Assignment 3: Language Modeling

Homework assignments will be done individually: Each student must hand in their own answers. Use of partial or entire solutions obtained from others or online is strictly prohibited. Electronic submission on Canvas is mandatory.

You will need to use the Penn Treebank corpus for this assignment. Three data files are provided: train.txt, valid.txt, and input.txt. You can use train.txt to train your models and use valid.txt for testing. File input.txt can be used for a sanity check on whether the model produces coherent sequences of words for unseen data with no next word.

1. **Preprocessing** (0 points) Run the functions including remove punctuation, remove url, remove number, lowercase and tokenize.

2. **N-gram** (50 points)
   (a) (10 pts) Implement a BiGram model for language modeling. Fill in the code for class `BiGram`.
   (b) (15 pts) Implement Good Turing smoothing. Fill in the code in the class `GoodTuring`. Hint:
   - Use power law to replace empirical N_c when c is greater than 100. (See page 69 in lecture slides of language modeling)
   - You will need to calculate frequency of frequency for both bi-gram terms and unigram terms. $P_{GT}(w_2|w_1) = \frac{P_{GT}(w_1,w_2)}{P_{GT}(w_1)}$

   (c) (15 pts) Implement Kneser-Ney smoothing using:

$$P_{\text{KN}}(w_i|w_{i-1}) = \frac{\max(c(w_{i-1},w_i) - d, 0)}{c(w_{i-1})} + \lambda(w_{i-1})P_{\text{CONTINUATION}}(w_i)$$

   where

$$\lambda(w_{i-1}) = \frac{d}{c(w_{i-1})}|\{w : c(w_{i-1},w) > 0\}|$$

$$P_{\text{CONTINUATION}}(w) = \frac{|w_{i-1} : c(w_{i-1},w) > 0|}{\sum_{w'}|\{w'_{i-1} : c(w'_{i-1},w') > 0\}|}$$

   Check the slides for specifying the value of d. Fill in the code for class `KneserNey`.
   (d) (10 pts) Implement Perplexity and use perplexity to evaluate BiGram, Good-turing, and Kneser-Ney. Fill in the code for function perplexity.
   (e) (0 pts) Choose the first 30 lines in the input.txt file and print the predictions of next words using your BiGram, Good-Turing, and Kneser-Ney models.

3. **RNN** (50 points) You can use libraries such as PyTorch or TensorFlow for this part.
   (a) (0 pts) Initialize parameters for the model.
   (b) (5 pts) Data preparation. You will need to build a vocabulary, prepare training and validation data.
   (c) (10 pts) Build your RNN model. The model should include an embedding layer(input layer using word embedding vectors), a hidden layer (RNN cells), an output layer (predict next word).

(d) (10 pts) Setup the training process and train your RNN model.

(e) (15 pts) Model evaluations: Prove that perplexity is the exponential of the total loss divided by the number of predictions. Calculate the perplexity score of your model predictions.

(f) (10 pts) Print the predictions of next words using the RNN model for the same 30 lines of input.txt as in N-gram.

4. **Submission Instructions** You shall submit a zip file named Assignment3_LastName_FirstName.zip which contains:

   - code files (.ipynb or/and .py) including all the code, comments, plots, and result analysis. You need to provide detailed comments in English.
   - If you submit your code in py files, you can include a pdf file to show your plots and result analysis.