Aloksai Choudari

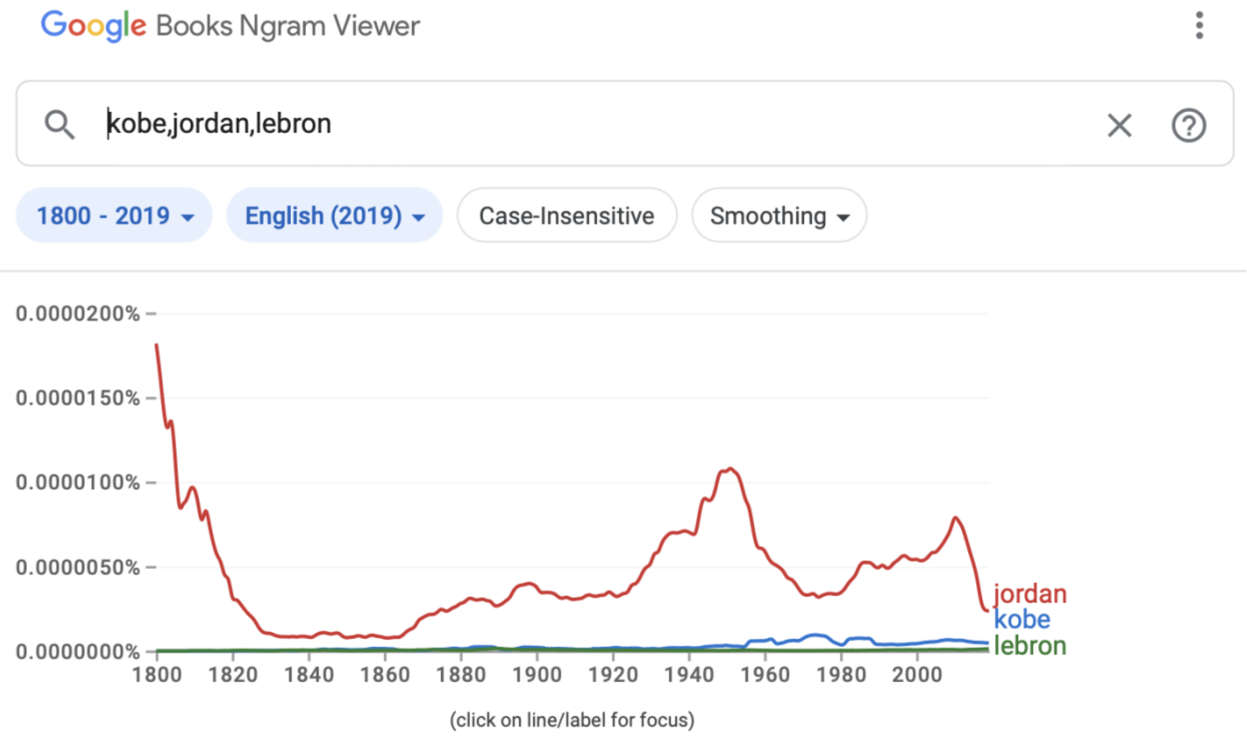
Dr. Karen Mazidi

CS 4395.001

4 March 2022

CS 4395 N-grams

N-grams are sequences of an n number of words from a given text that are used for building language models. They are used to build language models by counting the number of occurrences of each n-gram that comes from the text. This is used to guess the probability of a word showing up in the n-gram sequence. A few applications where n-grams could be used include speech recognition programs, spell checking programs, and language translation programs. The probability for unigrams is calculated through the formula P(a) = count(a) / N, and the probability for bigrams can be calculated using the formula P(a|b) = count(a,b) / count(a). In the first formula, the probability of a specific word in a text sequence is the number of times that the word appears in the sequence divided by the total number of words in the text. In the bigrams formula, the probability of a word given its previous word is the number of bigrams divided by the number of times the previous word appears in the text. The importance of source text in building a language model is that the model is only good as the quality and size of the text it is trained on. If the text is larger and more diverse, the model will likely be more accurate and stable. Smoothing is important for language modeling because it prevents n-grams from going unseen. An approach that is used for smoothing is add-one smoothing, which adds 1 to the n-gram count prior to calculation of the probabilities. Language models can be used for text generation based on the probabilities of a single word's occurrence given the words prior to it. The limitation to this is that there may be grammatical errors in the text generated if it has not been trained. Language models can be evaluated through their performance in modeling tasks through accuracy and precision. This will measure how well the model is able to predict text. Google's n-gram viewer is an NLP tool that gives users the capability to search and analyze frequency of occurrences of n-grams in large text. Attached is an image that shows an example of the usage of Google's n-gram viewer.



Google Books Ngram Viewer

kobe,jordan,lebron

1800 - 2019

English (2019)

Case-Insensitive

Smoothing

(click on line/label for focus)