# Q 1. Explaining how you approached the solution

## ***Dependencies Installation***

The Blackcoffer_Task.py file begins by installing necessary Python packages using pip. These packages include requests, beautifulsoup4, openpyxl, nltk, textblob, and python-docx. These libraries are essential for web scraping, text processing, Excel file handling, and natural language processing tasks.

## ***Function Definitions***

After installing dependencies, the script defines several functions:

- **extract_article_text(url):** This function takes a URL as input, extracts the article title and text content from the webpage using requests and beautifulsoup4 libraries, and returns the extracted data.

- **save_article_to_file(url_id, article_title, article_text):** This function saves the extracted article title and text to a text file named after the URL ID.

- **Analyze_text(article_text):** This function analyzes the linguistic features of the input text using nltk and textblob libraries. It computes metrics such as the number of words, sentences, average word length, parts of speech tags, polarity, and subjectivity.

- **extract_variables(docx_path):** This function extracts variables from a Docx file. It reads each paragraph of the document, splits it by ':', and stores the variable name and value in a dictionary.

- **read_words_from_folder(file_path):** This function reads words from a text file and returns a set of unique words. It handles different encodings to ensure compatibility with various text files.

## ***Main Functionality***

The script's main functionality is divided into two parts:

- The first part extracts text content from web pages specified in an input Excel file (Input.xlsx). It iterates through each row of the Excel file, extracts the article text from the corresponding URL, and saves it to a text file.
- The second part analyzes the text content of multiple text files in a folder, computes various linguistic features and variables, and saves the analysis results to an Excel file (Output Data Structure.xlsx).

# Q 2. How to run the .py file to generate output

***Prerequisites:*** Before running the Blackcoffer_Task.py, ensure you have the following:
Google Colab environment set up.

Necessary input files (Input.xlsx, Text Analysis.docx) and folders (StopWords, MasterDictionary) created in your Google Drive as per the specified paths in the script.

### ***Execution Steps:***

Mount your Google Drive by executing the command drive.mount('/content/drive') in Google Colab.

Upload the provided .py file (e.g., text_analysis.py) to your Google Colab environment.

Run the .py file using the command “!python text_analysis.py.”

The Blackcoffer_Task.py will start executing, processing the input data, performing text analysis, and updating the output Excel file with the calculated variables.

# Q 3. Dependencies required

### ***Installation Required***

Ensure you have access to a Python environment with pip installed.
Execute the command
!pip install requests beautifulsoup4,
!pip install openpyxl,
!pip install nltk,
!pip install textblob
!pip install python-docx
in your Google Colab environment to install the required dependencies.

### ***Explanation***

- **requests:** Used for making HTTP requests to fetch web page content.
- beautifulsoup4: Utilized for parsing HTML content retrieved from web pages.
- **openpyxl:** Employed for reading and writing Excel files for input and output data handling.
- **nltk**: Provides various natural language processing tools and resources.
- **textblob:** Offers tools for sentiment analysis and text processing.
- **python-docx:** Used for working with Microsoft Word files, specifically for extracting variables from a Docx document.