CS 6320: Natural Language Processing

Homework 1: N-grams (40 points)

Due date: 09/13/2019 4pm

An automatic speech recognition system has provided two written sentences as possible interpretations to a speech input:

S1: Milstein is a gifted violinist who creates all sorts of sounds and arrangements .

S2: It was a strange and emotional thing to be at the opera on a Friday night .

Write a C/C++/Java/Python/Perl program that trains a language model on the corpus (provided as an addendum to this homework) and finds out which of the two sentences is more probable.

Input: Your program must take two command line arguments from the user: an integer $N \in \{2, 3\}$ that indicates whether to use a bigram or trigram language model; and another integer $b \in \{0, 1\}$ that indicates whether the model should be trained without smoothing or with add-one smoothing.

Example command line usage: python Ngrams.py -N 2 -b 1

This python call requires your program to train a bi-gram language model on the corpus with add-one smoothing.

Output: Your program will output two matrices for each case:

1. A matrix showing the N-gram counts for the given sentences.
2. A matrix showing the N-gram probabilities for the given sentences.
3. The probability of both sentences, as computed using the language model.

To submit: Submit a single zip file that contains the following:

1. All your source code (include the corpus file as well).
2. A README file giving clear instructions on how to run the program.
3. A report containing the matrices and probability values output by your program for all aforementioned cases.