Instructions

Data Extraction:

- Read an Excel file named "Input.xlsx" that contains the URL_ID and URL columns.
- Iterate through each row of the Excel file and extract the article text from the corresponding URLs.
- Parse the HTML content of each URL using BeautifulSoup.
- Find the article title element and the paragraph elements containing the text.
- Write the title and paragraphs to separate text files named with the URL_ID.

Data Analysis:

- Remove the rows from the DataFrame where the title was not found (URL_IDs 44, 57, and 144).
- Define the path to the "StopWords" folder.
- Iterate through each file in the "StopWords" folder and extract the stopwords.
- Perform text cleaning by removing the stopwords from the extracted article text.
- Load positive and negative dictionaries from files.
- Calculate the positive and negative scores by counting the occurrences of positive and negative words in the cleaned text.
- Calculate the polarity and subjectivity scores based on the positive and negative scores.
- Analyze the readability of the text by calculating various metrics such as average sentence length, fog index, average word length, etc.
- Store the analysis results in a dictionary for each URL_ID.
- Append the analysis results to a list.

Output:

- Convert the list of analysis results into a DataFrame.
- Save the DataFrame to an Excel file named "Output Data Structure.xlsx" without including the index.

NOTE *- Please ensure that you have the necessary input files (Input.xlsx, positive-words.txt, negative-words.txt, and the files in the StopWords folder) in the specified paths.*