**BLACKCOFFER - DATA EXTRACTION AND NLP ANALYSIS**

**Objective:**

The objective of the project is to extract textual data articles from the provided URL and conduct text analysis to compute various variables.

**Approach:**

1. **Understanding problem statement:**

* Begin by thoroughly understanding the requirements outlined in the problem statement.

1. **Data extraction:**

* Utilized *BeautifulSoup* for web scraping to extract article titles and text from the provided URL.
* Excluded unnecessary contents such as website headers and footers.

1. **Data processing:**

* Save the extracted article content into a text file, naming it with the URL’s ID in the Extracted\_Articles folder.
* Incorporate provided stopwords from the stopwords folder into a set.
* Separate positive and negative words from the MasterDictionary folder into distinct sets.

1. **Text analysis:**

Referencing the methods and approach detailed in the Text Analysis.docx, proceeded with the following analyses:

* **Sentiment Analysis:** Calculate positive score, negative score, polarity score, subjectivity score.
* **Readability Analysis:** Determine average sentence length, Percentage of complex words, Fog Index.
* **Additional metrics:** Calculate average number of words per sentence, complex word count, word count, syllable count per word, personal pronouns, average word length.

1. **Implementation:**

* Utilized *nltk* package for tokenization, stop word removal, punctuation removal before computing scores.
* Organize the calculated metrics into a dataframe structured as described in Output Data Structure.xlsx

1. **Output Generation:**

* Load the calculated scores into the dataframe.
* Save the dataframe into the solution folder.

**Instructions:**

* Download the shared folder “BLACKCOFFER SOLUTION” folder in your local machine.
* There is a dataExtractionAndAnalysis.py script file.
* Run the script, output Excel file is generated.
* It will take some time, have a good network connection.
* Ensure you have the necessary dependencies installed or else install using

***pip install beautifulsoup4 pandas nltk openpyxl***

**Dependencies:**

* *beautifulsoup4 -* web scraping
* *pandas -* data manipulation
* *nltk -* natural language processing tasks
* *openpyxl -* for read and write Excel files