Homework 1

CS312 - Spring 2024

This homework is to help students practice the text processing, regular expressions, and n-gram language model. You are given two files (training_data.txt and testing_data.txt). You should use NLTK to split the text into words and sentences.
https://www.nltk.org/_modules/nltk/tokenize/regexp.html

Question 1: Find all sentences that contain “to be” verbs (i.e. “is”, “are”, …) in the training data file.

Question 2: Build a unigram model and a bigram model (both are with add-one smoothing) from the training data file. Then calculate and compare the perplexity score of these two models on the testing data file.

Submission: You need to submit both the code, some intermediate results (e.g., high-frequency words, high-frequency word-pairs, conditional probability, etc), and the final results (sentences with “to be” verbs, perplexity scores) on Canvas.