CS 6320: Natural Language Processing

Homework 1: N-grams (40 points)

Due: February 24 2023, 2:00 pm

For this homework, you will train and test the performance of a Bigram language model. You may use either C++, Java or Python to write your code. To test your code, please use the csgrads1/cs2 server[1].

- The corpus is a simple plaintext file. **Each line represents a new sentence**. Tokens in a sentence are **white-space separated**.
- Train a Bigram language model on the training corpus. The model must be trained for two scenarios: no smoothing and add-one smoothing.
- Evaluate your model against these 2 sentences:
  - mark antony shall say i am not well , and for thy humor , i will stay at home .
  - talke not of standing . publius good cheere , there is no harme intended to your person , nor to no roman else : so tell them publius

You are not allowed to use any internet software for this homework. You have to write your own software implementing the algorithm.

Your program should take as input the following two arguments:

```
.\program <training-set> b
```

where `<training-set>` represents the path to the training corpus; and `b` ∈ {`0, 1`} is an integer that indicates whether or not to use add-one smoothing.

For example, the call `.\program train.txt 1` indicates you have to train a bigram language model with add-one smoothing on the file 'train.txt' and evaluate it against the two sentences mentioned above.

For evaluation, your program must output the following:

- A table showing the bigram counts for both sentences
- A table showing the bigram probabilities for both sentences
- The probability of both sentences under the trained model

Submit the following bundled into a single zip file via eLearning:

1. Your code file(s)
2. A Readme giving clear and precise instructions on how to run the code.
3. A plaintext file showing the output of your code for the two sentences.

[1] https://cs.utdallas.edu/about/computing-facilities/