### **README FILE**

### **Approach to the Solution:**

I employed a Python script that accomplishes several tasks:

1. Data Extraction:
   1. Utilized Google Colab for environment setup and mounted Google Drive for file access.
   2. Installed necessary libraries (requests, beautifulsoup4, pandas, numpy, nltk, spacy, textstat, XlsxWriter).
2. Web Scraping:
   1. Read URLs from an Excel file and extracted content from respective web pages.
   2. Saved web content in text files named after their URL\_ID.
3. Text Processing:
   1. Removed punctuation, tokenized text, and filtered stop words from scraped content.
   2. Calculated various metrics like Positive Score, Negative Score, Polarity Score, Subjectivity Score, Average Sentence Length, Percentage of Complex Words, Fog Index, etc.
4. Excel Data Preparation:
   1. Created a structured data set with the extracted metrics for each URL.
   2. Loaded an existing Excel file and updated it with the collected data.

### **How to Run the .py File to Generate Output:**

1. Dependencies:
   1. The code requires the installation of specific Python libraries using pip:
      1. requests, beautifulsoup4, pandas, numpy, nltk, spacy, textstat, XlsxWriter, openpyxl.
2. Execution:
   1. The code is designed for Google Colab and might need adjustments for local execution due to file path references.
   2. Ensure that the necessary files (Input.xlsx, stop word files, positive/negative word lists) are stored in the specified locations on Google Drive.
3. Run Instructions:
   1. Open the script in a suitable environment (Google Colab or local Python environment).
   2. Make sure the dependencies are installed.
   3. Update file paths if necessary.
   4. Execute the script.

### **Dependencies Required:**

1. The code relies on the following libraries/packages:

1.1. requests, beautifulsoup4, pandas, numpy, nltk, spacy, textstat, XlsxWriter, openpyxl.

### **Note:**

1. Google Colab Usage:
   1. Ensure access to Google Drive and file paths according to the mounted drive in the Colab environment.
2. File Paths:
   1. Update the file paths (Input.xlsx, stop word files, positive/negative word lists) to match your directory structure.
3. Error Handling:
   1. Error handling is implemented for fetching web content; however, ensure robustness in handling various exceptions.