

**EXPERIMENT NO - 5**

**AIM:** Write a python script to implement N-Grams Smoothing

**THEORY:**

**N-GRAMS:**

N-grams are contiguous sequences of n items (words, characters, or other units) extracted from a larger sequence of text. The value of n is typically 1, 2, or 3, although it can be any positive integer.

For example, in the sentence "The quick brown fox jumps over the lazy dog," some possible 2-grams are "The quick," "quick brown," "brown fox," "fox jumps," "jumps over," "over the," "the lazy," and "lazy dog."

N-grams are commonly used in natural language processing (NLP) tasks such as language modeling, text classification, and machine translation. They can provide useful information about the structure and context of text and can be used to generate probabilistic models of language.

**N-GRAMS SMOOTHING:**

N-gram smoothing is a technique used to address the problem of sparse data in natural language processing (NLP) tasks. When using n-grams for language modeling, it's possible to encounter n-grams in the test set that were not present in the training set. This can result in zero probabilities for these n-grams, which can lead to inaccurate language models.

N-gram smoothing methods attempt to adjust the probability estimates of unseen n-grams to avoid zero probabilities. One common method is called add-one smoothing, where one is added to the count of each n-gram before computing the probability. Another method is called interpolation, where the probability of an n-gram is a weighted sum of its maximum likelihood estimate and the probabilities of its sub-ngrams.

The goal of n-gram smoothing is to produce more accurate language models that are less susceptible to overfitting to the training data. By adjusting the probabilities of unseen n-grams, the model can better generalize to new and unseen text.

**PROGRAM:**

Text, letter

Description automatically generated

**CONCLUSION:**

The N-gram model is a statistical language model that has various applications in natural language processing such as spell correction etc. The bi-gram model specifically has been studied in detail and has been implemented.