CSCI-LING 5832-001,001B: Natural Language Processing
Homework 2

Instructor: Katharina Kann
GSS: Beilei Xiang
Due: September, 21, 2020 at 10:00 am
Turn in electronically via Canvas.

Problem 1 (70 Points)

Write a python program to perform the following tasks:

1) Estimate a bigram language model on "hw2_training_sets.txt". Use your language model to compute:
   A. the probability,
   B. the probability normalized by sentence length,
   C. and the perplexity

   of the 5 sentences given in "test_set.txt". The training and test files have already been tokenized.
2) Explain how you have handled unknown words that appear in the test set.

Submission:
- Submit your python code as "lastname-firstname-part1.py".
- Submit an output file "output.txt" with each line containing all required values for the respective sentence, separated by comma.

Note: This time, you are not allowed to use NLTK.

Problem 2 (30 Points)

Sample a sentence from your language model and compute its perplexity.

Submission:
- Submit your python code as "lastname-firstname-part2.py".
- Submit your solution as a pdf file named "problem2".