OBJECTIVE

To scrape data from various URLs and perform analysis on text content to extract information such as positive score, word count, sentence count, syllable count, avg sentence length, etc.

APPROACH

1. Web Scrapping
   - Use Beautiful Soup to scrape data from each URL.
   - Retrieve text content from scraped data.
2. Data Storage
   - Store the scrapped text content in local Directory using ‘os module’
3. Retrieve Given Data
   - Data like stop words is retrieved and stored in a common array for further preprocessing.
   - Output.xlsx is retrieved as a DataFrame named as output.
4. Preprocessing
   - Scraped data is preprocessed like punctuation removal, stop word removal.
5. Parameter Calculation
   - Required features are Calculated for each URL.
   - Also make sure which parameters or features require preprocessing step or not.
6. Results
   - These calculated data is stored in output DataFrame.
   - Now this DataFrame is converted to comma separated File (csv) File.

How to run .py file to generate output

- Make sure all the dependencies are downloaded to generate output.
- Open file in vs code or PyCharm or any other IDE.
- Open the Terminal or command prompt
- Make sure all the Input and given files must be present in the same directory.
- Run command - Python main.py
- After running main.py, a new file called Output_data.csv will download in same directory, which contains the output.

Dependencies Required

- requests: To make HTTP requests to fetch web pages.
- pandas: To create DataFrame for storing output.
- os: To fetch data from our file system.
- beautifulsoup: To scrap web pages.
- nltk: To perform preprocessing tasks on data.
- re: re is for Regular Expression.
- string: to perform some preprocessing.
- openpyxl
- lxml