1. **Project Overview**

This project involves extracting article text from various URLs and performing sentiment analysis and text statistics. The tasks are divided into two main Python scripts:

* **Text Extraction Script:** Extracts text from URLs and saves it to a local directory.
* **Text Analysis Script:** Analyzes the extracted text for sentiment, complexity, and other linguistic features.

1. **Approach**

* **Text Extraction:**

Objective: Extract article text from a list of URLs provided in an Excel sheet.

Method:

A loop iterates through each URL, sending an HTTP request to fetch the page content.

BeautifulSoup is used to parse the HTML and extract text from relevant tags.

The script attempts to extract the text up to 5 times in case of request failures.

Extracted text is saved in a .txt file, named according to the URL\_ID in the input Excel sheet.

* **Text Analysis:**

Objective: Perform linguistic and sentiment analysis on the extracted text.

Method:

Text is tokenized, cleaned, and then analyzed to calculate sentiment scores (positive, negative, polarity, subjectivity) and text statistics (word count, sentence length, fog index, etc.).

Results are saved into a new Excel sheet, which includes both the original input data and the computed analysis metrics.

1. **Dependencies**

Make sure to have the following dependencies installed:

pip install pandas beautifulsoup4 requests nltk openpyxl

Note: openpyxl is required for reading/writing Excel files.

1. **How to Run the Scripts**

* **Text Extraction:**

1. Place the Input.xlsx file containing URLs in the same directory as the script.
2. Create a directory named Extracted\_Texts in the same folder if it doesn’t already exist.
3. Run the text\_extraction.py script:

python text\_extract.py

This will generate .txt files in the Extracted\_Texts folder.

* **Text Analysis:**

1. Ensure that all required .txt files are present in the Extracted\_Texts folder.
2. Run the text\_analysis.py script:

python text\_analysis.py

This will generate an Output Data Structure.xlsx file containing the analysis results.