Instructions

**1) Explaining how you approached the solution**

a.) Imports necessary libraries such as pandas, BeautifulSoup, requests and nltk.

b.) Reads data from an Excel file named "Input.xlsx" containing URLs and their corresponding IDs.

c.) Extracts text content from each URL and saves it in separate text files.

d.) Store the title in title as string and article in article and store both title and article in same file called TitleText

e) Now set the directory of textfile , stopword and Masterword as text\_dir , stopword\_dir and sentiment\_dir respectively

f) Store the stopword\_dir data in stop-word as set and store the text\_dir data in docs as list after tokenization and filtred the data.

g) Store the positive and negative word from sentiment\_dir

h) Similarly find all the feature given in the Objective.docx file

i) After analysis , Find the some url which produced 404 ERROR like url\_id 38 and 49 . After removal of this row , Merge the all list file as DataFrame in .CSV file formate. And store final result in Output\_Data.csv file.

e.) Stores the final ouput data in a CSV file named "Output\_Data.csv".

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

To run the provided Python script and generate the output, follow these steps

1. Make sure you have the required dependencies installed. You can install them using pip

**pip install pandas beautifulsoup4 requests nltk**

1. Ensure you have the input file "Input.xlsx" in the same directory as the Python script.
2. Execute the Python script by running the command:

**python blackcoffer\_test\_assignment.py**

### Dependencies Required:

pandas

BeautifulSoup (from bs4)

requests

nltk

stopword

numpy

re

You can install these dependencies using pip

pip install pandas beautifulsoup4 requests nltk

THANKS

AMIT KUMAR