Textual Analysis (Useful for NLP tasks)

-Sarthak Srivastava
(Email: sarthak.s.1603@gmail.com)

Objective of the project:

1. Scrape a website/websites for text
2. The Web scrapping section in the code can be edited as per requirement of the project. I have chosen websites to showcase the working of the code.
3. Upon changing the web scrapping part of the code the ultimate goal is to extract the main text from the website on which the analysis has to be done and save it as a string.
4. Once the code is saved, we can proceed with the Analysis.

The definitions of the various parameters are as under:

1. Extracting Derived variables

   We convert the text into a list of tokens using the nltk tokenize module and use these tokens to calculate the 4 variables described below:

   Positive Score: This score is calculated by assigning the value of +1 for each word if found in the Positive Dictionary and then adding up all the values.

   Negative Score: This score is calculated by assigning the value of -1 for each word if found in the Negative Dictionary and then adding up all the values. We multiply the score with -1 so that the score is a positive number.

   Polarity Score: This is the score that determines if a given text is positive or negative in nature. It is calculated by using the formula:
   Polarity Score = (Positive Score – Negative Score)/ ((Positive Score + Negative Score) + 0.000001)
   Range is from -1 to +1

   Subjectivity Score: This is the score that determines if a given text is objective or subjective. It is calculated by using the formula:
   Subjectivity Score = (Positive Score + Negative Score)/ ((Total Words after cleaning) + 0.000001)
   Range is from 0 to +1

2. Analysis of Readability

   Analysis of Readability is calculated using the Gunning Fox index formula described below.

Average Sentence Length = the number of words / the number of sentences

Percentage of Complex words = the number of complex words / the number of words

Fog Index = 0.4 * (Average Sentence Length + Percentage of Complex words) (aka Gunning Fox Index)

3. Average Number of Words Per Sentence
   The formula for calculating is:

   Average Number of Words Per Sentence = the total number of words / the total number of sentences

4. **Complex Word Count**
   Complex words are words in the text that contain more than two syllables.

5. **Word Count**
   We count the total cleaned words present in the text by
   removing the stop words (using stopwords class of nltk package).
   removing any punctuations like ? ! , . from the word before counting.

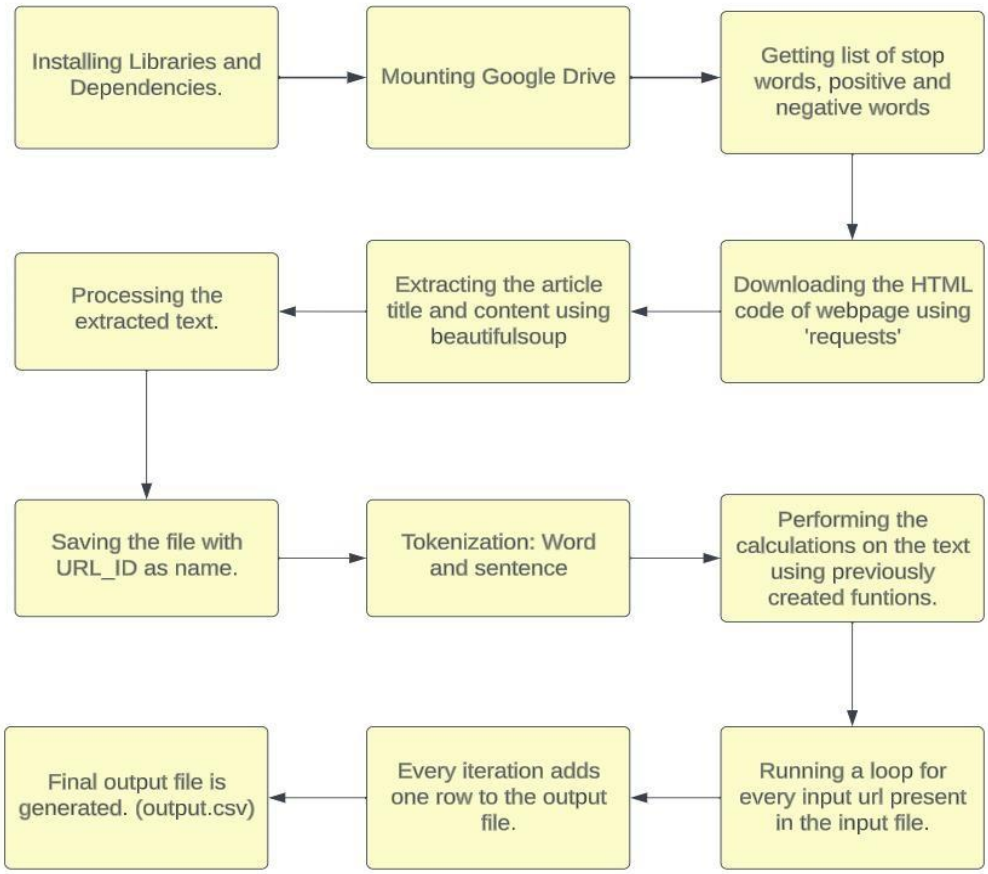
6. **Syllable Count Per Word**
   We count the number of Syllables in each word of the text by counting the vowels present in each word. We also handle some exceptions like words ending with "es","ed" by not counting them as a syllable.

7. **Personal Pronouns**
   To calculate Personal Pronouns mentioned in the text, we use regex to find the counts of the words - “I,” “we,” “my,” “ours,” and “us”. Special care is taken so that the country name US is not included in the list.

8. **Average Word Length**
   Average Word Length is calculated by the formula:
   Sum of the total number of characters in each word/Total number of words

Code flowchart:

(As per my code your implementation could be different on basis of how you import your target websites and what all parameters you wish to calculate as well as the format of the output file.)



A zip file containing list of Stop words, Positive words, Negative words is uploaded in the GitHub repository.