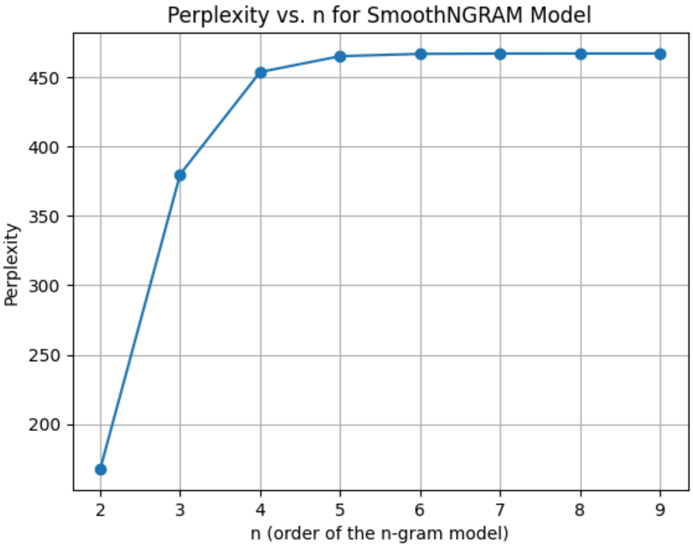
Report

N-gram models represent a category of statistical language models employed in natural language processing and machine learning. The 'N' in 'N-gram' denotes the quantity of words treated as a single unit. These models operate on the concept that the likelihood of a word in a sequence is contingent upon the preceding N-1 words. 'N' serves as a hyperparameter, and in this project, I conducted experiments with 'n' ranging from 2 to 10, assessing their respective perplexity scores.

It is noteworthy that I could calculate perplexity scores exclusively for Laplace-smoothed N-gram models. The computation of perplexity scores for non-smoothed models proved unfeasible, primarily due to recurring division by zero exceptions. This issue arose as I evaluated perplexity using the test data 'austen-persuasion.txt,' while the N-gram model derived its probabilities from the training data 'austen-emma.txt' of the Gutenberg corpus. The majority of N-1 grams in the test set did not occur in the training set, leading to the division by zero error during probability calculations for perplexity.

Upon calculating perplexity scores for the smoothed version across various 'n' values of N-gram models, it emerged that an increase in 'n' corresponded to an elevation in perplexity scores. The evaluation spanned from n=2 to 9, with n=2 yielding the lowest perplexity score of 167, while n=9 resulted in a score of 467. Below is a graph that shows perplexity score vs n.



Laplace smoothing adversely impacted the model's performance. Upon qualitative analysis of the generated text from both the unsmoothed and smoothed models, the unsmoothed model's text appeared more meaningful and coherent compared to the Laplace-smoothed version. For example I gave prefix as “emma by” to both a laplace smoothed and regular trigram model to generate a sentence.
Regular trigram model produced: emma by jane herself was infinitely superior he is the matter he understood what would be almost as much reason

Smoothed trigram model produced: emma by jane chose bountiful kind insulting brace languor yorkshire wander roused apiece unreasonable remedy hind disagreement fearfully harmless harry

When we try to read them both, the first one makes a lot more sense and seems a lot more coherent than the second one. Therefore it appears as though laplace smoothing is hurting the model’s performance.