Instructions Documentation

1. **data_extration.py** This python code used to scrape text from URLs listed in an Excel file and save the extracted content to text files.
    - Used **requests** library to library to fetch the webpage's content.
    - parsed HTML content using BeautifulSoup4
    - The clean text is extracted and saved into a text file, named after the URL_ID from the Excel file.
2. **Textual_analysis.py** This python performs various textual analyses on the extracted content and writes the computed metrics to an Excel sheet.
    - Tokenizing sentences and words using NLTK's sent_tokenize and word_tokenize.
    - Removing stopwords loaded from multiple files.
    - Calculating the metrics like POSITIVE SCORE , NEGATIVE SCORE , POLARITY SCORE, SUBJECTIVITY SCORE etc.
    - Added checks if the filename matches an entry in Column A before writing computed metrics.
3. **Challenges**
    - have encountered issues with **pkg_resources** while using **syllapy**. Tried troubleshooting by checking the Python environment and making sure all necessary libraries were installed.
    - ensuring directories exist before writing files
4. **Next Step**

    Deeper NLP Techniques: Can be used more advanced text processing techniques, such as using machine learning models for sentiment analysis.

How to run the .py file to generate output

1. **data_extration.py** → This code is used to extract the article text and save the extracted article in a text file with **URL_ID** as its file name.

    - **Install Required Libraries:**

        python3 install requests bs4 pandas openpyxl

    - **Run 'data_extraction.py'**

        python3 data_extraction.py

1. **textual_analysis.py** →

    - **Install Required Libraries:**
        python3 -m pip install nltk syllapy setuptools

    - **Run textual_analysis.py**
        python3 data_extraction.py