Austin Girouard and Aarya Patil

ATG180001 / AAP180014

Dr. Karen Mazidi

Assignment 4: Ngrams

Ngrams are contiguous sequences of N words that are used in natural language processing (NLP). Ngrams are a fundamental concept in language modeling, which is a way of modeling the probability distribution of sequences of words in natural language based on the frequency of each sequence's occurrence in a corpus of text (training data). Besides their use in building language models, Ngrams are also useful in spell checking applications, sentiment analysis, and text classification/generation. Calculating probabilities for unigrams is simple; the formula is the number of occurrences of the word divided by the total number of words in the corpus. Bigrams must be calculated by considering the probability of a word, x, while considering the word that precedes it, y. The formula is dividing the number of times the bigram (y, x) occurs in the corpus by the number of times the word y appears as a preceding word in the corpus. The source text is very important when building a model, because any errors in the source text will lead to inaccuracies in all text analytics models that are trained on it, irrespective of its application. Smoothing is important for handling probabilities that are extremely low. These low probabilities make it more difficult for a language model to make accurate predictions of words. A simple approach to smoothing is Laplace smoothing. Text generation can be done using language models by making predictions on the most likely set of words that follows a given set of words. This can be done using Ngrams, as mentioned above. One limitation of this approach is that it does not have the ability to generate text beyond what it was trained on.

Another is that it may generate nonsensical text, such as when you let word prediction on a texting app generate an entire sentence based on single-word predictions. Language models are commonly evaluated by a metric known as perplexity, but can also be evaluated using cross entropy and bits-per-character (BPC). Google's Ngram viewer shows a graph displaying the frequency of phrases as they have appeared in a corpus of books across a specified time frame. For example, here is a graph detailing change of occurrences in "The Flu" and "The Black Plague" throughout history. You can see a sudden spike in "The Flu" around the 1920's due to the Spanish Flu of 1918.