Tushar Prasad, Melissa Mannis, Larry S. Liebovitch, Peter T. Coleman
NOTES for Tuesday 2024_01_09, 11:00 AM – 12:30 PM AC4

1. Overview: Sustaining Peace, Peace Speech, powerpoint

2. I've set up (use your Columbia uni email for access)
Columbia Google Drive - for files
AC4_Spring2024_GDrive
https://drive.google.com/drive/folders/1RyAqvK5XITYOL38ZBSD1wHl7kdteOhB0?usp=drive_link
GitHub repository - You can use your own if you prefer
AC4_Spring2024
https://github.com/LarryLiebovitch/AC4_Spring2024

3. PLOS ONE paper
https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0292604

4. LexisNexis (LN) word frequency DATA in folder LN_Data on the Columbia Google Drive. The number of times each word occurs for the 10,000 words of highest occurrence, separately, in LOW and HIGH peace countries.
TAB DELIMITED spreadsheets, cols separated by TAB, rows by RETURN
    nopeace_10K.txt
    peace_10K.txt
There is also a short piece of these files (that I found helpful in debugging):
    TEST_peace_short.txt

4. First Task: Compute the NORMALIZED word frequencies for each word in the set of the LOW and HIGH peace countries.
    For EACH country, SUM the total number of occurrences of ALL words = N(country).
    Then the NORMALIZED word count for each word in that country is the number of occurrences of that word N(word,country) / N(country).
    Then for each word, find the AVERAGE of the N(word,country) / N(country) across countries. Do all this separately for the nopeace and peace data sets.
    The AVERAGE of the N(word,country) / N(country) will be the values of the features used in the Machine Learning.

5. Next Steps: use those feature values of each word to train a machine learning model (logistic regression, random forest, SVM, XGBoost) to classify a country as nopeace or peace.
NOTE: there is a different set of words in the nopeace v. peace files. That is, a small number of words in the nopeace file is missing from the peace file and a small number of words in the peace file is missing from the nopeace.
To start: it may be helpful to construct the machine learning model from:
    The 100 most frequent words in the nopeace and peace files.
    The 1,000, most frequent words in the nopeace and peace files.
    The 10,000 most frequent words in the nopeace and peace files.

7. Set up AWS account