**INSTRUCTIONS**

How I approached the solution

1. First I retrieved all the data from Input.xlxs file using openpyxl.
2. Extracted data from web pages of the given url using bs4 BeautifulSoup and stored in a text file of name equal to url\_id.
3. Then I created a list of all the stop words given in the files.
4. Clean the content of each file – removing punctuations and removing stop words and then tokenizing.
5. Sentiment Analysis –
6. Stored all the positive words into a list from positive-words.txt file
7. Similarly, stored all the negative words into a list from negative-words.txt file
8. Calculated positive scores, negative scores and then polarity score and subjectivity for each file
9. Readablity analysis –

using spacy and textstat modules calculated variables like

* 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

1. Then stored these score values inside the OutputDataStructure.xlsx file in the respective columns.

How to run the .py file

1. Open the folder in VS code or any other python supporting editor
2. Set up the python environment by installing a python interpreter of version 3 and above
3. Install python extension in VS code

Follow the below link for detailed information

<https://code.visualstudio.com/docs/python/python-tutorial>

1. After setting up the environment, open main.py file in the editor, run the following commands in the terminal

pip install openpyxl

pip install bs4

pip install spacy -U

pip install textstat -U

python -m spacy download en\_core\_web\_sm

1. Run the main.py file and see the output in OutputDataStructure.xlsx file

Dependencies required

1. openpyxl
2. requests, bs4, BeautifulSoup
3. spacy, textstat