**Data Extraction and NLP**

Instructions:

* Approach to the Solution:

1. Importing Libraries:

- Ensure that you have installed Python on your system.

- Import the necessary Python libraries as follows:

import numpy as np

import pandas as pd

import requests

import seaborn as sns

from bs4 import BeautifulSoup

from nltk import FreqDist

from nltk.corpus import stopwords

from nltk.tokenize import word\_tokenize

import re

import warnings

warnings.filterwarnings(action='ignore')

2. Reading Input Data:

- Use the input() function to enter the file paths for Input Data, Stopword Data, Positive and Negative Word Data, and Output Data.

**Please ensure that each file name matches the names of the files exactly.**

3. Fetching Text Content from URLs:

- Use the `fetch\_content()` function to gather text content from the URLs specified in the Excel file.

4. Preprocessing Stopwords:

- Extract and preprocess stopwords from various text files, and compile them into a list.

5. Tokenization and Text Cleaning:

- Tokenize and clean text content by removing stopwords and special characters.

6. Analysis of Readability:

- Compute metrics such as average sentence length, fog index, and percentage of complex words.

**Note** - the fog index percentage was not correct, so I multiplied it by one hundred to convert it to the correct form.

7. Additional Analysis:

- Calculate metrics such as the average number of words per sentence, complex word count, and syllable count per word.

8. Output Generation:

- Combine the processed data with the original input, and save the output to a CSV file at desired file path.

* How to run the .py file to generate output

Please follow the steps below to ensure successful execution of the Python script:

1. Install Dependencies:

- Confirm that the following dependencies are installed by using the pip command in the command prompt before running the Python script.

- Dependencies to be installed: numpy, pandas, requests, seaborn, beautifulsoup4, and nltk.

2. Run the Python Script:

- Open a command prompt or terminal.

- Navigate to the directory where the script is saved by using the **"cd"** command followed by the path to the folder where the script is located.

- Execute the script by using the command " **NLP.py**".

4. Enter the Required Information:

- When prompted by the command prompt**, enter the path** of the input file, stopwords folder path, positive words, negative words, and output data structure file in the same order.

**- For the stopwords folder path, you only need to enter the path to the folder where all the stopword files are located.**

5. Check the Output:

- After the script has finished executing, verify the output in the location you specified for the output data structure file.

* Dependencies Required:

Please ensure that you have the following dependencies installed before running the script:

**- Python 3.x**

**- numpy**

**- pandas**

**- requests**

**- seaborn**

**- beautifulsoup4**

**- nltk**

To install the necessary dependencies, please use the following command in the command prompt:

**pip install numpy pandas requests seaborn beautifulsoup4 nltk**

**Please note that it's important to have all the dependencies installed to ensure that the script runs smoothly.**

**\*\*The Jupyter Notebook is also given with the py file\*\***