Crystal Ngo and Kiara Madeam
Professor Mazidi

Narrative. Write a one-page or more narrative about Ngrams:

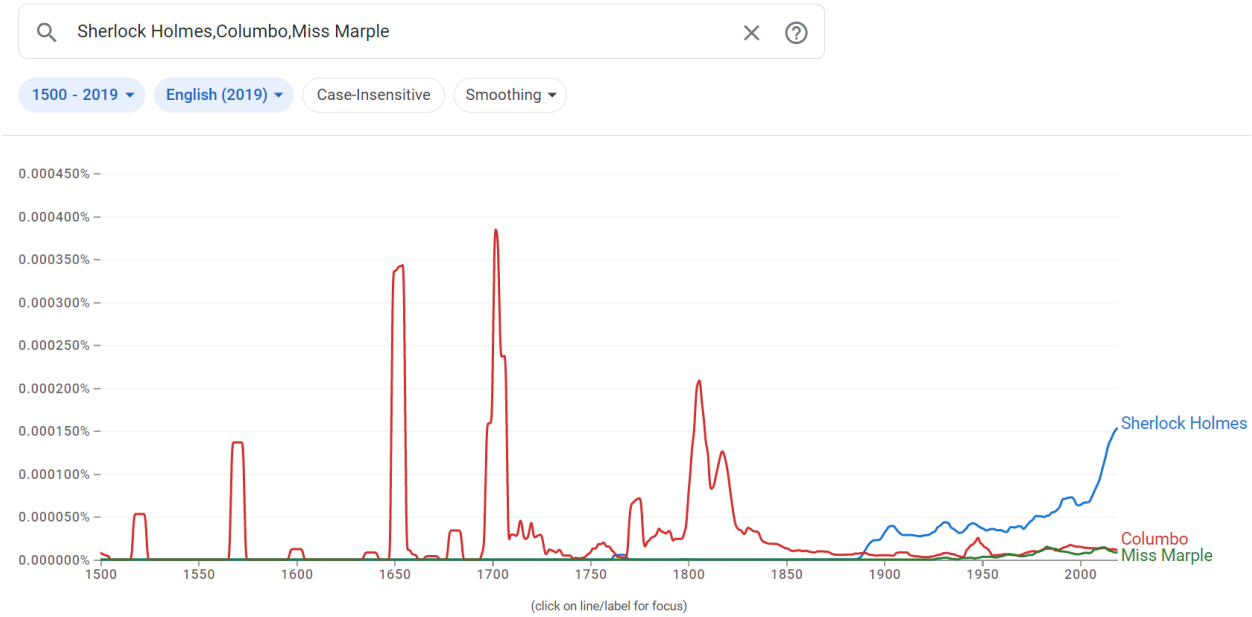
a. what are n-grams and how are they used to build a language model
b. list a few applications where n-grams could be used
c. a description of how probabilities are calculated for unigrams and bigrams
d. the importance of the source text in building a language model
e. the importance of smoothing, and describe a simple approach to smoothing
f. describe how language models can be used for text generation, and the limitations of this approach
g. describe how language models can be evaluated
h. give a quick introduction to Google's n-gram viewer and show an example

a. N-grams can be described as an N-sized selection of text from a string. They are useful for predicting how common certain text sequences are. This supports pattern recognition, which is important for language learning and identifying specific phrases.
b. N-grams can be used for things like error detection, text generation, or semantic analysis. Because of this, they would be useful in applications that need spell-check tools, spam mail or hate speech filters, or plagiarism detection.
c. For unigrams, probability is calculated by dividing the number of occurrences of each word by the total number of words. Bigram probability is calculated by dividing the number of times the bigram appears in the text by the number of times the first element of the bigram appears in the text.
d. The source text is qualitative information that a language model turns into quantitative text that a computer can understand. Although we can use a strictly rules-based approach, the reality of how people write is better learned by a language model using probabilistic methods. It is important to have text from a range of different sources so that the language model will be better at predicting text in the test data, also known as perplexity.
e. The importance of smoothing is that it simplifies noisy data so that things like language models have better pattern recognition. One simple approach to smoothing is modified Good-Turing smoothing, where zero counts or probabilities are replaced by the counts or probabilities of words with single occurrences.
f. Language models can be used for text generation, for example, by converting N-grams to probabilities and choosing the next n-gram to be displayed with respect to whichever has the highest probability of appearing next given the previous phrase. This method of text generation has its limitations because an adequately sized corpus would be time consuming to create dictionaries of counts for, but a smaller corpus size generally does not provide enough information to generate meaningful text. A slight solution to this problem is to pickle the dictionaries of counts so that it only needs to be done once.
g. Language models can be evaluated extrinsically, where human annotators will judge the language model based on some metric, or intrinsically, where the language model is compared based on some metric. A metric used commonly is perplexity, which is the

inverse probability of the words we see, normalized by the number of words. This is the formula for calculating perplexity:

$$PP(W) = P(w_1, w_2 \ldots w_N)^{-1/N}$$

h. Google's n-gram viewer is an accessible way to experiment with n-grams, available through Google books by the following link: https://books.google.com/ngrams/. Entering words and phrases into the n-gram viewer shows the user a graph of how often those words and phrases have been used over time in a corpus of books dating back to the 1500's. Here, I have entered the names of three detectives into the n-gram viewer: "Sherlock Holmes", a bigram, "Columbo", a unigram, and "Miss Marple", another bigram.



From this graph, we can see that Sherlock Holmes is by far the most popular detective now but the word "Columbo"'s usage was the highest before the show ever aired, exceeding Sherlock Holmes by far. I was able to find out that "Columbo" was a name used in a poem written by Matthew Prior, an extremely important English poet, by scrolling down to see the sources used by the n-gram viewer. Switching over to the British English corpus, Miss Marple is more popular than Columbo in the present, but less popular than Sherlock Holmes.