CS 4395.001
Motalib Rahim
Mxr170012
Dr. Karen Mazidi
3. Narration

Ngrams

N-grams is a very useful tool in natural language processing under the field of computational linguistics and probability when there exists a sequence that is contiguous in a corpus. It is so powerful that it can be utilized in tasks such as retrieving information, modeling language and in translating. With the statistical aim the ngram can be used to train a model that can produce the distribution of texts in a language like the one in our project. The probability and the sequence of texts in a language. It can used in modeling languages, retrieving information, classifying texts, analyzing texts. For probabilities it can be done through laplace smoothing which was used in our project. Another is good-turing and the log. With the number of tokenized data and the overall data the probability can be calculated. The source text is very significant in building a language model because that is what the model will be trained on. Therefore, the results are impacted by it. It is necessary for the source text to be rich in quality and a decent quantity to get the accurate data of the model. Smoothing assigns a non-zero probability so that when it faces data sparsity in the trained data. The smaller the value of output it means there was more smoothing and higher the value there was less smoothing. This way that data can be prevented from overfitting. The language models can be used for text generation through probability distribution with the prediction of the upcoming word. Not only words it can help generate long texts. It does not necessarily have to be very predictable always. Though there are limitations where the model can not interpret the underlying meaning or context of the text be used in or for. In terms of semantics the language model would be poor with it. The language models I believe can be evaluated through its statistics. Calculating through the probabilistic models provide sufficient data about the language. The Google's ngram viewer is where users can look up usage of words or a phrase in books. In books the cultural changes of world are apparent. So it gives an insight of how the usage trends is or are along with its timeline. The timeline helps see how the prompt by the user have been throughout the ages from the 1500's to 2019. The corpus is updated till 2019. There is an option for smoothing. With the smoothing it gives more clarity or clear view of the graph with smoothing the graph. Suppose the term soccer and football, football is widely used however in the US it is known to be a different sport. Here, football has been used the most since soccer is only used in US. It shows a trend of how words have been used throughout. But here lies a weakness as like in what the term 'football' was used in as it could be a sport in USA or somewhere else. These tools will have pros and cons and may not be the best or perfect, there will always be room for improvement. The tools and techniques can be improved furthermore and overcome such cons or maybe another tool and technique can be developed newly and with more features.