**1. Overview of the Approach**

The objective of this project was to extract article content from the URLs provided in the input.xlsx file, analyze the extracted text, and compute various sentiment and readability metrics.

1. **Web Scraping**:
   * **Input**: The input.xlsx file contains a list of URLs.
   * **Process**: For each URL, the script uses the **BeautifulSoup** and **Requests** libraries to scrape the article's title and body text. It ensures that only the main content of the article is extracted (excluding headers, footers, and other irrelevant sections).
   * **Output**: The article content is saved in .txt files. Each file is named using the URL\_ID from the input.
2. **Text Analysis**:
   * **Input**: The text files created by the web scraping process.
   * **Process**: The script analyzes the text content to compute several variables, including sentiment scores (positive, negative, polarity, subjectivity) and readability metrics (e.g., average sentence length, fog index, complex words).
   * **Output**: The results are stored in a CSV file (text\_analysis\_results.csv) containing the analyzed data in tabular format.

**Key Variables Calculated:**

* **Sentiment Analysis**:
  + Positive Score, Negative Score
  + Polarity Score (between -1 to 1)
  + Subjectivity Score (between 0 to 1)
* **Readability Analysis**:
  + 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
  + Average Word Length

**2. Dependencies Required**

To run the Python script, you need to install the following libraries:

* **BeautifulSoup4**: To scrape HTML content from web pages.
* **Requests**: To make HTTP requests and fetch the web page content.
* **TextBlob**: To perform sentiment analysis.
* **NLTK**: To process text and calculate certain readability metrics.
* **Pandas**: To save the output data into a CSV file.

**How to Install the Dependencies:**

Before running the script, you must install the required dependencies. Open your terminal (or command prompt) and run the following commands:

pip install beautifulsoup4

pip install requests

pip install textblob

pip install nltk

pip install pandas

These commands will install all the necessary libraries.

**3. Running the Python Script**

1. **Download the Files**:
   * Make sure you have the input.xlsx file, which contains the URLs to scrape.
   * Create a folder where the scraped text files will be saved.
   * Update the script to use the correct file paths (e.g., location of input.xlsx and the folder for saving .txt files).
2. **Save the Python Script**:
   * Save the provided Python code into a .py file. For example: web\_scraping\_and\_analysis.py.

**Running the Script**

Once everything is set up, you can run the script using the following steps:

1. Open the terminal (or command prompt).
2. Navigate to the directory where the Python script is saved.
3. Run the script using the following command: python web\_scraping\_and\_analysis.py

**Expected Output**

* **Scraped Text Files**:
  + After running the script, a folder will be created that contains .txt files for each article scraped.
  + Each text file is named according to the URL\_ID, which corresponds to the entry in the input.xlsx file.
* **Analysis Results**:
  + The script will also generate a **text\_analysis\_results.csv** file in the same directory.
  + The CSV file will contain the following columns:
    - URL\_ID: Unique identifier for each URL.
    - Sentiment and readability metrics.