Ngrams

a. N-grams are sliding windows of size N over text. A unigram has size 1, [cat] would be a unigram. A bigram is size 2, [cat has] is a bigram, and so on. They are used to build a probabilistic model of language.

b. N-grams could be used in language parsing, data compression, computational biology, and communication theory.

c. Probabilities are calculated through a variety of methods. The basis is that the probability of two words appearing alongside one another is the probability of one word given the other times the probability of the other word, shown as $P(w1)P(w2|w1)$. A larger sequence of words can be as follows (under a Markov assumption): $P(w1)P(w2|w1)P(w3|w2, w1)P(w4|w3, w2, w1)$. There are further expansions on this through the Good-Turing, Laplace, and logarithmic smoothing methods.

d. Source text is used to base patterns off of and is extremely important. The source text teaches the model what words and phrases may or may not appear in a given language. Without this basis, the model is lacking a significant amount of information leaving its further conclusions less grounded than before.

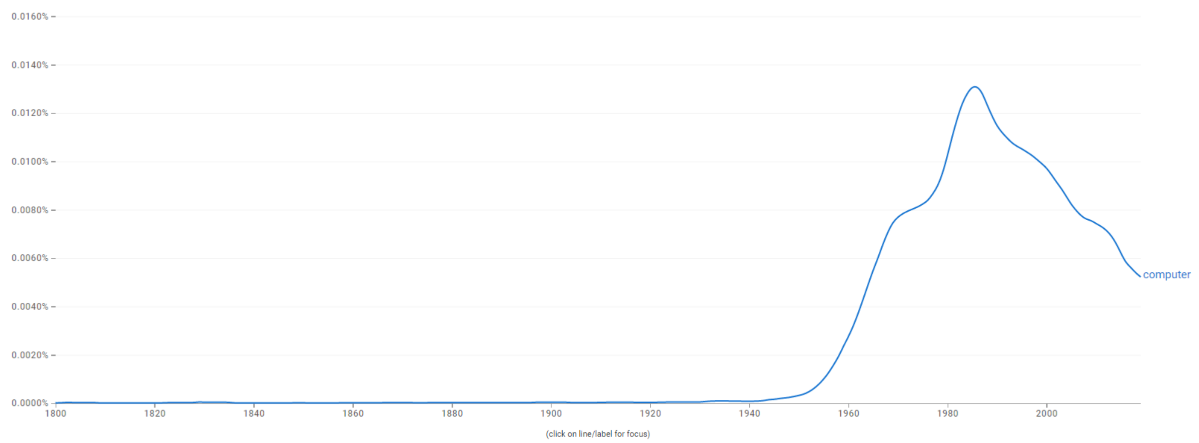
e. The importance of smoothing is to reduce stark contrasts in words or phrases with 0 occurrences. Smoothing is the slight changing of data to keep the overall model more in line with itself. An example of this is adding a + 1 to formulas where a variable could be 0, since dividing 0 by anything is still 0, leading to a loss of information. A 1 is close enough and still represents some data, reducing variance that is driven by the presence of 0s.

f. Language models can use its stored values of occurrence to attempt sentence creation, but it is initially greatly limited by amount of sample data, and later limited by not having a deeper understanding of language and instead relying solely on given patterns. This still reduces its scope of output, and also may fail to construct possible sentences with its known vocabulary.

g. Language models can be evaluated by both extrinsic and intrinsic metrics. Extrinsic evaluation could include human annotators, who judge and analyze model output based on their natural understanding of a language. Intrinsic evaluation could include perplexity identifiers. Perplexity, in this context, can be defined as the inverse probability of seeing the words that are observed, normalized by the number of words. A low perplexity is ideal.

h. Google's Ngram Viewer has access to one of the largest sets of data in existence, making it a fairly accurate and appealing tool in the world of N-grams.

Here is a sample image with example word "computer":



The probability that the word “computer” was used by itself or in a phrase was highest in 1985, an age where breakthroughs and discoveries surrounding computers opened our horizons and brought a new age of technology into the world. It wasn’t necessarily when computers themselves were the most popular, but it was when it was (relatively) very common to be written or said.