**Instructions Document for Data Extraction and NLP**

**1. Objective**

The purpose of this assignment is to extract textual data from specified URLs and perform text analysis to compute various linguistic metrics as outlined in the provided documentation.

**2. Approach to the Solution**

The solution consists of two main parts: Data Extraction and Data Analysis.

**2.1 Data Extraction**

1. **Input Handling**: The program reads a list of URLs from an Excel file (Input.xlsx), which contains an identifier for each URL.
2. **Web Scraping**: The code uses BeautifulSoup and requests to scrape the article title and text. It specifically targets the main content of the articles while ignoring headers, footers, and advertisements.
3. **File Output**: Extracted articles are saved as text files, named using their corresponding URL\_ID.

**2.2 Data Analysis**

1. **Text Processing**: The program processes each article's text to compute linguistic metrics:
   * Positive and Negative Scores
   * Polarity and Subjectivity Scores
   * Average Sentence Length
   * Percentage of Complex Words
   * FOG Index
   * Complex Word Count, Word Count, Syllable Count
   * Average Word Length and Personal Pronoun Count
2. **Output Generation**: The results are compiled into a structured format according to the specifications in Output.xlsx, and saved as a CSV file.

**3. How to Run the Python File**

1. **Setup Environment**:
   * Ensure you have Python installed (version 3.6 or above recommended).
   * Install required libraries.
2. **Place Input Files**:
   * Ensure the Input.xlsx file is in the same directory as the Python script.
   * Make sure to have “StopWords” directory and “MasterDictionary” directory in the same directory you are working in.
3. **Run the Script**:
   * Execute the Python script using the command line:

python NLP\_Analysis.py

* + After execution, the extracted articles will be saved as text files, and the analysis results will be saved as Output.csv.

**4. Dependencies Required**

* **Python**: Version (Any will work, however I used 3.11.9)
* **Libraries**: You can do this by running pip command on below given libraries
  + beautifulsoup4==4.12.0
  + nltk==3.8.1
  + pandas==1.5.3
  + numpy==1.24.3
  + requests==2.31.0
  + tqdm==4.64.1
  + xlrd==2.0.1
  + openpyxl

**Excel Files**:

* + Input.xlsx: Should contain the URL\_ID and URLs for data extraction.

**5. Output Files**

* **Text Files**: Each article will be saved in a text file named with its corresponding URL\_ID.
* **Output CSV**: The results of the text analysis will be saved in a file named Output.csv.