**Data Extraction and NLP**

**Instructions**

# **Approach to solution:**

**Data Extraction:**

* The Python script starts by reading the URLs and URL\_IDs from the 'Input.xlsx' file using the Pandas library.
* It then uses the requests library to retrieve the HTML content from each URL.
* BeautifulSoup is employed to parse the HTML and extract relevant information such as the article title and content.
* Extracted data is then saved to text files with filenames corresponding to the URL\_ID in the 'output\_file' directory.

**Data Analysis:**

* The analysis begins by reading the input Excel file ('Input.xlsx') to obtain the list of URLs and URL\_IDs.
* NLTK (Natural Language Toolkit) is utilized for tasks such as tokenization and sentiment analysis.
* Stop words are loaded from the 'StopWords' folder, and positive/negative words are loaded from the 'MasterDictionary' folder.
* The script defines functions to calculate derived variables, including sentiment scores, readability metrics, and linguistic features.
* It then iterates over the saved text files, analyzes each article's content, and computes the required variables.
* The final results are merged with the original input DataFrame, and the output is saved to 'Output\_Data.xlsx'.

# **How to run the .py file to generate output?**

**Data Extraction:**

* Ensure you have Python installed on your system.
* Install required libraries using: pip install pandas requests beautifulsoup4.
* Execute the script 'Data\_Extraction.py' using a command like python Data\_Extraction.py.
* The extracted data will be saved in the 'output\_file' directory.

**Data Analysis:**

* Ensure you have Python installed on your system.
* Install required libraries using: pip install pandas requests beautifulsoup4.
* Execute the script 'Data\_Analysis.py' using a command like python Data\_Analysis.py.
* The extracted data will be saved in the 'output\_file' directory.

# **Dependencies required**

**Data Extraction:**

* **pandas:** Used for handling DataFrames and reading/writing Excel files.
* **requests:** Utilized for making HTTP requests to retrieve webpage content.
* **beautifulsoup4:** Employed for HTML parsing and extraction of specific elements.
* **os:** Used for creating directories and handling file paths.

**Data Analysis:**

* **os:** Used for working with file paths and directories.
* **pandas:** Utilized for handling DataFrames and reading/writing Excel files.
* **nltk:** The primary library for natural language processing tasks, including tokenization and sentiment analysis.

The extracted data will be saved in the 'output\_file' directory.