Jonathan Ho

CS 4395

Ngram Narrative

a. Ngrams are the number of N words in sequence (example being a bigram being two words). By determining the probability of a ngram with a corpus, a language model can be trained to either help generate language, or to determine the semantics of a text.

b. Ngrams can be used in a variety of ways such as a translating a language into another, suggested searches, and spell-checking within a program.

c. Probabilities of a calculating a unigram is taking the count of all the unigram's occurrence within a text and divide it by the total number of words in the text. For a bigram, it would be the count of the phrase appearing within the text divided by the total number of times the first word in the phrase appears.

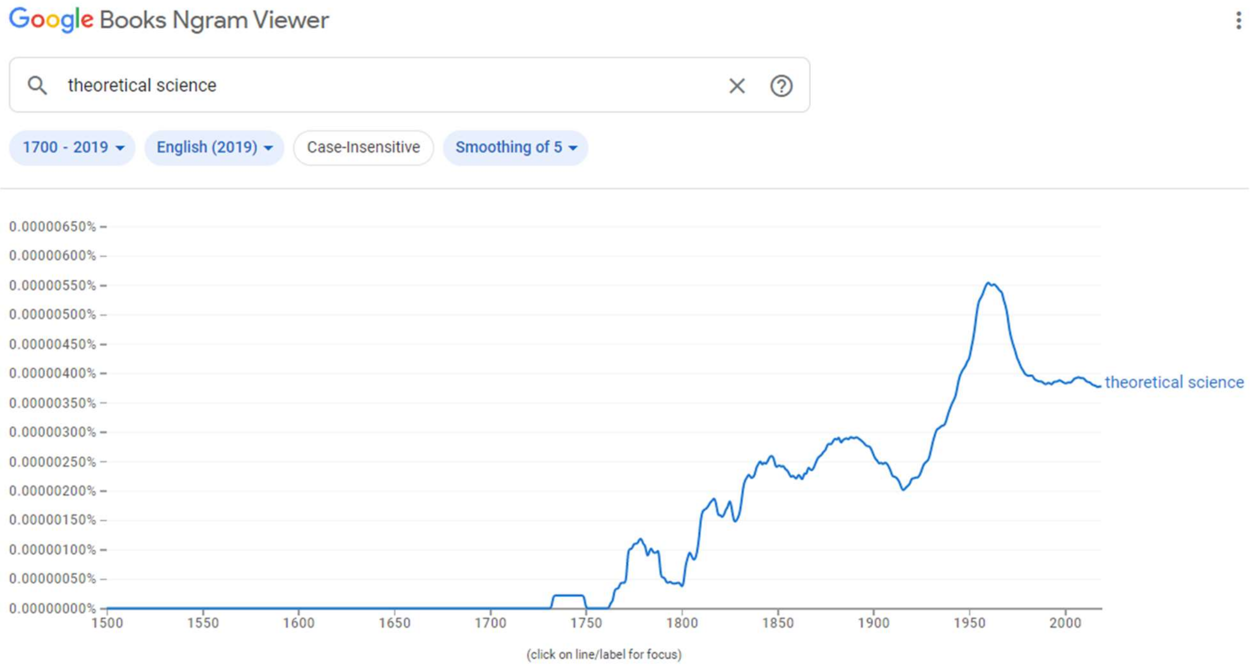
d. Based off the assignment we have done, the source text is important because this is how a computer can build a language model to interpret text similar to the source text. The English language model had its own set of unigrams and bigrams derived from the source text. If that were applied to the French or Italian languages, it would never be able to recognize it at all compared to any English texts.

e. When predicting probability of a bigram, smoothing is important since there will always be instances of text where a bigram would have a count of 0 since it may not show up. One simple way of smoothing is Laplace smoothing which is adding a 1 to any bigram count of 0.

f. Language models can be used to generate text by using bigram probabilities. If a start word is given, a computer can check for bigrams using the start word as the first parameter to look for a phrase that has the highest probability containing the start word and the second word in the bigram. This second word would then be inserted, and that now acts as the first word of the bigram to look for the next word likely to be generated. However, the limitation of this approach is the smaller the ngram, the less accurate the word generation would be.

g. Language models can either be evaluated extrinsically or intrinsically. Extrinsically could be humans evaluating the results of a language model given a predefined metric to check. Intrinsically would be comparing the language model to another more established model.

h. Google's n-gram viewer is a search engine that can find word frequencies of books printed between the years 1500 and 2019. It takes in an input of up to a 5-gram and finds the number of yearly appearances the ngram has had. Then, it divides each appearance by the range of years to look at. The result is a chart where you can see how frequent a word appeared in books based on the year. A smoothing value can also be chosen to make the distribution better as well as check a word regardless of capitalization. The picture below shows an example of looking at the ngram "theoretical science":



Google Books Ngram Viewer

theoretical science

1700 - 2019 ▾ English (2019) ▾ Case-Insensitive Smoothing of 5 ▾

(click on line/label for focus)