Project Report

-Vedant Poal

1. Explaining How the Solution Was Approached

The solution to the text analysis project was approached in a structured manner to ensure that all components were handled efficiently and effectively. Here are the key steps taken:

1. **Data Preparation**:
    - **Input Data**: The input data was loaded from an Excel file (Input.xlsx) containing URLs to be processed.
    - **Stop Words**: Multiple stop words files were loaded to filter out common words that do not contribute to the analysis.
    - **Positive and Negative Words**: Lists of positive and negative words were loaded to perform sentiment analysis.
2. **Text Processing**:
    - **Text Cleaning**: A function was created to clean the text by removing non-alphabetic characters, tokenizing the text, and filtering out stop words.
    - **Syllable Count**: A function was implemented to count syllables in words, with special handling for suffixes like "es" and "ed".
    - **Personal Pronouns**: A regular expression was used to count personal pronouns in the text.
3. **Text Analysis**:
    - **Sentiment Analysis**: The script calculated positive and negative scores, polarity, and subjectivity based on the presence of positive and negative words.
    - **Readability Metrics**: Metrics such as average sentence length, percentage of complex words, Fog Index, and average number of words per sentence were calculated.

- **Word Statistics**: Additional statistics like word count, syllable per word, and average word length were computed.

4. **Error Handling**:
    - The script included robust error handling to manage issues such as file encoding problems and missing NLTK data packages.
    - Errors during URL processing were logged, and the script continued processing other URLs.
5. **Output Generation**:
    - The analyzed data was compiled into a Data Frame and saved to an Excel file (Output Data Structure.xlsx).
    - Text files containing the raw article data were saved for each URL.

2. How to Run the .py File to Generate Output

To run the .py file and generate the output, follow these steps:

1. **Install Required Libraries**: Ensure you have the necessary Python libraries installed. You can install them using pip:

   pip install pandas requests newspaper3k nltk syllables

2. **Download NLTK Data**: Ensure the required NLTK data packages are downloaded. Run the following commands in your Python environment:

   import nltk

   nltk.download('punkt')

3. **Prepare the Environment**:
    - Ensure the input Excel file (Input.xlsx) and all required text files (stop words, positive/negative words) are in the specified directories.
    - Ensure the output directory (data) exists or the script will create it.
4. **Run the Script**: Navigate to the directory containing the script and run it using Python:

   python app.py

5. **Check the Output**:
    - The processed data will be saved in an Excel file (Output Data Structure.xlsx).
    - Raw article texts will be saved in the data directory.

3. Dependencies Required for the Project

The project relies on the following dependencies:

1. **Python Libraries**:
    - pandas: For data manipulation and analysis.
    - requests: For making HTTP requests.
    - newspaper3k: For extracting articles from URLs.
    - nltk: For natural language processing tasks.
    - syllables: For estimating syllable counts in words.
2. **NLTK Data Packages**:
    - punkt: Required for tokenizing text into sentences.
3. **File Paths**:
    - Ensure the following file paths are correctly set and accessible:
        - Input Excel file: D:/projects/new assignment/Input.xlsx
        - Stop words files: D:/projects/new assignment/StopWords/
        - Positive and negative words files: D:/projects/new assignment/MasterDictionary/
        - Output directory: D:/projects/new assignment/data/
4. **Environment Variables**:
    - Ensure the NLTK_DATA environment variable is set if you are using a custom data directory.

Conclusion

This report outlines the approach taken to solve the text analysis project, how to run the script, and the dependencies required. By following these steps, you should be able to successfully run the script and generate the desired output. If you encounter any issues, ensure all dependencies are correctly installed and file paths are accurately specified.