Tedi Dika (txd190008)
Ritvik Regulapati
3/4/2023

N-Grams Narrative

a. what are n-grams and how are they used to build a language model

N-grams are contiguous sliding windows of size n items over a given corpus and can be used to create probabilistic models. These probabilistic models can then be utilized to build a language model that predicts the probability of the next word given previously occurring text.

b. list a few applications where n-grams could be used

Some applications where n-grams could be used include,

1. Email Autocomplete - Generate new text in an email given previously occurring text i.e. “Given a start word, find the most likely next word.”
2. Grammar Correction - Identify misspelled text or incorrect grammar by comparing the text with a database of n-grams
3. Data Parser - Search through text to match query with relevant documents. Can be used in ATS, search engines, file search.

c. a description of how probabilities are calculated for unigrams and bigrams

The probability of a unigram is calculated by counting all the occurrences of the unigram and dividing by the total number of words in the corpus. The probability of a bigram is calculated by counting all the occurrences of a bigram and dividing by the number of occurrences of the first unigram in the given bigram.

d. the importance of the source text in building a language model

The quality of the source text is extremely valuable when building a language model. This is because the choice of source text will greatly change the model’s overall conclusions which as a result will then result in wildy different results. For example, a model who’s source text was from old literature will be wildly different in comparison to a model which learned from a modern textbook of some sort.

e. the importance of smoothing, and describe a simple approach to smoothing

Smoothing is a concept that is used as a solution due to something termed as the sparsity problem. The sparsity problem occurs when certain probabilities of occurrences end up being 0 which causes there to be a problem. The overall idea of smoothing is to fill in these 0 values of probability by using a little of the probability of the overall mass. The use of smoothing then as a result makes the overall distribution smoother.

One simple approach to smoothing is the Laplace smoothing approach(add-one approach). The idea behind this approach is to simply just add one to all the probability values in order to remove any chance of there being 0 values. Since there is no good way to predict which values will be 0 values and which will not, this adding of one is done to the entire distribution. And in order to offset that this happens, the total vocab count is then added to the denominator in order to fully balance it out.

f. describe how language models can be used for text generation, and the limitations of this approach

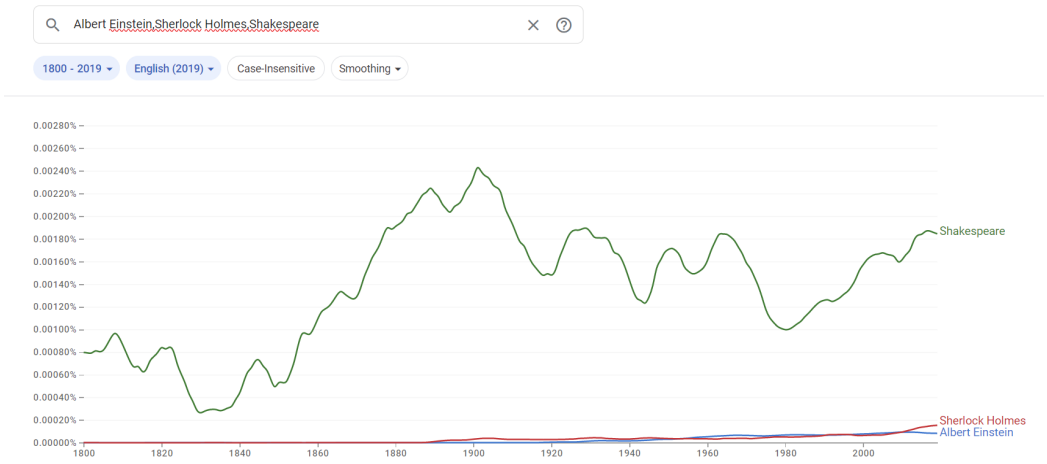
Language models can be used for text generation through the use of N-grams. What happens is that n-grams are converted into probabilities. These probabilities are simply just the count of the unigram divided by the total number of words. Using this, text can be generated as when given a start word, the model then looks at the ngrams with the highest probabilities of being next to in position and goes from there. The limitations of this approach are its small size as well as simple and naive approach as it can cause some big problems.

g. describe how language models can be evaluated

Language models can be clearly evaluated through having a human annotator evaluate the result using predefined metrics. As this is time consuming and can be difficult to do, another method is to use in-built metrics within a computer such as the concept of perplexity. What perplexity does is measure how well the language model predicts the text in the test data sample that is given to it.

h. give a quick introduction to Google's n-gram viewer and show an example

What Google's n-gram viewer does is given a set of phrases, it displays graph that shows how often those phrases have occurred in a corpus of books which are selected prior.



Here you can see that Google n-gram is giving a visual look at how often the phrases Albert Einstein, Sherlock Holmes, and Shakespeare are used in English books since the 1800s.