Tab 1

**Data Extraction and NLP**

**Objective**

The goal of this assignment is to extract textual content from a list of URLs, perform sentiment and readability analysis, and save the results in a structured Excel.

**Approach summary**

The solution is implemented in two main stages:

1. Data Extraction

* Loads `Input.xlsx` to retrieve `URL\_ID` and article URLs.
* Crawls each URL using `requests` and `BeautifulSoup`.
* Extracts the article title and main content only.
* Saves each article as a `.txt` file using `URL\_ID` as the filename.

2. Text Analysis

* Loads word lists from `MasterDictionary`:
* `positive-words.txt`
* `negative-words.txt`
* Tokenizes and cleans each article:
* Removes stopwords using NLTK.
* Computes sentiment scores (positive, negative, polarity, subjectivity).
* Calculates readability metrics (fog index, complex words, sentence length).
* Analyzes structural metrics (syllables, word count, personal pronouns, word length).
* Outputs the results to `final\_output.xlsx` in the exact required structure.

---

**How to Run**

1. Install Required Dependencies or import important libraries

Like pandas, requests, beautifulsoup, nltk, textstat, re, os, word\_tokenize, sent\_tokenize.

and also download some nltk resources (punkt, stopwords)

2. Now load the ‘Input.xlsx’ file.

3. After that, perform the data extraction method like:-

* Function to extract title and article text.
* Extract data for each URL.
* Save extracted data to a Dataframe.
* Save extracted text into ‘.txt’.

4. Create/Download/load the MasterDictionary folder containing the positive and negative word

list.

5. Now do the Text Analysis part like:-

* Load positive and negative word lists.
* Start Preprocessing the Text (Tokenize, Stopword Removal)
* Extract drive variables like:-
* Sentiment Analysis (Positive Score, Negative Score, Polarity Score, Subjectivity Score)
* Readability Metrics (Average Sentence Length, Percentage of Complex Words, Fog Index)
* Structural Metrics (Complex Word Count, Syllables Per Word, Personal Pronoun, Average Word Length)
* Return Result as List.

6. Define the output structure and save it as an Excel file.