Project Title	Web Scraping and Text Analysis of News Articles
Technologies	Data Cleansing, EDA, NLP
Domain	Data Science

Problem Statement:

The primary goal of this assignment is to develop a systematic process for retrieving textual data from a specified URL, which contains articles or written content. Once this data is collected, the next step is to conduct thorough text analysis to extract valuable insights and information. The assignment focuses on computing specific variables that have been defined below.

File Description:

Extraction Data:

Input.xlsx

For each of the articles, given in the input.xlsx file, extract the article text and save the extracted article in a text file with URL_ID as its file name.

While extracting text, please make sure your program extracts only the article title and the article text. It should not extract the website header, footer, or anything other than the article text.

Note: YOU CAN USE BEAUTIFULSOUP, SELENIUM OR SCRAPY, OR ANY OTHER PYTHON LIBRARIES THAT YOU PREFER FOR DATA CRAWLING.

Data Analysis:

For each of the extracted texts from the article, perform textual analysis and compute variables, given in the output structure excel file. You need to save the output in the exact order as given in the output structure file, “Output Data Structure.xlsx”

NOTE: YOU MUST USE PYTHON PROGRAMMING FOR THE DATA ANALYSIS

Variables:

Definition of each of the variables given in the “Text Analysis.docx” file.

Look for these variables in the analysis document (Text Analysis.docx):

1. POSITIVE SCORE
2. NEGATIVE SCORE

3. POLARITY SCORE
4. SUBJECTIVITY SCORE
5. AVG SENTENCE LENGTH
6. PERCENTAGE OF COMPLEX WORDS
7. FOG INDEX
8. AVG NUMBER OF WORDS PER SENTENCE
9. COMPLEX WORD COUNT
10. WORD COUNT
11. SYLLABLE PER WORD
12. PERSONAL PRONOUNS
13. AVG WORD LENGTH

Output Data Structure:

Output Variables:

1. All input variables in “Input.xlsx”
2. POSITIVE SCORE
3. NEGATIVE SCORE
4. POLARITY SCORE
5. SUBJECTIVITY SCORE
6. AVG SENTENCE LENGTH
7. PERCENTAGE OF COMPLEX WORDS
8. FOG INDEX
9. AVG NUMBER OF WORDS PER SENTENCE
10. COMPLEX WORD COUNT
11. WORD COUNT
12. SYLLABLE PER WORD
13. PERSONAL PRONOUNS
14. AVG WORD LENGTH

Checkout output data structure spreadsheet for the format of your output, i.e. “Output Data Structure.xlsx”.

Dataset:

Dataset_Link: Data_Link

Problem to be answered:

This assignment aims to equip you with practical skills in web scraping, text processing, and analysis. By performing text analysis and computing relevant variables, you gain insights into the content's characteristics, sentiment, and themes. These skills are valuable in fields such as data science, natural language processing, and content analysis, where understanding and deriving meaning from textual data are essential tasks.

Note:

After completion of all the task you need to create a PowerPoint presentation

That should contain the:

1. Problem Statement
2. Tools Used
3. Approaches
4. EDA Insights

Project Evaluation metrics:

- Project evaluation will be done in the live session and have to showcase the approaches done to complete the project
- You are supposed to write a code in a modular fashion (in functional blocks)
- Maintainable: It can be maintained, even as your codebase grows.
- Portable: It works the same in every environment (operating system)
- You have to maintain your code on GitHub.(Mandatory)
- You have to keep your GitHub repo public so that anyone can check your code.(Mandatory)
- Proper readme file you have to maintain for any project development(Mandatory)
- You should include basic workflow and execution of the entire project in the readme file on GitHub
- Follow the coding standards: https://www.python.org/dev/peps/pep-0008/