How to run assignmentnotebook.py file

Step - 1 :- Download all the files which are present in the assignment file.

Step - 2 :- Setup chrome driver and selenium for data Scraping from every website

Step - 3 :- Download assignmentnotebook.ipynb from assignment solutions file

Step - 4 :- Upload that file in the jupyter notebook and open it.

Step - 5 :- Upload all necessary files on jupyter notebook like stopwords files

and masterdict files carefully.

Step- 6 :- Update every path with your actual dir path.

Step- 7 :- Run every cell of the notebook one by one carefully.

About Assignment flow :

|--- Setup chrome driver and selenium for data Scraping of data from websites.

|--- Scraping all the data from the website and saving it to the OS as a txt file with url\_id name

|--- Collecting all the data as per mentioned in the objective.

|--- Data Preprocessing

| |--- Remove characters other than letters

| |--- Convert the sentence to lowercase

| |--- Tokenize the sentence into words

| |--- Remove stopwords

| |--- Join the words back into a sentence

| |--- Add preprocessed data as a new column in df

|--- Text Analysis & Sentiment Analysis

|---- POSITIVE SCORE

|---- NEGATIVE SCORE

|---- POLARITY SCORE

|---- SUBJECTIVITY SCORE

|---- AVG SENTENCE LENGTH

|---- PERCENTAGE OF COMPLEX WORDS

|---- FOG INDEX

|---- AVG NUMBER OF WORDS PER SENTENCE

|---- COMPLEX WORD COUNT

|---- WORD COUNT

|---- SYLLABLE PER WORD

|---- PERSONAL PRONOUNS

|---- AVG WORD LENGTH

Text preprocessing approach

* File Reading:
  + For each row in the data table, we identify the corresponding text file.
  + We read the content of that file.
* Sentence Tokenization:
  + The content of the file (paragraphs) is broken down into individual sentences.
  + This is important for understanding the structure of the text.
* Sentence Preprocessing:
  + Each sentence undergoes a cleaning process.
  + We remove anything that isn't a letter (like punctuation or numbers).
  + We convert all letters to lowercase for consistency.
* Stopword Removal:
  + Common words that don't carry much meaning (stopwords) are removed.
  + This includes both general English stopwords and any additional custom stopwords.
* Adding Preprocessed Text to Data Table:
  + The cleaned sentences are put back together and stored in a new column in the data table.
  + This new column is named 'preprocessed\_data'.
* Final Output:
  + The function returns the updated data table with the cleaned text.

Note :- For Text Analysis I used the same approach which is mentioned in the text analysis doc.

Requirement.txt

* Pandas
* Numpy
* Selenium
* Chrome driver
* Os
* Nltk
* String
* Re