NATURAL LANGUAGE PROCESSING

REPORT

TOKENIZATION

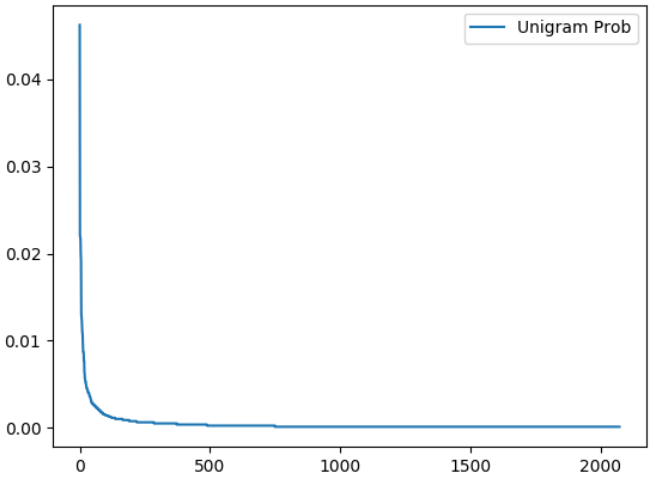
For tokenization, a hack that was used was that all the non-alpha numeric characters excluding space were removed that converted the data into a easy to parse dataset.

LANGUAGE MODELS

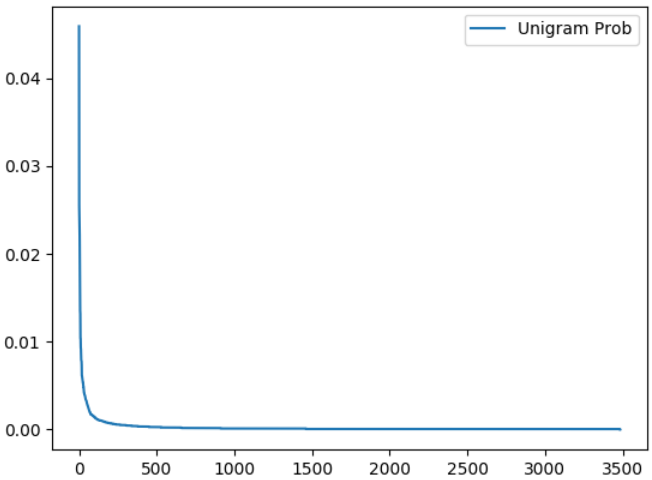
We implemented the unigram, bigram and trigram language models via the P(MLE) counts for the same. Below attached will be the graphs for the same

FOR UNIGRAMS:

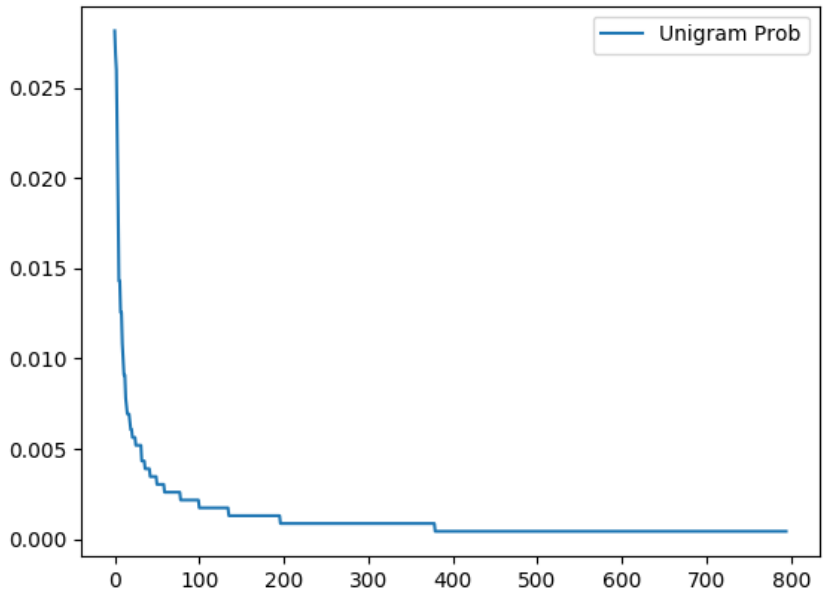
Entertainment:



News:

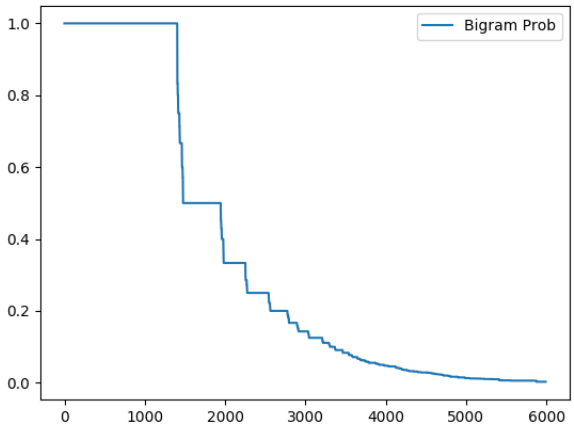


Lifestyle:

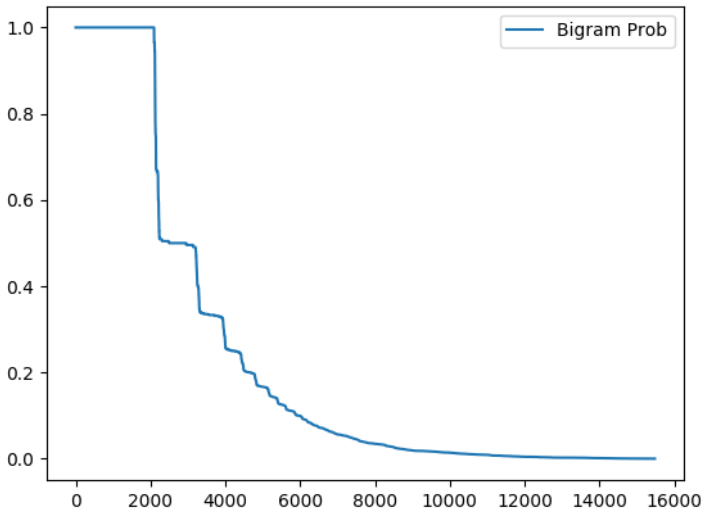


BIGRAM MODELS

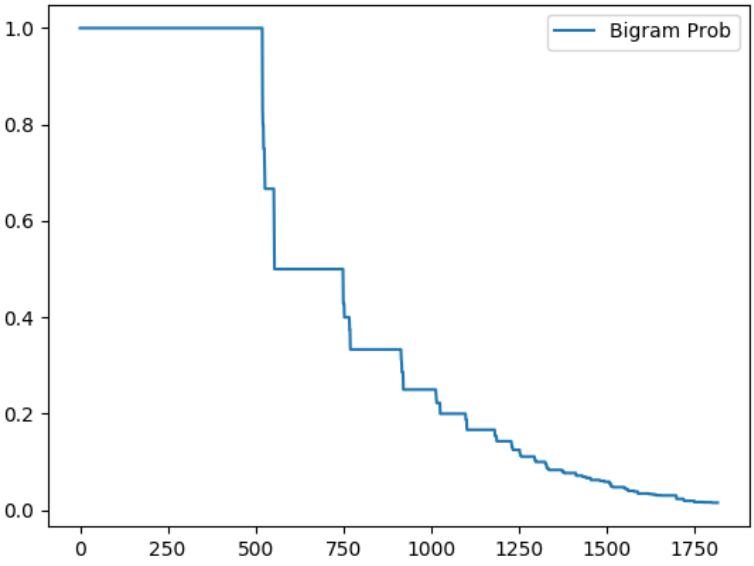
Entertainment:



News:

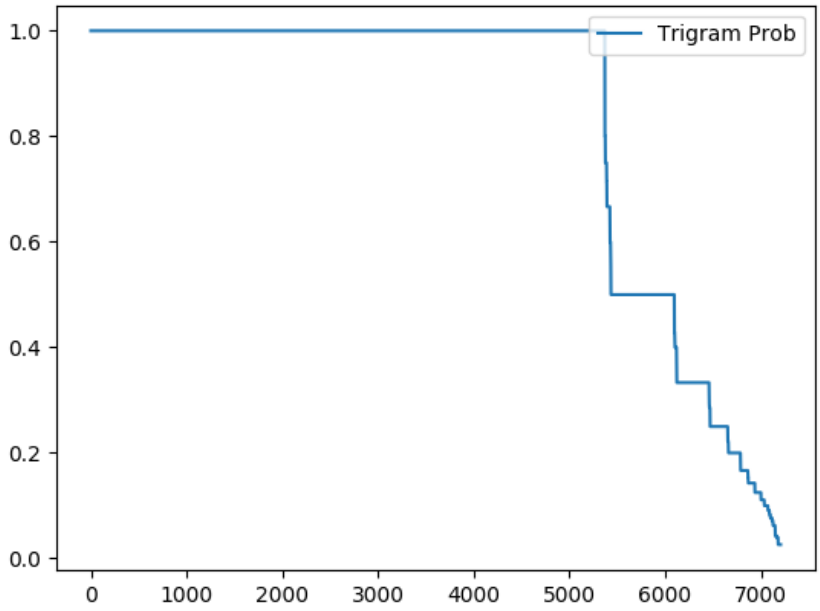


Lifestyle:

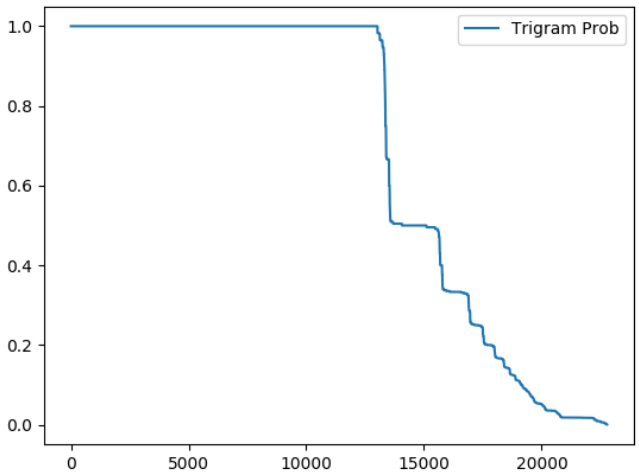


TRIGRAM MODELS:

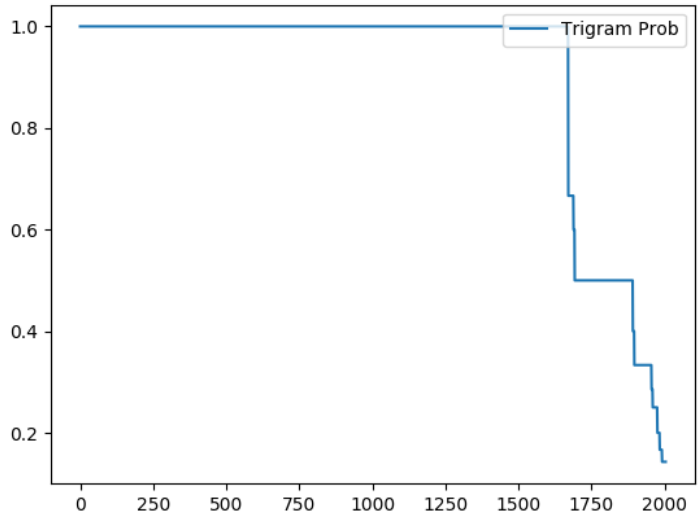
Entertainment:



News:



Lifestyle:



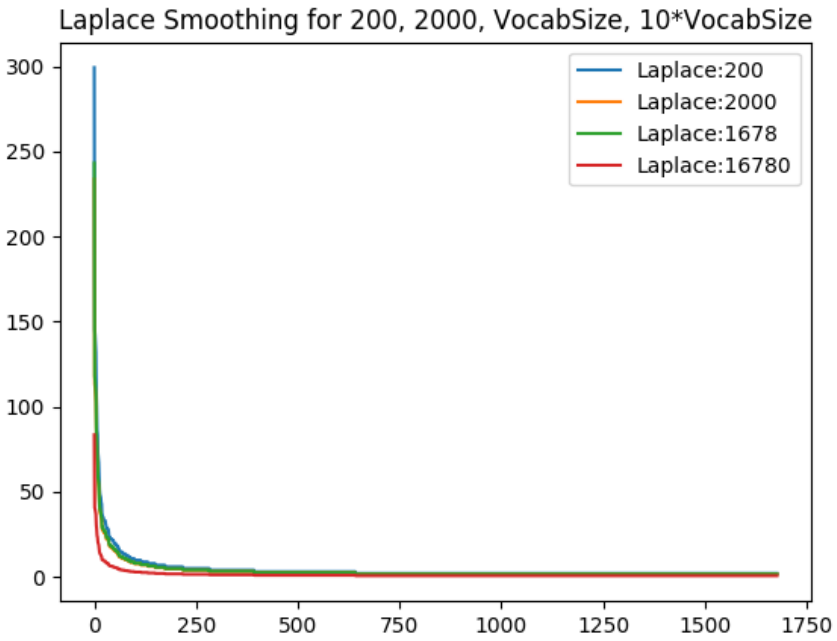
OBSERVATIONS:

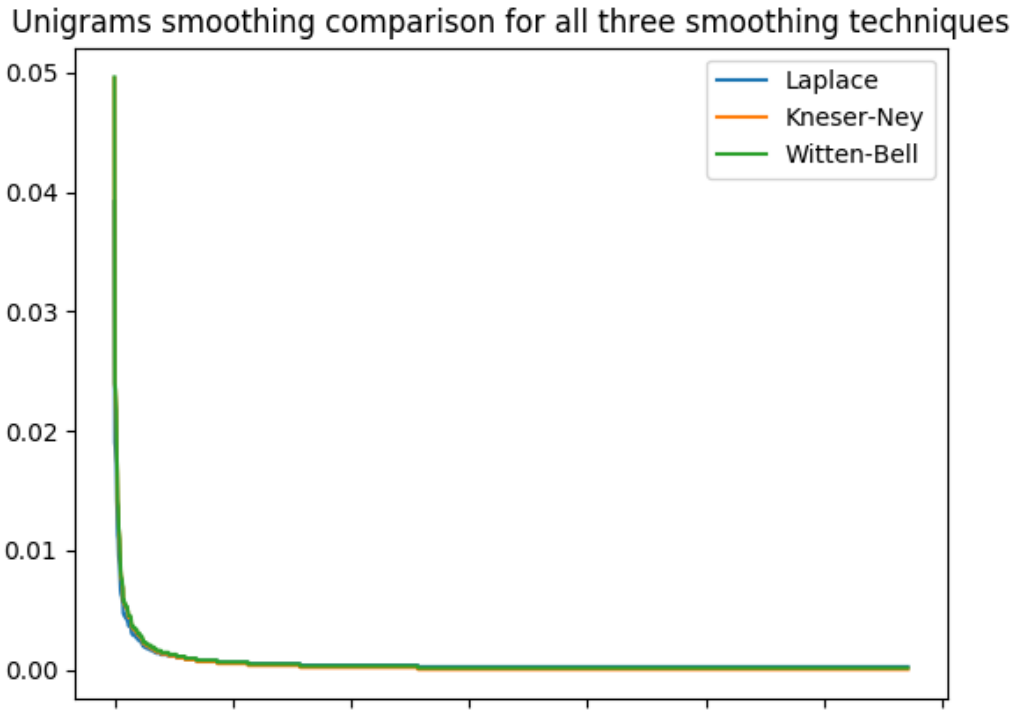
It is observed that the distribution of the unigrams, bigrams and trigrams follow the Zipf's Law. (Power Law).
Upon plotting on the log-log scale, it is found that at certain places, we can get an indication of a straight line but in some places where the curve is flat, we couldn't find the straight line, rather a curved line in those spots.

SMOOTHING TECHNIQUES:

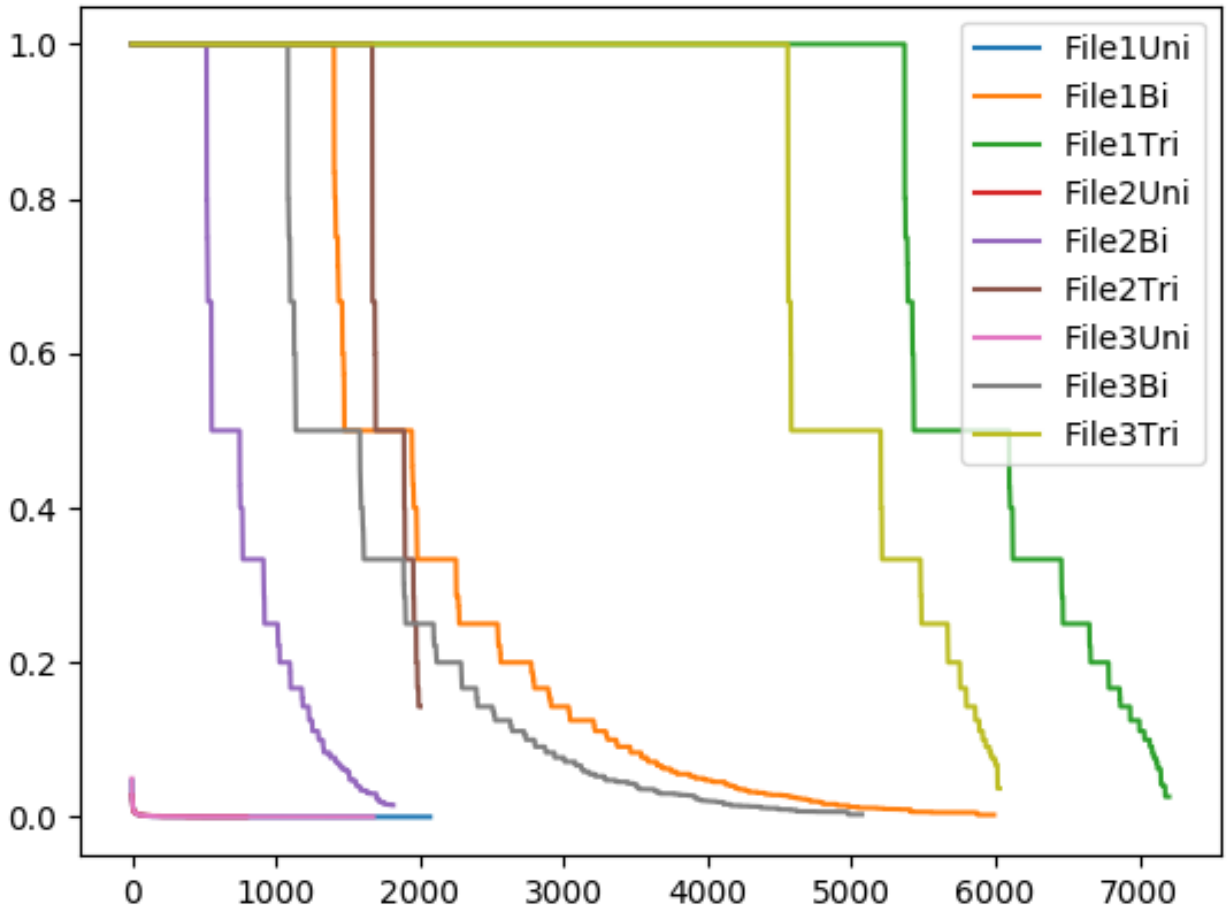
Three smoothing techniques were employed, namely Laplace, Witten-Bell and Kenser-Ney

Below, we can see that Laplace smoothing curve is smoother for V = 200, SizeVocab(1678 in this case)





The below graph shows us all the Zipf's curves for all 3 sources. It is seen that Unigram MLE probabilities coincide for all the 3 files whereas the Bigram and Trigram MLE probabilities are near each other but don't exactly overlap.



NAIVE BAYES

We implemented Naive Bayes using the IOB model. We used 5 classes, S, O, I, E, B to train the characters and calculated argmax of P(char|class)*P(class) to estimate our probs. for the remaining text. A sample is provided in the data folder (as _news.txt)

Although it failed to capture subtle points, but it acted as a decent classifier as it always classified spaces as O(which is nice!) and guessed start of token quite frequently. Thus, we achieved decent enough accuracy with this model, after a lot of thought process of going through other models.