N-grams

N-grams are separations of tokens of text into groups of one or more tokens. For example, separating a text into unigrams is separating it into a list of each token, while separating it into bigrams separates it into a list of each pair of adjacent tokens. N-grams help to build a language model by grouping together words which may have some connection, which offers more information than just each word individually. By attaching certain words to others, n-grams can be used to build a language model and generate text in any given language. One could do this by using bigram probabilities to choose the most likely token to come after another over and over. N-grams can also be used to determine important multi-word phrases, such as measuring concordance with bigrams to determine that the bigram “United States” has deeper meaning than the sum of its parts.

In this project, I calculated unigram probabilities by comparing how often a unigram appeared in one language versus how often it appeared in every language. I did the same for bigrams, but with the Laplace smoothed bigram probability I used the equation provided (b + 1) / (u + v) which calculates the count of the bigram plus 1 divided by the count of the first word of the bigram plus the total vocabulary size (of unigrams in every language). Smoothing is important when dealing with these counts because some n-grams don’t appear at all in the languages we’re evaluating, so that could cause multiplication or division by zero if one is naïvely performing their calculations, which would ruin the result. Smoothing by adding 1 or adding the length of the vocabulary to each number solves that issue by ensuring that each probability has a baseline value.

The source text is important when building any language model because it can bias the result. Language is not written in just one way, and the topic and context of a text source determines multiple aspects of the language model that derives from it. For example, a corpus from an anatomy textbook

will be heavily skewed towards anatomical terms and much more formal than typical language. On the other hand, a corpus of Tweets would be much more informal and contain a much wider range of topics.

Language models can be evaluated by how well they represent the language and how complex they are. For example, a language model which is heavily biased would not be as useful as one which was trained on a multitude of data and generally reflects the overall use of the language. Additionally, a language model which was only trained on simple sentences would not be as valuable as one which has been trained on complex sentences.

Google's n-gram viewer allows anyone to examine how often a certain n-gram has been used in all of Google's repository of books, going back some centuries. It can be very interesting to see how a phrase has risen or fallen in popularity over time. However, it isn't exactly reflective of language use as it only reflects books and not the spoken word, especially for modern terms which may not have been added to as many books, if any, yet.

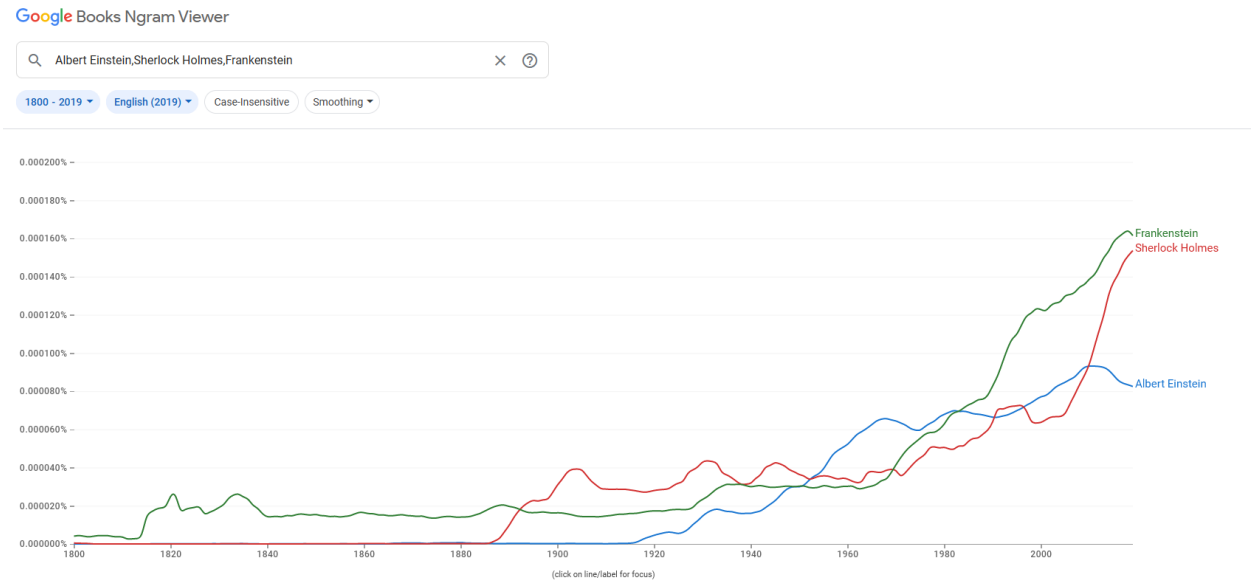

Figure 1: Google's Ngram viewer with some popular historical figures [1].

Bibliography

[1] Google, "Google Books Ngram Viewer," 2 October 2022. [Online]. Available: https://books.google.com/ngrams/. [Accessed 2 October 2022].