N-Grams in Language

N - a stand in for a natural number, being a positive integer.

Gram- A unit, in the context of language, a word.

Together, it means a unit of a certain number of words. A unigram is one word, a bigram is 2, and up to N. Ngrams are an important tool when thinking about language as being probabilistic- certain words are more likely to appear than others, and certain sequences are likely to reappear. By using ngrams, you can think of language as a Markov chain, where the next word is a probabilistic function of the words that preceded it.

Ngrams, thus, are useful in multiple scenarios. One simple use is the subject of this project- identifying the language a sample was written in. More in depth, if you have the corpus of a particular author, you can identify who wrote a specific piece even. It can also to some extent be used to generate language, if you use long enough ngrams.

The probabilities of n-grams is calculated by using a large corpus of input text, and locating every appearance of every bigram, and counting their frequency, which can be used to calculate the probability. With a large and diverse enough sample, you can generate a fairly exhaustive list, but never one that is fully complete.

Source text is important- when using bigrams for example, the subject matter of the source will create different frequencies. For example, if you were to have one source of text from he 50s, you might find ‘record disk’ to be the common bigram, while ‘floppy disk’ might be it in the 80s.

Additionally, lots of words only appear in some text sources, which means any model will have gaps. This is why smoothing is necessary- you can’t insert a zero, or it will dominate any probability arithmetic. Smoothing instead replaces the zero with a small value so as to not utterly dominate calculations.

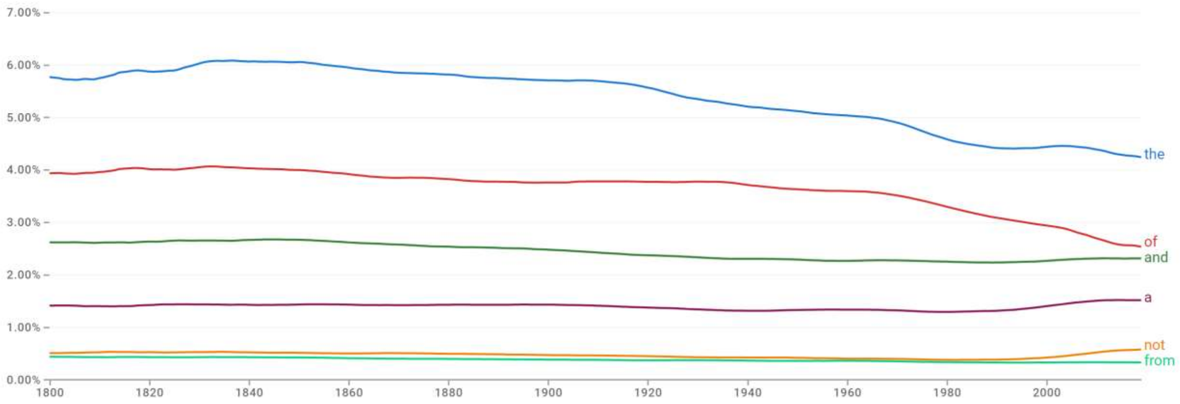
As previously described, Ngram language models can be used generatively, particularly with larger grams. The limitation of this however is that the underlying model while able to predict reasonable next words will not have an understanding of the subject matter, and thus without sufficiently long n-gram analysis will appear to have no connected train of thought, changing subjects and meandering.

The general idea in evaluating these models is that a higher probability should be given to sentences that are ‘valid’- eg, that follow grammatical and syntactic norms, and makes sense. It can also be evaluated by removing words from an existence sentences, and having it fill in gaps and checking if it matches the sample.

Finally, there is Google's N-gram viewer, which shows the frequency of n-grams in google's massive aggregated corpus over time in the form of a line graph. It is useful for monitoring how language has changed over time.

As an example, I chose some of the most common unigrams in English- the of not from and a

Supprisingly, there is a tren that the has declined from 6% to 5% of all words used since 1840.


the
of
and
a
not
from