**Blackcoffer**

**Data Extraction and NLP**

**Test Assignment Answer**

**Objective:**

To extract textual data articles from the given URL and perform text analysis to compute given variables.

**Steps :**

1. Divided problem into sub problems viz., extracting link part from Input file, revocation of the only content of an article, analysis of the exctracted text, and storing back all of the said records in a file.
2. Searched for different tools, packages, functions needed.
3. Created an algorithms to solve each sub problem individually.
4. Executed them independently.
5. Integrated all the modules in a program and received a desired output.

**Methodology:**

1. **Web Scraping**

**Tool:** Use a web scraping tool or library like BeautifulSoup (Python) to extract HTML content and prettify or interact with the content from the given URL.

**Considerations:** Check the website's terms of service to ensure compliance with web scraping policies.

**2. Data Cleaning and Preprocessing**:

* 1. HTML to Text Conversion: Strip HTML tags to retain only the text content.
  2. Remove Noise: Eliminate irrelevant characters, symbols, or numbers that may interfere with analysis.
  3. Tokenization: Break text into individual words (tokens).
  4. Lowercasing: Convert all text to lowercase for consistency.

**3.Text Analysis:**

* 1. Text Summarization: Generate a concise summary of the article to capture its main points.
  2. Sentiment Analysis: Determine the sentiment of the text (positive, negative, neutral).
  3. Keyword Extraction: Identify and extract key terms or phrases.
  4. Named Entity Recognition (NER): Identify and classify named entities such as persons, organizations, and locations.

**4. Variable Computation:**

* 1. Word Count: Compute the total number of words in the article.
  2. Sentence Count: Determine the total number of sentences.
  3. Average Word Length: Calculate the average length of words in the text.
  4. Most Frequent Words: Identify the most frequently occurring words.

**Steps of Data Analysis**

Step 1 – Get the Data into a Spreadsheet. ...

Step 2 – Scrub the Responses. ...

Step 3 – Assign Descriptors. ...

Step 5 – Repeat Steps 3 and 4. ...

Step 6 – Analyze.

**How to run the given code:**

Would need a python editor as well as some packages to be installed pre running the program. A command to install packages looks like:

>>pip install package\_name

**Required Packages:**

1. re

2. pandas

3. requests

4. BeautifulSoup word\_tokenize, sent\_tokenize

5. openpyxl

6. xlsxwriter

7. Nltk:

i. cmudict from nltk.corpus

ii. NLTKWordTokenizer from nltk

iii. stopwords from nltk.corpus

iv. word\_tokenize from nltk.tokenize

v. word\_tokenize, sent\_tokenize from nltk.tokenize

**Pre requisite to run the program:**

This program will require two files to be in **same folder** where the program is stored at or else would need to change path of the required two files namely(with same alias as given).

1. Input.xlsx
2. Positive and Negative Words List.xlsx
3. Output file.xlsx(only with **desired column names** added to it given in Output Stucture File.xlsx)

**Output:**

After successfully running the program will create an output file in a given format in the same folder as program named as “Output file.xlsx” or “Output file.csv”. In which a whole text analysis of every article can be seen from “Output file.xlsx”. And concurrent code can be seen in “Assign\_text\_analysis.py”.

“Output With Text Analysis.xlsx” file **after running** this program is attached with google link can have look upon.

Waiting for your positive response.

**Thanking you.**