Assignment 5: Narrative

A) A n-gram is a sliding window over text that looks at n words at a time. If it takes one word at a time, it is called a unigram. A bigram takes two words at a time, and trigrams take three words at a time. Anything above 3 is usually just refered to as a n-gram, where n is specified as a specific integer. N-grams can be used to build a probalistic based model, where we use existing words to calculate the probablity of a word showing up after them.

B) A n-gram can be used in any application where the goal is to predict what words might show up next given an existing series of words. Some examples include spelling correction, speech recognition, machine translation, and auto suggestion.

C) Probablity of a unigram is simply the percent of of many times the unigram occurs divided by the total number of unigrams. Bigrams are slightly more complicated, for example the probability of P(jack, jump) = P(jack) * P(jump|jack). This helps to increase the percent for two words that are often seen together.

D) The source of training data used on n-grams is highly important. The probablistic values of a n-gram are calculated using the relational ship of words in the training data. For example, a model trained on Shakesphere will probably have a hard time predicting text for a fantasy novel accurately, since that language and style is vastly different.

E) Smoothing is one possible solution to sparsity problem. When we are computing the probabiliy of a word, we may encounter a word that does not have a count since it did not apprear in the training text. Trying to find the probability for this world will yield a 0, and zero out the rest of the probabilities when we multiply. To avoid this, we use smoothing techniques. A very simply smoothing technique would be to fill in 0 value probabilities with a very small percent of the overall mass. This way we can avoid 0's and the probability from having a large effect on the overall probability.

F) We can use this probalistic model for language generate. By passing in a start word of phrase, we can find word of phrase that follows it with the highest probability. Then we simply add this text to our initial input text. This method is generally limited by its small size and simple approach. Using a trigram or higher n-gram model might prove to be more effective. This method is also heavily reliant on the training source.

G) The effectiveness of a n-gram language model can be evaluated using intrinsic evaluaions such as perplexity. The perplexity is the inverse of seeing the words we observe, normalized by the number of words. Perplexity is simply an exponentiation of low entropy. If a language model has low perplexity, it means there is less chaos in the data and our model is able to make meaningful non-random prdictions.

H) The Google Ngram Viewer is a search tool that lets you search books and research articles for a specified word of phrase of text. You can use it by going to https://books.google.com/ngrams/ and searhcing for any example text. You can then specify what year you want the text to be from. It will then give you a list of results