**Instructions**

Firstly, the text analysis was performed using Jupyter Notebook. Later, the .py file was opened in PyCharm and the code was run to generate the Output file. The .ipynb file is also attached.

To ensure the code runs smoothly, make sure that the stopword files, +/- words files, and Input.xlsx file are all copied and pasted into the default directory along with the .py file. This is done so that no one has to edit the path in the codes.

Next, install the required libraries:

* pip install pandas
* pip install bs4
* pip install newspaper3k
* pip install openpyxl

Now the steps undertaking for the Analysis :

* Importing necessary libraries
* Reading the URLs from the Input.xlsx and storing in a list
* Using Article library to create a user defined function to extract the article title and content text from each URL. Note: 3 URLs had 404 Client error and no longer contained the article information. So the code was updated to give an exemption to these links.
* Created an empty list and append it with URL\_ID obtained by iterating over the file names of the created text files, along with the title and text.
* Created a dataframe (df) containing URL\_ID, Title and Text.
* Saved the dataframe as data.xlsx
* Started with Text preprocessing, Did analysis on 1st row of the Text Column by normalized the data.
* Imported re library to use regular expression to clean the data
* Added the stopwords files to compare and remove the stopwords from the text
* Created a user defined function called text\_preprocessing based on the text preprocessing done on 1st row, then applied it to all rows of Text Column by using the apply function
* Computed the positive, negative, polarity, subjectivity score based on the formula and applied it to all columns with apply function
* Made a user defined function called count\_syllables which can count the syllables
* Another user defined function called compute\_readability which calculates the average sentence length, percentage of complex words and fog index
* A user defined function called avg\_words\_per\_sentence to calculate the average number of words per sentence.
* A user defined function called count\_complex\_words to count the complex words.
* A user defined function called word\_count to count the total number of words in the text
* A user defined function called count\_syllables\_per\_word to count syllables in each word of the text.
* A user defined function called count\_personal\_pronouns to find if there is any personal pronouns(I|we|my|ours|us)
* A user defined function called avg\_word\_length to find the average word length
* Applied all these functions to each row along with making the required columns
* Inserted the Column URL and updated it with the URLs from the list created in step 2. Keep in mind that the 3 URLs which had error 404 was not added so that the rows will match and the URL will be accurate to its other columns.
* Dropped the Title & Text Column, since its not there in the output format.
* Then finally renamed the columns properly as per the format and saved the file as Output.xlsx