# **BLACKCOFFER**

# **Assignment Title: Web Data Extraction and Natural Language Processing for Data Analysis**

**By: BHOOMIKA AGRAHARI**

### **1. Objective**

The objective of this assignment is to extract textual data from a series of provided URLs (websites) and perform Natural Language Processing (NLP) tasks on the extracted text. The goal is to clean, preprocess, and analyze the text data to derive insights such as sentiment scores, readability metrics, word complexity, and more. The final output will provide useful metrics for further analysis or reporting.

### **2. Problem Statement**

Given a list of URLs, the task is to:

* Extract article text and titles from the webpages.
* Ensure that only relevant content (article text and title) is extracted, excluding headers, footers, and advertisements.
* Perform a series of NLP tasks such as sentiment analysis, readability analysis, word count, syllable count, and other text metrics.

### **3. Key Steps Involved**

**3.1 Data Extraction from Websites**

* **Input**: An Excel file containing a list of URL\_ID and their corresponding URLs.
* **Task**: Extract the article title and main content from each URL using web scraping techniques.
* **Libraries/Tools Used**: Python's requests and Selenium for scraping and HTML parsing.

**3.2 Text Cleaning and Preprocessing**

* **Task**: Clean the extracted text by removing stopwords, special characters, and unnecessary white spaces.
* **Tools Used**: Python libraries like nltk for tokenization, stopword removal, and string manipulation.

**3.3 Natural Language Processing Tasks**

The following NLP tasks are performed on the cleaned text:

**3.3.1 Sentiment Analysis**

* **Task**: Calculate positive, negative, polarity, and subjectivity scores.
* **Formula/Tools**: Use sentiment analysis libraries such as TextBlob, VADER, or custom lexicons for financial or other specific domains.

**3.3.2 Readability Analysis**

* **Task**: Calculate readability metrics like the Gunning Fox Index, average sentence length, and percentage of complex words.
* **Formula**: Use the Gunning Fox formula and additional metrics for readability.

**3.3.3 Word and Sentence Metrics**

* **Task**: Compute metrics like word count, sentence count, average number of words per sentence, complex word count, etc.
* **Tools Used**: nltk for tokenization and sentence segmentation.

**3.3.4 Syllable Count and Complex Word Count**

* **Task**: Count the number of syllables per word and identify complex words (words with more than two syllables).
* **Custom Code**: Implement a function to count vowels and handle exceptions like words ending in "es" or "ed".

**3.3.5 Personal Pronouns**

* **Task**: Identify and count the usage of personal pronouns like "I", "we", "my", "ours", "us".
* **Regex Matching**: Use regular expressions to accurately identify personal pronouns while excluding irrelevant matches like the country "US".

### **4. Tools and Libraries Used**

* **Python**: The main programming language used for all steps.
* **pandas**: For reading, updating, and saving Excel files.
* **nltk**: For text preprocessing, tokenization, and stopword removal.
* **Selenium**: For web scraping and HTML parsing.
* **TextBlob/VADER**: For performing sentiment analysis.
* **Regex**: For identifying and counting personal pronouns and handling specific text patterns.

### **5. Output and Deliverables**

* **Extracted Text Files**: The cleaned text content from each URL stored in text files, named according to the URL\_ID.
* **Excel Output File**: An updated Excel file (Updated\_output\_data\_structure.xlsx) with all the NLP-derived metrics filled in for each URL\_ID.
* **NLP Metrics**: Final outputs such as sentiment scores, readability scores, word counts, syllable counts, and other text-based metrics.

### **6. Conclusion**

This assignment demonstrates how to automate the process of extracting and analyzing web-based content using NLP techniques. The steps outlined help in transforming unstructured web data into structured insights, which can be used for further data analysis and decision-making.