**Sentiment Analysis Project Documentation**

**Objective**

The objective of this project is to extract textual data from specified URLs, perform sentiment analysis, and compute various metrics to assess the sentiment and complexity of the text.

**1. Data Extraction**

* **Libraries Used:**
  + pandas: For handling and saving data to Excel.
  + openpyxl: For Excel file operations.
  + requests: For making HTTP requests to fetch web content.
  + beautifulsoup4: For parsing and extracting text from HTML.
* **Process:**
  + Extract text content from specified URLs.
  + Handle various HTML structures to retrieve article text.

**2. Data Analysis**

* **Process:**
  + Load text files from the directory.
  + Process data by parsing and analyzing text files.

**3. Sentiment Analysis**

* **Libraries Used:**
  + nltk: For text processing and sentiment analysis.
  + External Files:
    - positive-words.txt: Contains a list of positive words.
    - negative-words.txt: Contains a list of negative words.
* **Approach:**
  + **Data Extraction:**
    - Extract text content from URLs and handle HTML structure variations.
  + **Text Preprocessing:**
    - Tokenize text into words and sentences.
    - Remove stopwords and punctuation.
    - Convert text to lowercase.
  + **Sentiment Analysis:**
    - Compute sentiment scores using predefined positive and negative word lists.
    - Calculate additional metrics including word count, average word length, and personal pronoun count.
  + **Output Generation:**
    - Save the analysis results to an Excel file (Output Data Structure.xlsx).

**Sentiment Analysis Metrics**

1. **Positive Score**
   * **Definition:** Frequency of positive words in the text.
   * **Calculation:** Count of positive words based on a predefined list.
   * **Purpose:** Indicates the positive sentiment of the article.
2. **Negative Score**
   * **Definition:** Frequency of negative words in the text.
   * **Calculation:** Count of negative words based on a predefined list.
   * **Purpose:** Indicates the negative sentiment of the article.
3. **Polarity Score**
   * **Definition:** Sentiment polarity ranging from -1 (very negative) to 1 (very positive).
   * **Calculation:** Computed using a sentiment analysis library (e.g., TextBlob).
   * **Purpose:** Provides a numerical sentiment orientation.
4. **Subjectivity Score**
   * **Definition:** Degree of subjectivity or objectivity, ranging from 0 (very objective) to 1 (very subjective).
   * **Calculation:** Computed using a sentiment analysis library (e.g., TextBlob).
   * **Purpose:** Reflects the level of personal opinion vs. factual information.
5. **Avg Sentence Length**
   * **Definition:** Average number of words per sentence.
   * **Calculation:** Total number of words divided by the number of sentences.
   * **Purpose:** Helps understand text readability and complexity.
6. **Percentage of Complex Words**
   * **Definition:** Percentage of words with three or more syllables.
   * **Calculation:** (Number of complex words / Total number of words) \* 100
   * **Purpose:** Indicates language complexity.
7. **Fog Index**
   * **Definition:** Readability index estimating the years of formal education needed.
   * **Calculation:** 0.4 \* [(Average Sentence Length) + (Percentage of Complex Words)]
   * **Purpose:** Assesses text readability.
8. **Avg Number of Words Per Sentence**
   * **Definition:** Average number of words in each sentence.
   * **Calculation:** Total number of words divided by the total number of sentences.
   * **Purpose:** Provides insight into sentence structure.
9. **Complex Word Count**
   * **Definition:** Number of words with three or more syllables.
   * **Calculation:** Count of complex words.
   * **Purpose:** Indicates advanced vocabulary usage.
10. **Word Count**
    * **Definition:** Total number of words in the text.
    * **Calculation:** Count of all words.
    * **Purpose:** Measures text length.
11. **Syllables Per Word**
    * **Definition:** Average number of syllables per word.
    * **Calculation:** Total number of syllables divided by the total number of words.
    * **Purpose:** Assesses vocabulary complexity.
12. **Personal Pronouns**
    * **Definition:** Count of personal pronouns (e.g., I, we, my).
    * **Calculation:** Count occurrences of personal pronouns.
    * **Purpose:** Reflects personal perspective in the text.
13. **Avg Word Length**
    * **Definition:** Average number of characters per word.
    * **Calculation:** Total number of characters divided by the total number of words.
    * **Purpose:** Provides insight into word complexity.

**Running the Script**

1. **Environment Setup:**
   * Ensure Python and required libraries are installed (requests, BeautifulSoup4, nltk, textblob, pandas, etc.).
2. **File Setup:**
   * Place text files (positive-words.txt, negative-words.txt) in the same directory as the script.
3. **Execute the Script:**
   * Run the Python script using the command: python script\_name.py.
   * The script processes URLs, extracts article content, performs sentiment analysis, and saves results to Output Data Structure.xlsx.
4. **Dependencies:**
   * **Libraries:**
     + requests
     + BeautifulSoup4
     + nltk
     + textblob
     + pandas
   * **Files:**
     + positive-words.txt
     + negative-words.txt
   * **Output File:**
     + Output Data Structure.xlsx

**Notes**

* Verify that URLs are valid and accessible.
* Ensure all required text files are present and correctly formatted.