Instructions for Internship Assignment

1. Introduction:

Welcome to the instructions document for the internship assignment. This document provides guidance on how to use and reproduce the text analysis code and output.

2. Dependencies:

- Python 3.7 or higher

- Libraries: [List the required libraries and their versions]

  - requests

  - BeautifulSoup

  - os

  - pandas

 - nltl

 - glob

- string

Please make sure to install the required libraries before running the code.

3. File Structure:

 - Main Python File: BlackCoffer-Jeevak Moon.ipynb

   - Purpose: Contains the code for extracting textual data from the given URL and performing text analysis.

 - Output File: Output Data Stucture.xlsx

   - Purpose: Stores the analysed results in Excel format.

4. Instructions for Running the Code:

 1. Open a terminal or command prompt.

 2. Navigate to the directory where the BlackCoffer-Jeevak Moon.ipynb file is located.

 3. Run the following command:

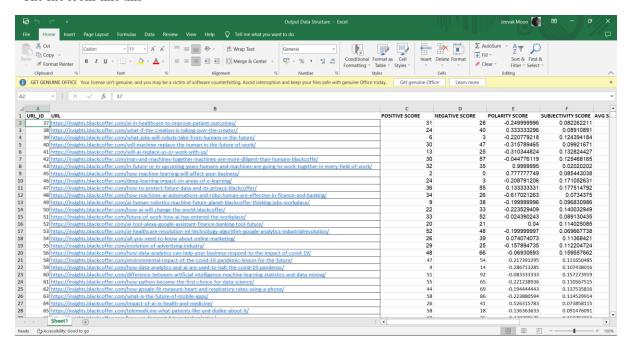   ```

    python BlackCoffer-Jeevak Moon.ipynb

    ```

 4. If applicable, provide the URL for the text data extraction when prompted.

 5. Wait for the code to execute and complete the text analysis.

 6. Once the analysis is finished, the output file Output Data Structure.xlsx will be generated in the same directory.

## 5. Output Structure:

- The output file Output Data Stucture.xlsx will have the following structure:

- The Output Data Structure.xlsx file will contain columns

["URL_ID","URL"," POSITIVE SCORE"," NEGATIVE SCORE"," POLARITY SCORE"," SUBJECTIVITY SCORE"," AVG SENTENCE LENGTH"," PERCENTAGE OF COMPLEX WORDS"," FOG INDEX"," AVG NUMBER OF WORDS PER SENTENCE"," COMPLEX WORD COUNT"," WORD COUNT"," SYLLABLE PER WORD"," PERSONAL PRONOUNS"," AVG WORD LENGTH"]

- The Column name is self-explanatory of the information it is carrying

- The file looks like this



## 6. Expected Results:

- The expected result is the result of the text analysis done on the extracted text from the webpages saved in the exact order as expected

## 7. Additional Notes:

- In the Input.xlsx file there were few webpages who gave error 404 while performing the assignment. So, to handle those corrupted data I removed those links from the data frame and manually removed from the Output Data Structure.xlsx file for smooth run of the process.

- The path of the folders/files specified are according to my system. Please ensure to change them according to your system.

## 8. Contact Information:

If you have any questions or need further assistance, please feel free to contact:

- Name: Jeevak Moon

- Email: jeevakmoon03@gmail.com

- Phone: +917058470924

Please ensure that you carefully follow these instructions to reproduce the text analysis results accurately.