**Bonus Assignment**

**Language Modeling:**

As explained during the lecture and the tutorials, Language Modeling is the task of predicting next word given previous sequence of words. Moreover there are different variations of Language Models: Unigram (Probability of word itself), Bigram (Probability of word given previous word), Trigram (Probability of word given previous two words), 4-gram (Probability of word given previous three words) and so-on.

The task is to create a 4-gram language model that is going to generate text. This task can be broken down into main pillars:

1. Creation of 4-gram function that takes an input line , append <START> , <END> tokens and return all the possible 4-grams available from that line as function Bigram, Trigram in the tutorial notebook.
2. Creation of a dictionary suitable for storing the language model probabilities and be able to query using 3 words. For example: Quadgram is our dictionary.

Quadgram([‘I’,’am’,’eating’]) -> should return a list/dictionary containing list of different words probabilities for this conditional P( Next word | I am eating) as explained in the tutorial.

1. Generate 2-4 sentences by inserting <START> token and two other words from your choice (but make sure they are in your dictionary) and then use the output and the previous two words to generate next word and so on.

For Example:

* Using <START> ,They ,are -> use them as input to the Quadgram dictionary.
* The Quadgram dictionary will return a list/dictionary of different words with different probability. You should select a word from this list randomly to create different sentences if you are using same tokens.
* Then using They, are , “New Generated Token” -> this is going to be the new input to your language model which will follow the same previous step and append it to are, Previously Generated Token, New Token and so on.

**Notice:**

1. More Bonus:  
   If you created 5-gram Language Model and compare both of the results , you will get a big bonus.
2. Link to colab/github: You will open the notebook , and download the dataset named “9053.txt” and put it in the same directory so the first cell which reads the dataset work.  
   If using colab , make sure to upload it in the correct directory.  
   https://colab.research.google.com/github/ahmedbahaaeldin/Introduction-to-NLP-Class/blob/main/Language\_Model.ipynb https://github.com/ahmedbahaaeldin/Introduction-to-NLP-Class/blob/main/Language\_Model.ipynb
3. Link to Dataset: <https://github.com/ahmedbahaaeldin/Introduction-to-NLP-Class/blob/main/9053.txt>

Don’t Hesitate to contact me or Eng. Roba for further Inquires.