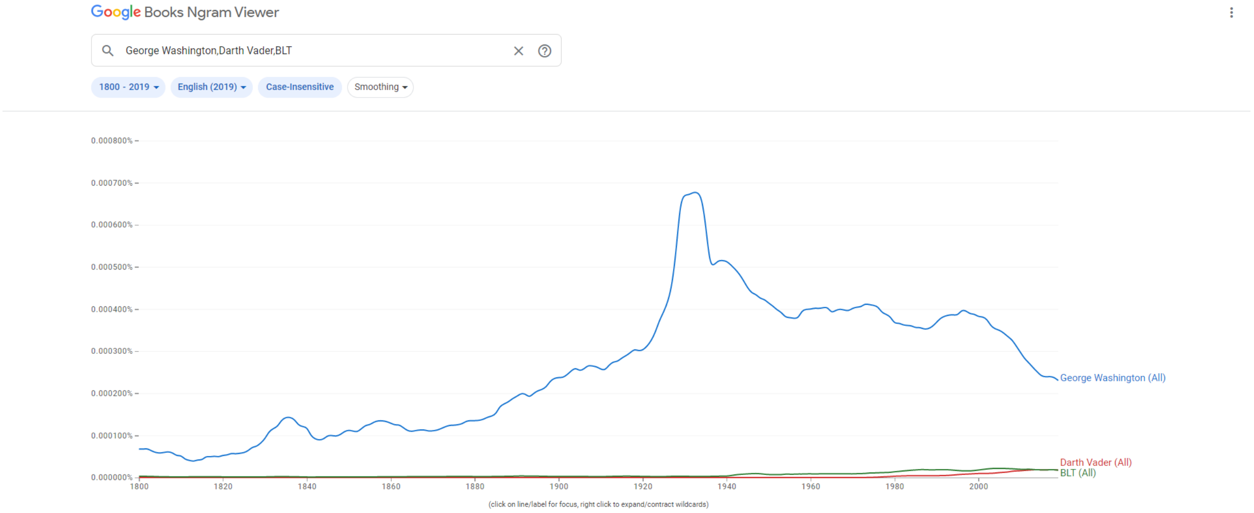
Joshua Olowoyeye, Bill Dozier
CS 4395.001

An n-gram is a sequence of items from a sample text or speech corpus that can be used as the base of a language model. It functions as a sliding window over a body of text n words at a time. They can be used as a language model to predict the probability of a specific n length sequence of words appearing in any sequence of words in the chosen language. It is important to keep in mind that the chosen corpus has influence over how well a model will perform. N-grams can be very useful in a handful of applications, such as generating probabilistic models to auto complete statements, for spelling correction, text summarization, and part of speech tagging. To calculate the probabilities of a unigram occurring in a corpus we can just take the number of unigrams of a specific word and divide it by the total number of unigrams (P(w) = number of occurrences of w in unigram list / total number of unigrams). Calculating bigrams follows a similar formula where it is the probability of the first word as a unigram times the count of the bigram divided by the count of the first word (P(w1, w2) = P(w1) * P(w1 | w2) = P(w1) * count of bigram / count of w1 as unigram). As stated briefly before the source text of an n-gram model is of great importance to the accuracy and performance of the model. Using poetry as the source text for an anatomy based language model would likely result in lackluster accuracy. This leads into the idea of smoothing, where even with an appropriate corpus not all word sequences can possibly be captured. Looking at the calculated probabilities this could result in issues stemming from a count of zero for some specific word. This will lead to a zeroing out of the probability and is referred to as the sparsity problem. To deal with this an approach called smoothing is potentially applied. Smoothing is the act of replacing zeros with a little bit of the overall mass. A simple example would be Laplace smoothing which applies an addition of 1 to all of the counts ahead of time but also adds the total vocabulary count to the denominator to balance this change out. While Laplace smoothing is simple it isn’t very effective in its performance because it tends to be quite aggressive in its adjustments of probabilities. N-grams and their probabilities can be used for text generation given a start word. Doing so uses the probabilities of the bigrams in a model to concatenate the highest probability next word based off of a starting word in the bigram. This generation continues until a last token is added, generally being a period. With a limited corpus this method is itself quite limited in the results it can return and using it with bigrams specifically would not yield as good results as higher order n-grams. The main limitation is the size of the corpus required to yield viable results from this method, without it the generation is not exactly inteligible or diverse. Language models are evaluated using an intrinsic evaluation metric known as perplexity. Perplexity measures how well a language model predicts the text in the test data, where some of that data was set aside specifically for this calculation. The formula for perplexity is the inverse probability of seeing the words, normalized by the number of words. Perplexity is an exponentiation of the number of bits needed to encode the information in a random variable or the entropy. We strive for low perplexity, or entropy in the data which essentially means less chaos in the data. Perplexity can be seen as the number of branching choices we have moving from word to the next. Google has an n-gram visualization that uses the vast corpus of the Google Books platform to track the use of specific n-grams in literature from the 1950s to 2019. The system displays a graph showing how terms have been used in that period by showing the percentage of the total n-grams for which that n-gram encapsulates. For example we can compare the percentages of George Washington, Darth Vader, and BLT over the timespan and see that the latter two lack the same reach as Washington does in the corpus. Likely because George Washington was the first president of the United States while Darth Vader is a character from a 1977 film and a BLT is a (debatably popular) sandwich.



Google Books Ngram Viewer

George Washington,Darth Vader,BLT

1800 - 2019 ▾ English (2019) ▾ Case-Insensitive Smoothing ▾