Sentiment Analysis & Readability Metrics Script Documentation

1. Objective

This script analyzes the extracted article text from `.txt` files to compute various text analysis metrics such as sentiment scores, readability indices, and word complexity. The results are saved into the **Output Data Structure.xlsx** file in the required format.

2. Approach

2.1. Loading Required Data

- Reads `StopWords` from the `StopWords/` directory to filter out unnecessary words.
- Loads `positive-words.txt` and `negative-words.txt` from `MasterDictionary/` for sentiment analysis.
- Uses **spaCy** (`en_core_web_sm`) for tokenization and sentence segmentation.
- Reads the extracted article `.txt` files from the `extracted_articles/` directory.

2.2. Text Processing

- **Tokenization**: Uses `spaCy` to extract words while ignoring punctuation and stopwords.
- **Complex Words**: Identified as words containing more than two syllables.
- **Syllable Count**: Counts vowels while handling exceptions (`es`, `ed` endings).
- **Personal Pronouns**: Identifies occurrences of `I, we, my, ours, us` using regex.

2.3. Sentiment Analysis

- **Positive Score**: Counts occurrences of words found in `positive-words.txt`.
- **Negative Score**: Counts occurrences of words found in `negative-words.txt`.
- **Polarity Score**: `(Positive - Negative) / (Positive + Negative + 0.000001)`.
- **Subjectivity Score**: `(Positive + Negative) / (Total Words after Cleaning + 0.000001)`.

2.4. Readability Metrics

- **Average Sentence Length**: `Total Words / Total Sentences`.
- **Percentage of Complex Words**: `Complex Words / Total Words`.
- **Fog Index**:
  `0.4 * (Avg Sentence Length + Percentage of Complex Words)`.
- **Syllable Count per Word**: `Total Syllables / Total Words`.
- **Personal Pronouns Count**: Regex-based matching.
- **Average Word Length**: `Sum of character count / Total Words`.

2.5. Saving Results to Excel

- Reads `Output Data Structure.xlsx` and merges new computed values.
- Removes duplicate columns (_x, _y) caused by merging.
- Ensures correct column ordering before saving.
- Adjusts **column widths** for better readability.

3. Running the Script

Ensure you have Python installed and run the following command to install required packages:

```
pip install pandas spacy openpyxl
python -m spacy download en_core_web_sm
```

Run the following command in the terminal:

```
python sentiment_analysis.py
```

Ensure `extracted_articles/`, `MasterDictionary/`, and `StopWords/` directories are present.

4. Dependencies & Notes

Package	**Purpose**
`spacy`	Tokenization, sentence segmentation
`pandas`	Data manipulation and merging
`openpyxl`	Excel reading and writing

Package	Purpose
`re`	Regex for personal pronoun detection