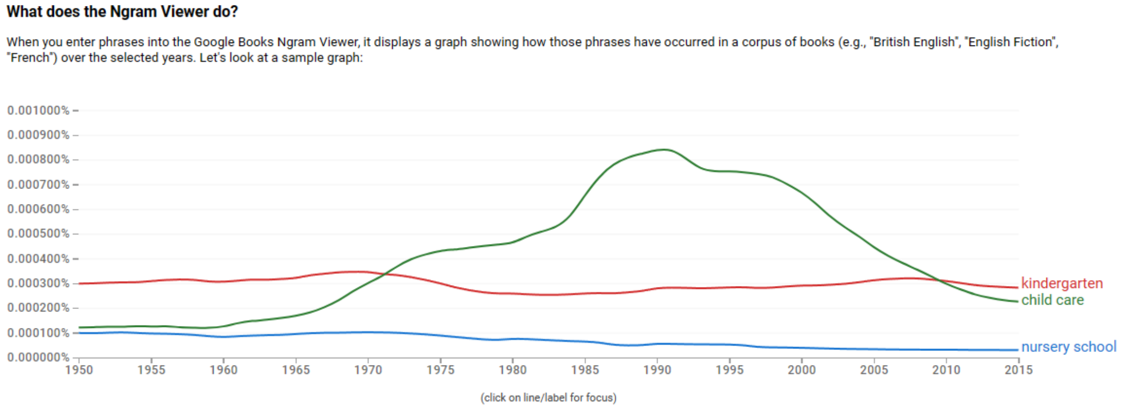
Cris Chou
CS 4395

N-grams are used to get the frequency of a word and analyze it through pattern recognition. They are a probabilistic model which requires a large amount of corpus data to train. n-grams can be used in speech recognition software, online translation services, and auto correctors. Unigrams are the frequency that a word appears in a given text. Bigrams are the frequency that a pair of words appears together in a given text. To calculate this probability, P(go | lets) would be the probability of the word “go” appearing after “lets”. You would first calculate the probability of the unigram “lets”, then multiply it by the probability of the P(go | lets) to get the probability of the bigram “lets go”. Source text is extremely important for building a language model because it is the data that is training the model, meaning bad input becomes bad output. Having good, accurate, and robust data, will likewise make the model more accurate and robust. Smoothing helps eliminate zero probabilities, which would mess up probabilistic models, when having to multiply by zero. It also helps with the *sparsity problem* of a corpus being unable to contain every single combination of words. A simple idea to smoothing is to simply add 1 to the count rather than zero, known as add-one smoothing. Language models can sometimes generate text by attempting to predict the next word in a sequence based on previous words. A naive approach would be to look through and see which bigrams include the current word. This works better on smaller sentences, but with a longer string of words, the generated text may not be coherent. Language models can be evaluated based on their perplexity. Human annotators can give an evaluation based on certain metrics. However, this takes too much time and resources, and an easier method is to use intrinsic evaluation. Using perplexity as a metric, one could compare different language models to each other. Google’s n-gram viewer shows the frequency of unigrams and bigrams throughout history, as far as they have data for.



What does the Ngram Viewer do?

When you enter phrases into the Google Books Ngram Viewer, it displays a graph showing how those phrases have occurred in a corpus of books (e.g., "British English", "English Fiction", "French") over the selected years. Let's look at a sample graph: