**Text Analysis Project Documentation**

**1. Approach to the Solution**

**Overview**

This project is designed to perform text analysis on a collection of text files. The analysis includes various linguistic metrics such as sentiment scores, readability indices, word and syllable counts, and more. The results are then saved in an Excel file.

**Steps Taken**

1. **Text Preprocessing**:
   * The text files are read and tokenized into sentences and words.
   * Custom stopwords are removed from the text to focus on meaningful words.
   * Sentences and words are stored separately for further analysis.
2. **Sentiment Analysis**:
   * Positive and negative sentiment scores are calculated using predefined sets of positive and negative words.
   * The polarity score is calculated based on the ratio of positive to negative words.
3. **Readability Metrics**:
   * The project computes the FOG index and other readability metrics based on the number of complex words, sentences, and syllables.
4. **Word and Syllable Count**:
   * The total number of words and syllables are calculated.
   * The percentage of complex words (words with more than three syllables) is determined.
5. **Excel Integration**:
   * The analysis results are merged with an existing Excel file, preserving columns like URL\_ID and URL alongside the new metrics.
6. **Error Handling**:
   * Robust error handling ensures that incorrect data types or missing data do not cause the program to crash. This is especially crucial when merging data frames or performing calculations.

**2. How to Run the .pynb File to Generate Output**

1. **Install Jupyter Notebook:** Ensure you have Jupyter Notebook installed. You can install it using the following command:

pip install notebook

Alternatively, if you're using Anaconda, Jupyter Notebook comes pre-installed.

1. **Launch Jupyter Notebook:**
   * Open your terminal or command prompt.
   * Navigate to the directory containing your .pynb file.
   * Launch Jupyter Notebook by typing:

jupyter notebook

* + This will open the Jupyter Notebook interface in your web browser.

1. **Open the Notebook:**
   * In the Jupyter Notebook interface, navigate to the .pynb file you want to run.
   * Click on the file name to open it.
2. **Run the Cells:**
   * To run the entire notebook, go to the menu at the top and select **Cell** -> **Run All**.
   * Alternatively, you can run individual cells by selecting the cell and pressing Shift + Enter.
3. **Generating the Output:**
   * The notebook will execute the code in each cell sequentially.
   * Once the notebook is run, the results (e.g., analysis metrics, dataframes) will be displayed directly beneath each cell.
   * If the notebook saves output to a file (like an Excel or CSV file), you can find the generated files in the same directory as your notebook.
4. **Saving Your Work:**
   * After running the notebook, you can save your work by clicking **File** -> **Save and Checkpoint** or by pressing Ctrl + S.
   * To export the notebook in a different format (e.g., PDF, HTML), go to **File** -> **Download as** and select the desired format.
5. **Dependencies:** Ensure all necessary dependencies are installed before running the notebook. You can install them using:

pip install -r requirements.txt

If a requirements.txt file is not provided, install the dependencies mentioned in the "Dependencies" section of this documentation.

By following these steps, you should be able to successfully run the .pynb file and generate the desired output.

**3. Dependencies**

**Required Python Libraries**

* **Pandas**: For data manipulation and merging results into the Excel file.
* **Numpy**: For numerical operations.
* **Openpyxl**: For working with Excel files.

**Installation of Dependencies**

To install the required dependencies, use the following pip commands:

pip install pandas numpy openpyxl

Make sure all these packages are installed before running the script to avoid import errors.

This document should help you understand the approach taken in the project, guide you through running the script, and ensure you have all the necessary dependencies installed.