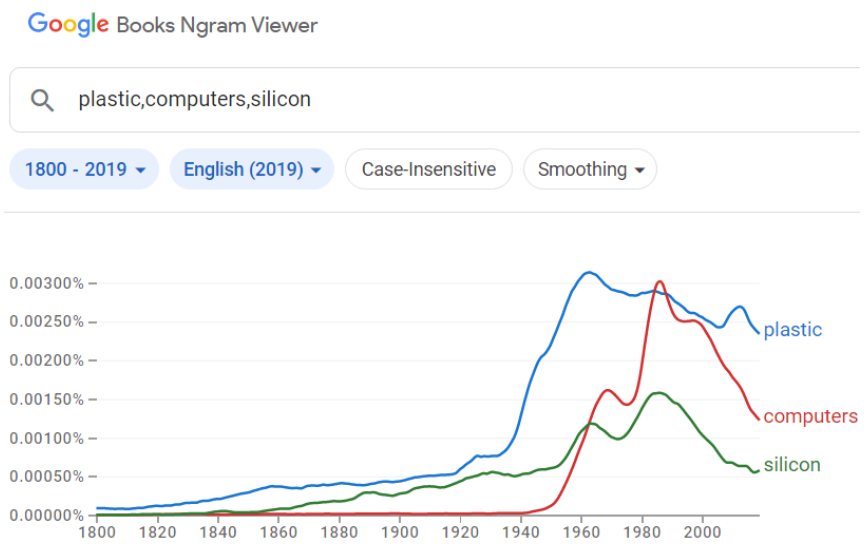
My Experience with N-grams

N-grams are considered to be a “window” into a given amount of text. They are typically separated into single tokens or pairs, and can be much larger than that. A popular usage of n-grams is when creating language models and needing to determine probabilities. Dictionaries can be created with these n-grams and a count can be created to help see how often certain combinations appear in a given text. Like with the n-gram program I had written, they can be used to detect certain languages which can be useful with translator apps. And given a chance at predicting text, such as when someone is typing, the computer can try to predict what will be said based off a large corpus. Probabilities can be calculated in different ways for unigrams and bigrams. Good-Turing smoothing, LaPlace smoothing, and logarithmic probabilities are all different ways for determining probabilities. The source text can be incredibly important when trying to build a language model. If the data being used is not accurate or contains issues within the text, the language model will be less accurate and useful. Smoothing is useful for filling in empty spots in the dictionaries being created. Something like LaPlace smoothing can be useful to help keep probabilities from becoming inaccurate. It simply adds 1 into the probability equation to account for missing counts. Language models can attempt to generate text based off its training data. But in order for it to do a good job, a large and accurate corpus is necessary. The main downside is that it only knows what it is given, which is why a smaller corpus will be very limiting. Language models can be evaluated on complexity and accuracy mainly. Google has created an N-gram viewer to graphically represent the occurrence of given phrases throughout written and recorded texts. It is helpful for knowing phrase usage especially

for a given time period and having a sophisticated program is useful here. An example is shown below:



It can be seen that plastic, computers, and silicon all had a rise from about the 1960s to the present. This is due to the rise in computing and materials used for it. A n-gram viewer like this is quite useful for people studying Natural Language Processing.