##WEB\_SCRAPPING##

1. Imported beautifulsoup and loaded the input.xlsx file to a dataframe. ##Installed requests, BeautifulSoup4 and lxml libraries.
2. Looped through all the URLs and web scrapped the article title and article text.
3. Saved all the article text files with name as URL\_ID.txt in a separate folder.

##ANALYSING THE TEXT DATA##

## ## Readability Analysis

1. Imported the web-scrapped text files from the local disk into a list.
2. Removed all the extra white-space from the articles.
3. Removed stopwords using nltk english stopwords.
4. Defined a sentence counter function followed by storing the sentence count of all the cleaned articles in a list. Note: Almost same number of sentence count for both raw and cleaned corpus.
5. Defined a word counter function followed by storing the word count of all the cleaned articles in a list. Note: To count the words, the articles were further cleaned by using the remove\_punctuation function.
6. Defined a character counter function to count the total characters in an article.
7. Defined syllable counter using textstatistic library, other manually rule based codes do not count a lot of corner cases.
8. Defined complex word counter using the definition of complex word and syllable counter.
9. Gunning Fog index was calculated using its formula.
10. Personal Pronoun counter function is used to count the personal pronouns in each article. NOTE: raw corpus without excluding stopwords have been used, only removed punctuations. Used regex to find the words “I,” “we,” “my,” “ours,” and “us” in the article. Special care is taken so that the country name US is not included in the list.
11. Saved them to the output data structure

## ##Sentiment Analysis

1. Imported positive and negative word dictionaries from nltk corpus, opinion\_lexicon module.
2. Defined a sentence's positive and negative scorer function. Loop through all the words of the sentence and return the positive and negative score.
3. Defined an article positive and negative scorer function.Loop through all the articles, return the overall positive and negative score.
4. Imported textblob to extract the polarity and subjectivity of articles. The Manual subjectivity score was not giving good results.
5. Saved them to the output data structure

Finally saved the output data structure to an excel file.