Instructions for Data Extraction and Analysis with NLP

Approach to the Solution:

1. **Data Extraction:**
   - Utilized the Python programming language for data extraction and analysis.
   - Used Beautifulsoup library for web datascraping.
2. **Sentiment Analysis:**
   - Loaded stop words from given files to create set for filtering.
   - Employed the NLTK library for natural language processing tasks such as tokenization.
   - Utilized pre-existing dictionaries for positive and negative sentiment words.
   - Calculated positive and negative scores based on predefined sentiment dictionaries.
   - Computed polarity and subjectivity scores using the obtained positive and negative scores.
3. **Additional Variables Calculation:**
   - Computed average sentence length, percentage of complex words, and Fog Index for each document.
   - Measured the number of personal pronouns and calculated average word length.
   - Consider all instructions given in text analysis file for calculating variables.
   - Extracted information such as URL and URL_ID to associate sentiment scores with specific documents.

Running the .py File:

1. **Environment Setup:**
   - Ensure that Python is installed on system.
   - Install required libraries.
2. **Execution:**
   - Place the provided .py file in the desired directory.
   - Upload Input file and other Stopwords and MasterDictionary files in the same directory containing the .py file.
   - Run the script
3. **Output:**
   - The output will be saved in a CSV file named **output_results1.csv**.
   - The CSV file will contain sentiment scores, additional variables, and relevant information for each URL.

Dependencies:

- Python
- NLTK (Natural Language Toolkit)
- Pandas
- BeautifulSoup
- Requests
- Os
- Nltk.tokenize
- Textstat