Anushka Karthikeyan
Brianna Murray
CS 4395.001
Professor Mazidi

NGram Narrative

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

N-Grams view text through sliding windows that show n-words at a time. There are three main types of N-Grams: unigram, bigram, and trigram. The unigram shows one word, bigram shows two words, trigram shows 3 words. The trigram is also known as N-Gram because the window can show any n amount of words within the window. N-Grams are used to build language models by counting how often certain word patterns appear within an entire body of text. Then the language model is built by calculating the probability from the frequency a word pattern is found compared to the amount of words within the given text.

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

N-Grams can be most frequently seen being used for spell-checking software and in tandem auto-complete for sentences. For example, when one is typing an email in Outlook sometimes there will be a suggestion for how to finish the sentence or even suggestions for how to reply to an email. N-Grams can even be used by those in the marketing space to help predict the best return on investment for high stake decisions. Marketers would use the probability percentages to determine key performance indicators which help to determine the best option to invest in based on the volume of words searched together. For example, a search such as “womens shoes” could help marketers know to invest more into womens shoes than shoes for children.

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

The probabilities are calculated by determining the fraction of time the word will appear in a body of text. The formula for a bigram is as follows: To calculate the probability of the first word and the word before it you must calculate how many times the two words appear in the given order together (serves as numerator), then count how many times the first word in the bigram is seen (serves as the denominator). After solving the fraction you will get a decimal value that can be converted into a percentage for the final result. P(wi | w i-1) = count(w i-1, w i) / count (w i-1). To calculate the probability of a unigram the formula is the number of times the word appears (numerator) over the number of times the word appears plus the unigram vocabulary size. P(w) = w / w +VocabSize.

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

Source text helps the language model learn. By having sound source text, the language model is able to accurately process the text to create a desired result for the user. Also, source text is used as a reference for the language model to build its predictions off of.

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

Smoothing is used to solve the zero count problem and ensures you avoid multiplying by a zero probability. One approach to smoothing is laplace smoothing (aka add-one smoothing). This approach adds 1 to all counts to eliminate any zero counts and then adds the vocabulary count to the denominator, so the Laplace smoothing probability is:

$$P(w_i) = \frac{C(w_i)+1}{N+V}$$

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

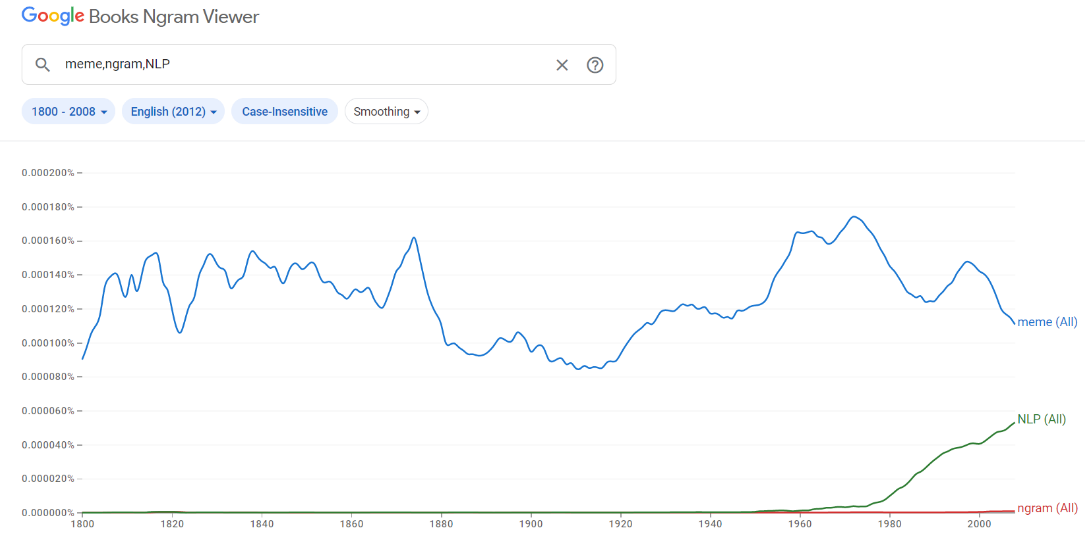
Language models can be used for text generation by creating probability dictionaries and using the n-gram probabilities to find what n-gram has the highest probability to be the start word. From there, it concatenates phrases based on the probability it will be next and finally ends when the last token is a period. However, using language models for text generation can be inaccurate as they are limited by the training corpus and ngram number. Higher ngrams provide better results than smaller one like unigrams and bigrams and a larger corpus also helps provide better results.

g. describe how language models can be evaluated

Language models can be evaluated using extrinsic and intrinsic evaluations. Extrinsic evaluations use human annotators and are typically very slow to perform. Intrinsic evaluations use an internal metric, like perplexity, and use it to compare models. These evaluations are done on a smaller set of data and perform faster than extrinsic evaluations.

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

Google's n-gram viewer provides a chart of word frequencies that have occurred in a corpus of books. Users type in phrases and the viewer charts the frequencies over a specified timeline. The n-gram viewer also has advanced features like case insensitive search, search for inflections, looking for particular POS, etc.
An example of Google's n-gram viewer is:



Google Books Ngram Viewer

meme,ngram,NLP

1800 - 2008 · English (2012) · Case-Insensitive · Smoothing