**Project Documentation**

**1. Solution Approach**

**Objective:**

The objective of this project is to extract textual data from articles provided via URLs in an Excel file and perform various text analysis operations to compute specific variables. The analysis results are saved in a structured format.

**Approach:**

1. **Data Extraction:**
   * **Input:** The URLs are provided in an Excel file (Input.xlsx).
   * **Processing:**
     + Each URL is accessed using the requests library.
     + The HTML content of the page is parsed using BeautifulSoup.
     + The article title and main content are extracted, excluding headers, footers, and any irrelevant sections.
     + The extracted text is saved in a text file named after the URL\_ID from the Excel file.
2. **Data Analysis:**
   * **Sentiment Analysis:**
     + Sentiment analysis is performed using TextBlob, which calculates the polarity and subjectivity of the text.
     + Positive and negative sentiment scores are inferred from the polarity score.
   * **Readability Metrics:**
     + Several readability metrics are computed, including the Gunning Fog Index, average sentence length, percentage of complex words, and others.
     + The number of syllables in each word is calculated using a custom function.
     + Complex words (words with three or more syllables) are identified.
   * **Additional Textual Features:**
     + Metrics such as word count, syllable count per word, personal pronouns, and average word length are calculated.
3. **Output:**
   * The computed variables are saved in a structured format in an Excel file (Output.xlsx) according to the provided output structure.

**2. How to Run the Script**

**Steps to Run:**

1. **Setup:**
   * Ensure that Python 3.x is installed on your machine.
   * Install the required libraries (listed below) using pip.
2. **Dependencies Installation:**
   * Open a terminal or command prompt.
   * Navigate to the directory containing the script assignment.py.
   * Run the following command to install dependencies:
     + py -m pip install -r requirements.txt
     + py -m pip install openpyxl beautifulsoup4 requests textblob
3. **Dependencies Installation:**
   * Ensure the Input.xlsx file is present in the same directory as assignment.py.
   * Run the script by executing:
     + py assignment.py
   * The script will:
     + Extract text from each URL.
     + Perform sentiment analysis and readability analysis.
     + Save the results in Output.xlsx in the same directory.

**Expected Output:**

The script will generate two main outputs:

* 1. Text files for each article, named using the URL\_ID.
  2. An Excel file (Output.xlsx) containing the analysis results.

**3. Dependencies Required**

Ensure the following Python libraries are installed:

* **openpyxl**: For reading and writing Excel files.
* **beautifulsoup4**: For parsing HTML and extracting data.
* **requests**: For making HTTP requests to retrieve web pages.
* **TextBlob**: For performing sentiment analysis.