Approach to the Solution

Name: Abhay Biradar

Objective

The goal was to perform automated textual analysis on articles retrieved from URLs provided in an input Excel file. Each URL was associated with a unique URL ID, and the results needed to be stored in an output file with the columns ordered as **URL ID**, and **URL**, followed by computed metrics.

Steps in the Solution

1. Data Input Handling

- **Input File**: The script reads the input Excel file (`input.xlsx`) using the `pandas'` library. The file contains two columns: **URL ID** and **URL**.
- **Column Extraction**: Both `URL ID` and `URL` are extracted for further processing.

---

2. Data Extraction

- **Web Scraping**: For each URL, the script uses the `requests` library to fetch the webpage content.
- **Text Parsing**: The `BeautifulSoup` library extracts the main text content, including the title and paragraphs.

---

3. Text Cleaning and Preprocessing

- **Stopwords Removal**: The text is tokenized using NLTK's `word_tokenize`, and stopwords are removed using NLTK's `stopwords`.
- **Text Normalization**: Only alphabetic tokens are retained for analysis.

---

4. Feature Computation

- **Sentiment Analysis**:
  - Positive and Negative Scores are computed using predefined dictionaries (`positive-words.txt` and `negative-words.txt`).
  - Polarity and Subjectivity Scores are derived from these scores.
- **Readability Metrics**:
  - Average Sentence Length, Percentage of Complex Words, and Fog Index are calculated.
- **Linguistic Features**:
  - Metrics such as Word Count, Complex Word Count, Syllable Count per Word, Personal Pronouns Count, and Average Word Length are computed.

---

5. Output Generation

- **Results Compilation**:
  - All computed metrics, along with the corresponding **URL ID** and **URL**, are stored in a dictionary for each article.
- **Reordering Columns**:
  - The output columns are arranged to place **URL ID** first, followed by **URL**, and then the computed metrics.

This is achieved with the following code:

Python

```python
columns_order = ["URL ID", "URL"] + [col for col in output_df.columns
if col not in ["URL ID", "URL"]]
output_df = output_df[columns_order]
```

  -
- **Output File**:
  - The results are saved in an Excel file (`output.xlsx`) using `pandas`.

Challenges and Solutions

1. **Unbalanced Parentheses in Regex**:
   - Fixed by refining the regex for personal pronouns with non-capturing groups.
2. **Column Reordering**:
   - Handled dynamically by extracting and reordering columns after processing.
3. **Missing Libraries**:
   - Ensured all required libraries (`nltk`, `pandas`, etc.) were installed and configured.

Outcome

The script successfully processes the input file, analyzes each URL, and generates an output file with results in the desired format. The column order ensures clarity and ease of tracking.