**Data Extraction and NLP**

**Test Assignment**

**How I approached the solution**

**Step 1:Data Extraction**

1-Read URLs from ‘input.xlsx’.

2-For each URL ,fetch the web content using ‘requests’ and parse it with ‘BeautifulSoup’

3-Extarct the title and article text ,and save them in individual text files.

**Step2:Text Analysis**

1-Load stop words from NLTK and custom stop words from the ‘StopWords’ directory.

2-Clean the text by removing stop words and non-alphanumeric characters.

3-Tokenize the text into words and sentences.

4-Perform various text analyses including sentiment analyses ,readability metrics ,word and sentence counts, syllable counts and more.

5-Use TextBlob for polarity and subjectivity scores and custom dictionaries for positive and negative scores.

**Step3:save Output**

1-Collect all analysis results in a list and save them to output.xlsx’ as according to required format

**How to run the .py file to generate output—**

**ensure you have python and required library are installed**

1-Make sure you have python installed in your system.

2-install all required libraries using pip package installer



requests- to fetch webpage content.

beautifulsoup4- for parsing HTML content.

pandas- for handling data frames and reading/writing excel files.

nltk-for tokenizing text and loading stop words.

Textblob-for sentiment analysis.

Pyphen-for counting syllables in words.

Openpyxl-for working with excel files

**Ensure Directory Structure**

1-make a file script.py in main directory.

2-Ensure StopWords , MasterDictionary and input.xlsx present in same directory.

**Run the script**

1-Open a terminal .

2-navigate to directory containing script.py.

3-Run the command python script.py

4-this script will generate output.xlsx in the same directory.