Alekhya Pinnamaneni
axp190109
CS 4395.001

Ngrams

An N-gram is a set of N consecutive words in a text. They are used to create a probabilistic language model by capturing the most likely sequences and groupings of words.

N-grams can be applied and used for many purposes, such as analyzing text, generating text, classifying text, sentiment analysis, and language detection.

For unigrams, the probability is calculated by dividing the number of occurrences of the unigram by the total number of unigrams in the vocabulary. For bigrams, the probability is calculated by multiplying the unigram probability of the worst word in the bigram by the unigram probability of the second word in the bigram.

The source text that is used for building the language model is very important because it determines how useful the model will be when used on other texts. A source text that is very small and specific will not generate a very useful model.

Smoothing removes the sparsity problem to minimize the possible n-grams for consideration. A simple approach to smoothing is Laplace smoothing, which simply adds 1 to the counts used in the probability calculation.

Language models can be used for text generation by using a naïve approach. This involves finding the next most likely word based on a given start word until the sentence end is reached. However, this is a naïve greedy approach with limited accuracy in its results.

Language models can be evaluated extrinsically, for example by humans, or intrinsically, used internal measures such as perplexity (PP). Lower perplexity indicates a better language model.