Afaf Guesmia

CS 4395.001

Assignment 4

N-grams are defined as a sliding window over text, n words at a time. In other words, n-gramas are continuous sequences of words or symbols or they also can be tokens, in a text or document. They also can be defined as the items neighboring another sequence in a text. N-grams are used for different types of applications depending on if they are suitable or not. To be able to build a language model using n-grams, first, we need to tokenize a text as we have done in the different assignments using the word\_tokonize method, then use the tokens to extract the different unigrams bigrams, etc. After extracting, we count the frequency of their occurrences in the given text, which is then used to calculate the probability of each n-gram occurring in the text. This probability can help us have more information about the ability to predict the likelihood of the occurrences of different sequences in a text.

One of the examples that I can think of, is the keyboard on smartphones. While texting or using search engines to browse the internet, the application can generate the words that are more likely to appear with the previous word, it can be used from the likelihood of how vaguely the word is used commonly in society or it can be more personalized, depending on how the individual usually use those words. An example, can be if someone frequently texts his friends, “How are you doing today” however if the individual is trying to search “how are volcanoes formed?” the keyboard might suggest the word you after the word are.

In python, unigrams and bigrams are simply calculated using nltk.util import grams. As discussed in the lecture, for unigrams, we used the word\_tokenize and we token the text, we can ignore characters and numbers, then for bigrams, we use ngrams(unigrams, n), n is usually the number of the words one that follows another one. The source text is extremely important when it comes to forming a language model, it provided the most important data to build the model. A language model that is built using a large more diverse text is more likely to perform better than a model that is trained on a limited text. In summary, the source text is crucial when it comes to a language model for the reason that it provides the likelihood and predictions of the words.

This is one of the approaches used in the context of n-grams. One of the common issues with n-grams as discussed in lectures is they can be assigned to zero, which can lead to issues when using unigrams bigrams etc. To solve this problem, we can apply the smoothing method, with a little bit of probability of the overall mass, which makes the distribution smoother.

The language model can be a powerful method to use for text generation. They can predict the likelihood of words or sentences happening together, which can help save the user time, and be more efficient with the application. The language model can be used on smartphones for texting or search engines, they can be used for chatbots or even IA. However, we are seeing the language model being used in our days, however, there is the limitation of this algorithm, such as lack of context, I do believe there are other words that are used that the language model has not used yet, such as informal English or what is known of slang, they can also be offensive or problematic depending on the data they have been trained for.

The langue model can be improved by adding more data to them, such as metaphors or slang words. Also if a human was to evaluate and judge the generated text, it can help better the performance of the model as well as have a better insight into how the words are generated and what can be fixed.

Google n-gram viewer is defined as a tool that allows users to search for the frequency of the words, sentences, and phrases that occur in large data. This tool scans, books, and articles, which are the sequences and words that occur in a given text. The user can adjust the parameter of the tool, the sentences can be separated by commas and the tool will output the graph that contains those words, or sentences and their frequency in the text. An example of a sentence I search was global warming, I found it interesting that it never talked about it until around 1980 as shown in the figure below.

In conclusion, I believe the language model has a huge impact on human lives currently and it will still have a bigger impact on the world in the future. It is important to learn about the algorithm and its features as it is keeping to evolve over time.

Chart, line chart

Description automatically generated