Test Assignment Instructions

By : Garvit Jain ,IIT Mandi CSE

To run the provided code successfully, you can follow these step-by-step instructions:

Step -1 : Mount Google Drive:

- Ensure you have a Google account and access to Google Drive.
- Launch a Google Colab environment.

Step -2 : Change Directory (if necessary):

- If your working files are stored in a specific directory within your Google Drive, you can change the current working directory.
  Replace '/gdrive/MyDrive/ test' with the path to your desired directory within Google Drive.

Step -3 : Import Libraries:

- Import the necessary libraries for web scraping, data processing, and natural language processing.

Step -4 : Download NLTK Data:

- Download NLTK data required for tokenization and stopwords. You only need to run this once.

Step -5 : Load DataFrame:

Load the DataFrame from the Excel file named 'Input.xlsx'. Make sure 'Input.xlsx' contains the URLs and URL_IDs you want to scrape.

Step -6 : Web Scraping and Text Extraction:

- ◆ The code block starting with the for loop will iterate through the URLs in the DataFrame, make HTTP requests, and extract text content from web pages. This includes extracting titles and main article text.

Step -7 : Define Directory Paths:

- ◆ Set the directory paths for various data and resources within your Google Drive. Replace these paths as needed.

Step -8 : Load Stopwords:

- ◆ Load stopwords from the specified directory and create a set of stop words.

Step -9 : Tokenization and Data Preprocessing:

- Tokenize the text and remove stopwords for further analysis.

Step -10 : Sentiment Analysis:

- Perform sentiment analysis by identifying positive and negative words in the text content.

Step -11 : Readability Analysis:

- Analyze readability by calculating metrics like average sentence length, percentage of complex words, Fog Index, and more.

Step -12 : Cleaned Words Analysis:

- Analyze word count and average word length in the text content.

Step -13 : Personal Pronoun Count:

- Count personal pronouns in the text content.

Step -14 : Update the Output DataFrame:

- Update the output DataFrame with the calculated metrics.

Step -15 : Save Output Data:

- Save the updated DataFrame to an Excel file named 'Output_Data.xlsx'.

Now, you can execute the entire script in a Jupyter Notebook or Google Colab environment by copying and pasting each code block sequentially. Ensure that you have the required input data ('Input.xlsx') and directories correctly set up in your Google Drive before running the code.