N-grams Narrative

A. N-grams are a sequence of words from a text, where the n represent the number of words in this sequence. They can be used to estimate the probability of words (or sequences) occurring. To do this, you first calculate the frequency of each sequence in a large text (called the corpus) and then the frequency can be used to generate and/or identify new text.

B. N-grams can be used to predict a sequence of words (used for applications like speech recognition). They can also be used for applications like spell-checking and text classification.

C. To calculate the probability of a unigram, you simply divided the amount of times the unigram occurs by the total number of words in the training text (corpus). To calculate the probability of the bigram, you divide the number of times the bigram appears and divide it by the number of times the first unit of the bigram occurs in the training text (corpus).

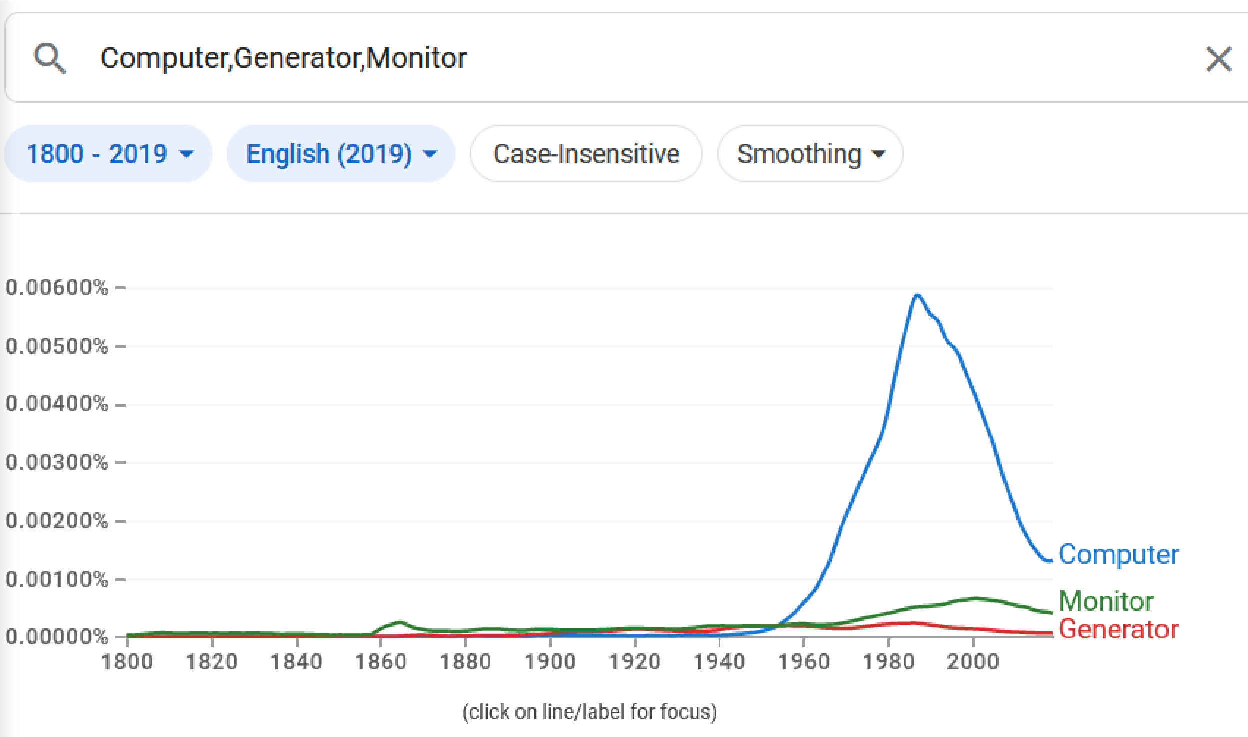
D. The source text is the basis of probabilities of n-grams. It can change factors in the language model like its quality and accuracy. Trying to use a source text which is different from its end application can give unoptimal predictions as the source text has different qualities from the application which the model is to be used.

E. Smoothing is a technique which is used to get rid of edge cases (like when real data has not been seen in the source text). Smoothing can help resolve cases that are predicted to have a zero probability of occurring and can be done by simply adding some small value to the count of each n gram in source text so that the probability is no longer zero.

F. Language models can be used for text generation by using the probabilities they have already calculated to generate new sequences of text. Some limitations include the text not making sense semantically, grammatically, or even be biased. Some of these limitations can be solved with post-processing and being careful with what is chosen for the source text, however, it is hard to make the text always be coherent.

G. A language model can be evaluated in categories like accuracy, human evaluation, and other forms of calculated scores. There is a formula for perplexity (which is how well the model predicts the source text) which is also a good indicator. Humans can also evaluate the generated text on aspects like quality and coherence (although this method may be expensive in terms of cost as it can be time-consuming).

H. Google's n-gram viewer is an online tool that lets people view the frequencies of n-grams (which they specify) across several centuries. The large corpus includes millions of books over this time period. This can help visualize trends in language over time. Here is an example of the visualization it produces for the words computer, generator, and monitor:



Computer,Generator,Monitor

1800 - 2019 English (2019) Case-Insensitive Smoothing