Data Extraction and NLP Assignment

# How to Run

1. Ensure you have Python installed.

2. Install required libraries:

pip install requests beautifulsoup4 pandas nltk textstat openpyxl

3. Place the Input.xlsx file in the same directory as data\_extraction.py.

4. Run the script:

Data Extraction and NLP.py

5. The output will be saved as Output Data Structure.xlsx.

# Approach

1. Data Extraction: Used BeautifulSoup to extract article text and title from given URLs.

2. Text Analysis: Computed various text analysis metrics such as positive/negative scores, polarity, subjectivity, etc., using Python and relevant libraries.

# Dependencies

requests: Used to send HTTP requests to retrieve the HTML content of web pages.

beautifulsoup4: Used for parsing HTML and extracting the article title and text from the web pages.

Pandas: Utilized to read the input Excel file and manage the data efficiently.

nltk: Leveraged for text processing tasks such as tokenization and stopword removal.

textstat: Used for calculating readability scores and other text complexity metrics.

openpyxl: Employed to read from and write to Excel files, ensuring proper data structure for input and output.