Assignment 1: Language Modeling

Course Name: Natural Language Processing
Course No: CS60075
Instructor: Prof Sudeshna Sarkar
TA's: Alapan Kuila, Aniruddha Roy, Goutam Sagar, Sai Chaitanya
Assignment Date: 5th August 2022
Due-Date: 20th August 2022, 11:59 PM
Platform: Google collab/Kaggle/Machine

Task Definition: Language modeling is the task of predicting the next word or character in a document. This technique can be used to train language models that can be applied to a wide range of natural language tasks like text generation, classification, and question answering. The common language modeling techniques involve N-gram Language Models and Neural Langauge Models. A model's language modeling capability is measured using cross-entropy and perplexity.

N-gram Language model:

1. Implement an N-gram language model with Laplace smoothing and sentence generation.
2. Write a language model class with the below function.
   i) A smoothing function for applying Laplace smoothing to n-gram frequency distribution.
   ii) A model function to create a probability distribution for the vocabulary of the training corpus. Please note if you use a unigram model, then probabilities will be simple frequencies of each token with the entire corpus. Otherwise, the probabilities are Laplace-smooth relative frequencies.
   iii) A perplexity function to evaluate the language model on the test dataset.
   iv) A generate-sentence function to generate a sentence using the language model.

Neural Network Based Language Model :

1. Implement a neural network-based language model, and calculate the perplexities for each input sentence based on the trained model. Please split your train data between 90% and 10% to create a dev set and fine-tune the model based on the dev dataset. Please prefer to use a PyTorch framework to implement the above code.
2. Please use Word2Vec/Glove for the initialization of the model.

Evaluation Metrics: Perplexity (For evaluating both the models)
Dataset: English
Submission Materials: Python file, Google drive link for the trained model, a doc-file with results, and your observation.