N-Grams Narrative

N-Grams are essentially sliding windows of size n over a text, start and stop symbols could be included. It can create a probability score to the next word based off considering previous words. The model is trained off a corpus of text and can then be used for a variety of different applications. N-grams could be used for speech recognition, where it can detect what language is being spoken, predicting text input, to determine the most likely words the user is going to say next, spelling detectors, and many more applications. There are different probabilities that can be used to calculate the probabilities for unigrams and bigrams. The way I used to calculate the probabilities was using LaPlace smoothing. For unigrams, the probability was simple enough, it was the word in the text with the count of how many times the word appeared in the text. For Bigrams, the count of the bigram in the text is calculated along with the count of the first word of the bigram. These two values are then used to plug into the formula to get the probability: (Count of Bigram + 1) / (Count of First word in Bigram + Number of Unique Tokens). Selecting the right source text is crucial to having a good model and getting good predictions. The text must be relevant to what the model is trying to predict. Both being English is not enough, they have to have the same style and grammar as well. Smoothing allows the gaps in the text to be covered as there's going to be bigrams where the list is nothing so to counteract that, we can add 1 to the count of the bigrams. This allows every bigram to be counted for at least once in the text. N-grams can help us in generating text by giving us the highest probability of a word occurring next after the previous words have been stated. This is obviously going to be limited to the corpus

and the quality and size of the size text. The bias of the corpus can influence the model greatly. Language models can be evaluated with test datasets, that already have correct output values. The Language model can be run on this test dataset, and then be judged on accuracy based off its predictions with the correct output results. Google's N-gram viewer is a data visualization tool that allows you to search for the frequency of desired ngrams based off that current year.