### 1. Approach to the Solution

**Data Extraction (Web Scraping):**

* I will use Python for web scraping using the requests and BeautifulSoup libraries.
* Each URL from Input.xlsx will be visited, and only the article text (title and main content) will be extracted. This ensures that unnecessary webpage elements like headers and footers are avoided.
* The extracted text will be saved into individual text files named by their respective URL\_ID.

**Text Analysis:**

* For text analysis, I will utilize libraries like NLTK (Natural Language Toolkit) for various linguistic operations such as sentence splitting, word tokenization, part-of-speech tagging, etc.
* I will compute the required variables such as positive score, negative score, polarity score, subjectivity score, average sentence length, percentage of complex words, FOG index, average number of words per sentence, complex word count, word count, syllables per word, personal pronouns count, and average word length.
* These calculations will be performed per article and stored in memory.

**Output Generation:**

* Finally, I will format the results according to the structure provided in Output Data Structure.xlsx using the openpyxl library to write to Excel files.
* Each row in the output Excel will correspond to an article's analysis results, including the input variables from Input.xlsx and the computed text analysis variables.

### 2. How to Run the .py File to Generate Output

To run the Python script in Google Colab and generate the required output:

* Upload Input.xlsx to your Google Colab environment.
* Create a new Python notebook or script.
* Install necessary dependencies (see below).
* Write and execute the Python script for web scraping, text analysis, and output generation.
* Download the Output Data Structure.xlsx file once the script completes.

### 3. Dependencies Required

Make sure to install the following libraries in your Python environment (Google Colab supports installing packages via pip):

* requests: To fetch web pages.
* BeautifulSoup (from bs4): For parsing HTML and extracting article text.
* nltk: For natural language processing tasks like tokenization and part-of-speech tagging.
* openpyxl: To read and write Excel files.

You can install these dependencies using the following commands in your Google Colab notebook:

!pip install requests

!pip install nltk

!pip install openpyxl

!pip install html5lib

!pip install bs4