CS 6320: Natural Language Processing

Homework 1: N-grams (40 points)

Due: Feb 17, 2020 2:30 pm

For this homework, you will train and test the performance of a bigram language model. You can use either C/C++, Java, Python or Perl to write your code. To test your code, please use the csgrads1 server.

- Download the training and test corpora available on this link:

  http://www.hlt.utdallas.edu/˜moldovan/CS6320.20S/homework.html.

  Datasets are simple plaintext files. Each line represents a new sentence in the corpus.
- Train a bigram language model on the training corpus. The model must be trained for two scenarios: no smoothing and add-one smoothing.
- Evaluate the model by computing the probabilities of sentences present in the test corpus.

Your program should take as input the following three arguments:

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

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

For example, the call `.\program train.txt test.txt 1` indicates you have to train a bigram language model with add-one smoothing on the file 'train.txt' and evaluate it on 'test.txt'. A sample sentence in the test corpus looks like:

I do not think this young lady is so Celtic as I had supposed.

Your program must output the following:

- A matrix showing the bigram counts for each sentence
- A matrix showing the bigram probabilities for each sentence
- The probability of each sentence

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

1. Your code files
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 test set.