Zach Allen & Nebil Weber
CS 4395.001 Human Language Technologies
Dr. Karen Mazidi

N-grams Report

A. N-grams are a sequence of any number (N) of tokens in a written form of human language. They can be used to predict the following token from a given set of tokens. N-grams can be used to build a language model by analyzing the given tokens and assigning a probability to the options available for the following token, showing what is most likely to come next.
B. N-grams could be used in applications involving text prediction such as Apple iMessage or Microsoft Word. Spelling/grammar applications such as Grammarly could also benefit from N-grams.
C. For unigrams, probability is calculated by taking the (count of the occurrence of the token) / (total number of tokens). It is just the probability of picking the token from the document. For bigrams, the probability is calculated by (count of both tokens occurring in succession) / (count of the first token). It is the probability of the tokens occurring together divided by the count of the first token in the document.
D. When building a language model, the source text is very important. If the source text is poorly written or has many grammatical errors, the probabilities will not be very accurate. With a well written source text, the probabilities will be more accurate, and the N-grams program will have a better chance of correctly predicting the output.
E. Smoothing is important in N-grams programs because not every N-gram will have occurred in the training set, so the program will not be able to set a probability for the following token. Smoothing will fill in the missing probability with a value so the program will be able to better predict output. A basic approach to smoothing is Modified Good-Turing Smoothing. This replaces N-grams with 0 probability with the probabilities of words that only occur once.
F. Language models can be used for text generation because these models are able to predict tokens for output. The text can be generated from these models based on how well the source text is written/how large the source text is. This is the limitation of models in text generation, as it depends on the source text of the model.
G. Language models can be evaluated either by humans (usually linguists), or through computational methods such as perplexity.
H. Google's N-gram viewer shows the probability of N-grams on the y-axis and a timeline on the x-axis. It is a way to view N-grams in books from 1800-2019. Here is an example, with Titanic and Tom Brady.