**Instructions for Data Extraction and NLP Assignment**  
 **Prepared by: Sujit Kumar**

**1. Objective**

The objective of this assignment is to extract textual data from 147 article URLs provided in **Input.xlsx**, perform sentiment and readability analysis, and save results in the required format as per **Output Data Structure.xlsx**.

**2. Approach**

**Data Extraction**

* Loaded URLs from **Input.xlsx**.
* Extracted the article title and body text using Python libraries **requests** and **BeautifulSoup**.
* Saved each article as a separate text file named <URL\_ID>.txt.
* Zipped all text files into **Articles.zip**.

**Data Cleaning**

* Converted text to lowercase.
* Removed punctuation and numbers using **regex**.
* Removed stopwords using the provided **StopWords** files.

**Text Analysis**

* Counted positive and negative words using the given **MasterDictionary**.
* Calculated the following metrics for each article:
  + Positive Score
  + Negative Score
  + Polarity Score
  + Subjectivity Score
  + 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
  + Average Word Length

**Output Generation**

* Stored the calculated metrics into **Final\_Output.xlsx** in the same format as **Output Data Structure.xlsx**.

**3. How to Run**

1. Install required libraries:

bash

CopyEdit

pip install pandas requests beautifulsoup4 nltk openpyxl

1. Download NLTK resources (first time only):

python

CopyEdit

import nltk

nltk.download('punkt')

nltk.download('punkt\_tab')

1. Ensure the following files are in the same folder:
   * assignment.py
   * Input.xlsx
   * Output Data Structure.xlsx
   * StopWords folder
   * positive-words.txt
   * negative-words.txt
2. Run the script:

bash

CopyEdit

python assignment.py

1. After execution, the following outputs will be generated:
   * **Final\_Output.xlsx** → containing all sentiment and readability metrics.
   * **Articles.zip** → containing all extracted articles as <URL\_ID>.txt.

**4. Dependencies**

* Python 3.9 or above
* Libraries:
  + pandas
  + requests
  + beautifulsoup4
  + nltk
  + openpyxl
  + re (built-in)
  + zipfile (built-in)

**5. Deliverables**

* assignment.py
* Final\_Output.xlsx
* Articles.zip
* Instructions.docx