Approach to the Solution:

- **Data Extraction:**
  - Utilized BeautifulSoup for web scraping.
  - Identified three types of articles during extraction:
    - Empty articles (URL no. 36, 50)
    - Two other types with distinct HTML structures, handled separately for flawless text extraction.
  - Saved extracted text in .txt files named after "URL_ID" from the Input excel sheet.
- **Text Analysis:**
  - Developed three functions for cleanliness and modularity:
    - **"perform_sentiment_analysis":** Conducted sentiment analysis.
    - **"perform_readability_analysis":** Performed readability analysis.
    - **"perform_basic_analysis":** Undertook the rest of the text analysis.
  - Common approach in each function: tokenization, cleaning, and analysis.
  - Returned analyzed values at the end of each function.
- **Function Call and Output Generation:**
  - Loaded the Output excel sheet into a DataFrame.
  - In a loop, extracted URL_ID and created file paths for extracted text files.
  - Read the corresponding text file and called all three functions for analysis.
  - Written the returned values into the DataFrame.
  - Finally, updated the Output excel sheet with the DataFrame.

Dependencies:

- **BeautifulSoup4:** Used for web scraping and extracting data from HTML
- **NLTK:** Natural Language Toolkit for tokenization, stop words, and text analysis.

Instructions to Run the .py File:

- Ensure all dependencies are installed using the provided commands.
- Replace directory paths in the script with your system-specific paths. All the required path related variables are listed at the start of the code.
- Ensure that all the input files namely: “MasterDictionary”, “StopWords”, “Input”, “Output Data Structure” are present in your system.

Additional Notes:

- The script is designed to handle different HTML structures for articles.
- Extracted text is saved in .txt files named after "URL_ID."
- Functions are modular for easy maintenance and readability.
- The output Excel sheet is updated with sentiment, readability, and basic analysis.