# Blackcoffer NLP Assignment - Text Extraction & Analysis

## Overview

This project fulfills the test assignment provided by **Blackcoffer** to automate:

1. Extraction of textual content from a given set of article URLs.
2. Textual analysis using NLP to compute a defined set of linguistic metrics.

All extracted data and computed values are exported to a structured output file in .csv format.

## Objectives

* Scrape the article title and body text from each URL listed in Input.xlsx.
* Save each article as a separate .txt file using its URL\_ID.
* Analyze each article's content using NLP and calculate the following metrics:
  + POSITIVE SCORE
  + NEGATIVE SCORE
  + POLARITY SCORE
  + SUBJECTIVITY SCORE
  + AVERAGE SENTENCE LENGTH
  + PERCENTAGE OF COMPLEX WORDS
  + FOG INDEX
  + AVERAGE NUMBER OF WORDS PER SENTENCE
  + COMPLEX WORD COUNT
  + WORD COUNT
  + SYLLABLES PER WORD
  + PERSONAL PRONOUNS
  + AVERAGE WORD LENGTH
* Save all computed metrics in the exact order and format of Output Data Structure.xlsx.

## Folder Structure

Blackcoffer\_NLP\_Project/

├── Input.xlsx

├── Output.csv

├── main.py

├── text\_files/ # Extracted articles (.txt)

├── stopwords/ # Stopword and lexicon lists

│ ├── positive-words.txt

│ ├── negative-words.txt

│ ├── stopwords\_auditor.txt

│ ├── stopwords\_currency.txt

│ ├── stopwords\_datesandnumbers.txt

│ ├── stopwords\_generic.txt

│ ├── stopwords\_names.txt

│ ├── stopwords\_places.txt

│ └── stopwords\_connectors.txt

└── README.md # Instructions and explanation

## Setup Instructions

### 1. Clone or Download

Ensure all required files and folders (Input.xlsx, stopwords/, main.py, etc.) are in the same directory.

### 2. Install Required Libraries

Run the following command:

pip install -r requirements.txt

**Dependencies:**

* pandas
* nltk
* textblob
* beautifulsoup4
* requests
* syllapy
* openpyxl

### 3. Run the Project

python main.py

This script will:

* Create a text\_files/ folder and save .txt articles
* Compute linguistic features
* Generate Output.csv as the final deliverable

## Notes

* Internet connection is required for web scraping.
* Some URLs may return empty due to invalid links or inaccessible pages.
* The script handles exceptions and continues processing all available URLs.
* All stopword and sentiment lexicons must be present inside the stopwords/ folder.

## Submission

Upload the folder to Google Drive and share the folder link via the form: [Google Form Link](https://forms.gle/nvWAgrCBdq1JkKou8)

Make sure your submission includes:

* main.py
* Output.csv
* README.md
* stopwords/ folder

## Author

Hariharasudhan — built and structured for exact match to the Blackcoffer assignment.