Data Extraction and NLP

Test Assignment

Objective:

The objective of this assignment is to extract textual data articles from the given URL and perform text analysis to compute variables.

Explain how you approached the solution?

To solve the problem of extracting, cleaning, and analyzing text from URLs, and updating an Excel file with required variables, we followed a structured approach that addressed various requirements step by step. Here's a detailed breakdown of how we approached the solution:

1. Understanding Requirements

Input: URLs from an Excel file.

Processing: Extract and clean text, perform sentiment analysis, and calculate various text metrics.

Output: Save the results into an output Excel file.

2. Initial Setup

Libraries: We used pandas for handling Excel files, newspaper3k for extracting text from URLs, re for regex operations, and string for punctuation handling.

Word Lists: We needed files for stopwords, positive words, and negative words.

3. Defining Functions for Each Task

Loading Word Lists: A function to load stopwords, positive, and negative words from text files into sets.

Extracting Content: Using newspaper3k to extract the main content and title from URLs.

Cleaning Text: Removing punctuation and stopwords, and counting total words after cleaning.

Sentiment Analysis: Counting occurrences of positive and negative words in the cleaned text.

Text Metrics Calculation: Functions to calculate various metrics like polarity score, subjectivity score, average sentence length, percentage of complex words, fog index, etc.

Handling Complex Words: Counting words with more than two syllables.

Syllable Counting: A helper function to count syllables in a word.

4. File Handling and Updating Mechanism

Initialize Excel File: Ensure the output Excel file has the necessary columns.

Updating Excel File: A function to update the Excel file with new data continuously.

5. Processing URLs and Integrating All Steps

Reading Input File: Load URLs from the input Excel file.

Processing Loop: Iterate over each URL, perform extraction, cleaning, sentiment analysis, and metrics calculation.

Updating Output File: Update the output Excel file with the results for each URL.

How to run the .py file to generate output

1. Kindly save the folder uploaded on the google drive in the local system.
2. Extract the files. The folder contains '.py' file, final code Jupyter notebook, code break up Jupyter notebook, txt files for stop words, positive words and negative words, and input and output excel files. These files could have been stored in different folders but then the paths might need to update. That's the reason all files are in same folder for easy access and no changes are required in the code.
3. The Jupyter files are executed once for the first four urls and results are displayed to give the idea about the output.
4. Python dependencies are needed to be installed before running the code else it might throw an error.
5. The libraries that need to be install included Newspaper3k, OpenPyXL, pandas, collections, re, string, and some libraries as per the system requirements.
6. Open Windows Power shell or linux terminal and change the directory path to the folder where the '.py' files is saved.
7. Run the python file titled 'NetClan_code.py' from the windows power shell or linux terminal. For Windows type 'python NetClan_code.py'.
8. The python file will iterate the urls one by one and each result is displayed in the power shell window. The output excel file will be generated after the completion of the code. And will be saved in the same folder.