HLT – Fall 2019 – Mazidi

Homework 1: Word Frequencies

Objective: Use Python to read files from a directory and calculate frequency statistics on words. Read about NLTK's FreqDist() function here: https://www.nltk.org/book/ch01.html

Turn in: your Python .py file (Do not turn in a Jupyter notebook)

Instructions:

1. Download the hm1_dir from Piazza and save it in the same folder as your Python program. Within your main() function, get the directory name from a system argument. Print an appropriate error message if the sys.argv is missing and end the program.
2. In a for loop, (a) read in each file, (b) use the string.replace() method to replace newlines with spaces, (c) use the string.lower() method to lower case the text, then use NLTK tokenizer to extract tokens, (d) make a FreqDist from the tokens, (3) print the filename and 5 most common words, (e) on each iteration through the files, add the FreqDist to a cumulative FreqDist for later. You can literally add += the distributions.
3. Repeat the same loop as above but add the following: (a) remove punctuation symbols before tokenizing, (b) remove stopwords.
4. For your cumulative FreqDist for steps 2 and 3, create a cumulative frequency graph of the 50 most common words. Note that you may have to install matplotlib, and the first time matplotlib runs it takes a while.
5.

Element	Points
Python script runs without error	30
Appropriate comments and white space	10
Steps 1-4	15
Total	100

The file path names vary from Mac to Windows computers. In order to make things universal, so that the TA can grade no matter what system you use, please try the following method to get the file names from the directory.

```
import os
import sys
cwd = os.getcwd() # get current working directory
parameter = sys.argv[1]  # get folder name
# use join instead of concatenating strings
# it will use either / or \\ depending on your os
print(os.path.join(cwd, parameter)) # print to confirm
```