Contents

1 How I Approached to solution

1.1 Part 1 : Web Scraping

Procedure:

1. Firstly, I looked at the website and inspected it to **locate the title and text area**.
2. After identifying the desired areas, I utilized the **request** module in Python to fetch the website, opened an input file using the Python **open** module, and employed **BeautifulSoup** for parsing.

Note: During parsing, I encountered two errors:

1. The HTML code for the text area and title **differed** on some websites, prompting me to create separate if-else statements to rectify this issue.
2. Certain websites displayed **"Page not Found"** errors, requiring me to implement additional if-else conditions to address this issue.

1.2 Part 2 : Data Analysis

Procedure:

1. Firstly, I imported **positive words, negative words, and stopwords** from the shared drive link.
2. I created a **function** to remove stopwords from sentences using both the drive and stopwords class of the NLTK package, as well as punctuation.
3. After that, I **created all the required functions** for text analysis:
    - 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
4. Then, I **looped** through every stored text data obtained from web scraping. For each text, I computed the values using the functions created in step 3, and stored these values in a **DataFrame**. Finally, I exported the DataFrame as an Excel file.

2 How to run the .py file

2.1 Part 1 : Web Scraping

Procedure:

1. Install the required libraries listed in the **'requirement.txt'** file.
2. Download the necessary folders shared in the drive.
3. Initially, clear all files from the `Solution\Text_File` folder, as it will be used to store the scraped files.
4. Execute the '**Web_Scraping.ipynb**' notebook. This will generate text files in the '**Solution** ' directory. Simply copy and paste these files into the '`Solution\Text_File`' folder.

2.2 Part 1 : Data Analysis

—

Procedure:
Note: Ensure that the folder structure remains unchanged throughout the process.

1. Run all the required functions for data analysis.
2. Upon completion, the resulting Excel file will be stored in the parent folder as '`Output_File.xlsx`'.