Name: Shiva Kumar

Net id: sak220007

Course: CS 6320

Assumptions Made:

Assumptions 1: Tokens are white space separated. So, to get a list of all the tokens I can split the Corpus by white space.

Assumption 2: Punctation are not words. So, the unigram will compose of all the tokens without punctuation.

How to run my program:

Set up: Put the train corpus in a .txt file named train

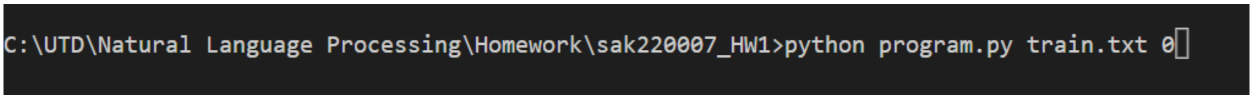
Note: I am using python3 to run my code.

No-smoothing:

In the cmd/terminal window enter:

python program.py train.txt 0

Example:



```
C:\UTD\Natural Language Processing\Homework\sak220007_HW1>python program.py train.txt 0
```

Name: Shiva Kumar

Net id: sak220007

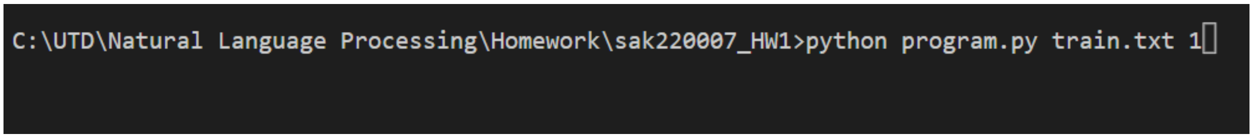
Course: CS 6320

Smoothing:

In the cmd/terminal window enter:

python program.py train.txt 1

Example:



```
C:\UTD\Natural Language Processing\Homework\sak220007_HW1>python program.py train.txt 1
```

Output:

The program will display the count table, probability table, and probability for each test sentence in the terminal and will put them in nosmooth.txt or smooth .txt depending on user input.