Objective:

Extract textual data articles from the given URL and perform text analysis to compute variables.

Approach:

1. Data extraction:

    - Load the Input.xlsx file
    - This file contains the URL_ID and URLs from where we need to extract the article title and content.
    - Before extracting the contents, we check if we face any issues while accessing the URLs by checking the status code of the URL.
    - We check the statues code for each URL and print if any issues are detected or not.
    - We extract the contents by using “BeautifulSoup” python library which is used for web scraping.
    - Extracting content for each URL separately is time consuming so we use the conditional statement for loop to iterate the process for all the URL_ID and URLs in the Input.xlsx file.
    - Then we create an output directory where we store all the extracted article contents.
    - The contents for each URL are saved in separate files in the output directory with the URL_ID as the file name.
    - Each file contains the title and content of the article.

2. Removing stop words:

    - We load the stop words directory “StopWords”.
    - The files in this directory include seven more text files containing stop words that are to be removed from the extracted text.
    - All the stop words in StopWords directory are converted to lower case.
    - Then we clean all the text files that are present in “output” directory.
    - This process involves punctuations and stop words removal.
    - After cleaning all the text files we then save them in a separate directory “cleaned_output”.
    - All the cleaned files are names by their corresponding URL_ID.

3. Sentimental Analysis:

    - In this step we perform the textual analysis on the extracted texts and compute variables that are given in Output Data Structure file.
    - Firstly, we load the csv file where we have to save the calculated data.
    - Each variable is calculated by declaring a function named after the variable that is to be calculated.
    - In this step we also display the value for the respective variable for each article.

- After passing a function to calculate the values of a variable, we then load these values for that variable in the csv file for all URL_IDs.
- This csv file has similar structure as the Output Data Structure file.
- After calculating all the values, we can check the csv file by directly accessing it from our main directory or using the command “df” in the python notebook.

Dependencies:

Import all the libraries mentioned below before running the files:

1. requests
2. bs4
3. beautifulsoup4
4. pandas
5. nltk

How to run the files to generate output:

1. Create a folder with all .ipynb files, StopWords, MasterDictionary, Input.xlsx and a copy of Output Data Structure file in it.

2. Run the “article_extraction.ipynb” file to extract the articles from the URLs and saving then in the output directory.

3. You can check the extracted articles in the output directory that is created after running the ”article_extraction.ipynb” file

4. Run the “clean files.ipynb” file to clean the extracted files for stop words and punctuations.

5. You can check the cleaned articles in the cleaned output directory that is created after running the ”clean files.ipynb” file

6. Run the “sentimental_analysis.ipynb” file to calculate all the variables and store then in the csv file.

7. To check the data calculated for each variable open the copy of Output Data Structure that you have created.