

Website Sentiment Analysis

We are going to make a web scraper in this assignment that visits each and every link in the list and gets us the required metrics so that we can know more about the article. We will look for metrics like polarity score, positive-negative score, word count etc and then analyse the article.

```
In [68]: #importing necessary packages
import requests
from bs4 import BeautifulSoup
import spacy
from textblob import TextBlob
from nltk.tokenize import sent_tokenize, word_tokenize
from nltk.corpus import stopwords
import nltk
from nltk.corpus import stopwords
import re
import pandas as pd
```

We are going to make a web scraper in this assignment that visits each and every link in the list and gets us the required metrics so that we can know more about the article. We will look for metrics like polarity score, positive-negative score, word count etc and then analyse the article.

```
#importing necessary packages
```

```
import requests
from bs4 import BeautifulSoup
import spacy
from textblob import TextBlob
from nltk.tokenize import sent_tokenize, word_tokenize
from nltk.corpus import stopwords
import nltk
from nltk.corpus import stopwords
import re
import pandas as pd
```

We are going to make a web scraper in this assignment that visits each and every link in the list and gets us the required metrics so that we can know more about the article. We will look for metrics like polarity score, positive-negative score, word count etc and then analyse the article.

```
#importing necessary packages
```

```
import requests
from bs4 import BeautifulSoup
import spacy
from textblob import TextBlob
from nltk.tokenize import sent_tokenize, word_tokenize
from nltk.corpus import stopwords
import nltk
from nltk.corpus import stopwords
import re
import pandas as pd
```


[illegible]


```
In [88]: #scraping the dataset
if __name__ == "__main__":

    dataset_file = 'C:\\Users\\sujoydutta\\Desktop\\Data analysis\\Projects\\Text scoring Blackcoffer\\Output 0

    df = pd.read_excel(dataset_file)

    url_column_name = 'URL'

    scraped_text_from_column(df, url_column_name)

    data = {
        'URL': df[url_column_name],
        'Scraped_Text': df['Scraped_Text']}
    }

Failed to retrieve content from https://insights.blackcoffer.com/how-neural-networks-can-be-applied-in-various-areas-in-the-future/. Status code: 404
Failed to retrieve content from https://insights.blackcoffer.com/covid-19-environmental-impact-for-the-future/. Status code: 404

In [89]: #examining the new dataset
new_df = pd.DataFrame(data)
new_df.head()
```

	URL	Scraped_Text
0	https://insights.blackcoffer.com/rise-of-tele... Automate the Data Management Process Realtime ...	
1	https://insights.blackcoffer.com/rise-of-e-hea... Automate the Data Management Process Realtime ...	
2	https://insights.blackcoffer.com/rise-of-e-hea... Automate the Data Management Process Realtime ...	
3	https://insights.blackcoffer.com/rise-of-tele... Automate the Data Management Process Realtime ...	
4	https://insights.blackcoffer.com/rise-of-tele... Automate the Data Management Process Realtime ...	

```
In [90]: #viewing the datatypes
new_df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 114 entries, 0 to 113
Data columns (total 2 columns):
 #   Column      Non-Null Count  Dtype
---  --
 0   URL         114 non-null    object
 1   Scraped_Text 112 non-null    object
dtypes: object(2)
memory usage: 1.9+ KB
```

```
In [91]: #dropping null values
new_df = new_df.dropna()
new_df
```

	URL	Scraped_Text
0	https://insights.blackcoffer.com/rise-of-tele... Automate the Data Management Process Realtime ...	
1	https://insights.blackcoffer.com/rise-of-e-hea... Automate the Data Management Process Realtime ...	
2	https://insights.blackcoffer.com/rise-of-e-hea... Automate the Data Management Process Realtime ...	
3	https://insights.blackcoffer.com/rise-of-tele... Automate the Data Management Process Realtime ...	
4	https://insights.blackcoffer.com/rise-of-tele... Automate the Data Management Process Realtime ...	
...
109	https://insights.blackcoffer.com/coronavirus-l... Automate the Data Management Process Realtime ...	
110	https://insights.blackcoffer.com/coronavirus-l... Automate the Data Management Process Realtime ...	
111	https://insights.blackcoffer.com/what-are-the-... Automate the Data Management Process Realtime ...	
112	https://insights.blackcoffer.com/marketing-dif... Automate the Data Management Process Realtime ...	
113	https://insights.blackcoffer.com/continued-dem... Automate the Data Management Process Realtime ...	

112 rows x 2 columns

```
In [92]: # Function to combine all variables into a single list
def calculate_all_variables(text, positive_words, negative_words):
    cleaned_text = clean_text(text)
    positive_dict, negative_dict = create_sentiment_dict(cleaned_text, positive_words, negative_words)
    positive_score = calculate_positive_score(positive_dict)
    negative_score = calculate_negative_score(negative_dict)
    polarity_score = calculate_polarity_score(positive_score, negative_score)
    total_words = len(cleaned_text.split())
    subjectivity_score = calculate_subjectivity_score(positive_score, negative_score, total_words)
    avg_sentence_length = calculate_avg_sentence_length(text)
    percentage_complex_words = calculate_percentage_complex_words(text)
    fog_index = calculate_fog_index(avg_sentence_length, percentage_complex_words)
    avg_words_per_sentence = calculate_avg_words_per_sentence(text)
    complex_word_count = count_complex_words(text)
    total_word_count = count_total_words(text)
    syllables_per_word = calculate_syllables_per_word(text)
    personal_pronoun_count = count_personal_pronouns(text)
    avg_word_length = calculate_avg_word_length(text)

    return positive_score, negative_score, polarity_score, subjectivity_score, avg_sentence_length, \
           percentage_complex_words, fog_index, avg_words_per_sentence, complex_word_count, \
           total_word_count, syllables_per_word, personal_pronoun_count, avg_word_length
```

```
In [94]: #applying the function on the new dataset
new_df[['Positive_Score', 'Negative_Score', 'Polarity_Score', 'Subjectivity_Score',
        'Avg_Sentence_Length', 'Percentage_Complex_Words', 'Fog_Index',
        'Avg_Words_Per_Sentence', 'Complex_Word_Count', 'Total_Word_Count',
        'Syllables_Per_Word', 'Personal_Pronoun_Count', 'Avg_Word_Length']] = new_df['Scraped_Text'].apply(
    lambda text: pd.Series(calculate_all_variables(text, positive_words, negative_words)))
```

```
In [95]: #examining the new dataset
new_df = new_df.drop(['Scraped_Text'], axis=1)
new_df.head()
```

	URL	Positive_Score	Negative_Score	Polarity_Score	Subjectivity_Score	Avg_Sentence_Length	Percentage_Complex_Words	Fog_Index	Avg_Words_Per_Sentence	Complex_Word_Count	Total_Word_Count	Syllables_Per_Word	Personal_Pronoun_Count	Avg_Word_Length
0	https://insights.blackcoffer.com/rise-of-tele... Automate the Data Management Process Realtime ...	368	-93	1.676364	0.202653	24.773810								
1	https://insights.blackcoffer.com/rise-of-e-hea... Automate the Data Management Process Realtime ...	162	-39	1.634146	0.218472	30.758621								
2	https://insights.blackcoffer.com/rise-of-e-hea... Automate the Data Management Process Realtime ...	177	-68	2.247706	0.115957	19.767123								
3	https://insights.blackcoffer.com/rise-of-tele... Automate the Data Management Process Realtime ...	264	-69	1.707692	0.182927	24.968750								
4	https://insights.blackcoffer.com/rise-of-tele... Automate the Data Management Process Realtime ...	264	-69	1.707692	0.182927	24.968750								

```
In [106]: # Taking only the necessary columns
subset_df = df.iloc[:, :12]
subset_df
```

	URL_ID	URL
0	123.0	https://insights.blackcoffer.com/rise-of-tele...
1	321.0	https://insights.blackcoffer.com/rise-of-e-hea...
2	2345.0	https://insights.blackcoffer.com/rise-of-e-hea...
3	4321.0	https://insights.blackcoffer.com/rise-of-tele...
4	432.0	https://insights.blackcoffer.com/rise-of-tele...
...
109	50921.0	https://insights.blackcoffer.com/coronavirus-l...
110	51382.8	https://insights.blackcoffer.com/coronavirus-l...
111	51844.6	https://insights.blackcoffer.com/what-are-the-...
112	52306.4	https://insights.blackcoffer.com/marketing-dif...
113	52768.2	https://insights.blackcoffer.com/continued-dem...

114 rows x 2 columns

```
In [107]: # Merging the DataFrames based on the 'URL'
merged_df = pd.merge(subset_df, new_df, on='URL', how='inner')
merged_df.head()
```

	URL_ID	URL	Positive_Score	Negative_Score	Polarity_Score	Subjectivity_Score	Avg_Sentence_Length	Percentage_Complex_Words	Fog_Index	Avg_Words_Per_Sentence	Complex_Word_Count	Total_Word_Count	Syllables_Per_Word	Personal_Pronoun_Count	Avg_Word_Length
0	123.0	https://insights.blackcoffer.com/rise-of-tele...	368	-93	1.676364	0.202653	24.773810								
1	321.0	https://insights.blackcoffer.com/rise-of-e-hea...	162	-39	1.634146	0.218472	30.758621								
2	2345.0	https://insights.blackcoffer.com/rise-of-e-hea...	177	-68	2.247706	0.115957	19.767123								
3	4321.0	https://insights.blackcoffer.com/rise-of-tele...	264	-69	1.707692	0.182927	24.968750								
4	432.0	https://insights.blackcoffer.com/rise-of-tele...	264	-69	1.707692	0.182927	24.968750								

```
In [109]: #dropping na values
merged_df.dropna()
```

	URL_ID	URL	Positive_Score	Negative_Score	Polarity_Score	Subjectivity_Score	Avg_Sentence_Length	Percentage_Complex_Words	Fog_Index	Avg_Words_Per_Sentence	Complex_Word_Count	Total_Word_Count	Syllables_Per_Word	Personal_Pronoun_Count	Avg_Word_Length
0	123.0	https://insights.blackcoffer.com/rise-of-tele...	368	-93	1.676364	0.202653	24.773810								
1	321.0	https://insights.blackcoffer.com/rise-of-e-hea...	162	-39	1.634146	0.218472	30.758621								
2	2345.0	https://insights.blackcoffer.com/rise-of-e-hea...	177	-68	2.247706	0.115957	19.767123								
3	4321.0	https://insights.blackcoffer.com/rise-of-tele...	264	-69	1.707692	0.182927	24.968750								
4	432.0	https://insights.blackcoffer.com/rise-of-tele...	264	-69	1.707692	0.182927	24.968750								
...
107	50921.0	https://insights.blackcoffer.com/coronavirus-l...	92	-28	1.875000	0.120983	29.653846								
108	51382.8	https://insights.blackcoffer.com/coronavirus-l...	254	-154	4.080000	0.068634	39.166667								
109	51844.6	https://insights.blackcoffer.com/what-are-the-...	385	-87	1.583893	0.215786	28.770270								
110	52306.4	https://insights.blackcoffer.com/marketing-dif...	275	-88	1.941176	0.161765	28.809524								
111	52768.2	https://insights.blackcoffer.com/continued-dem...	132	-61	2.718310	0.104412	31.935484								

112 rows x 15 columns

```
In [110]: # Converting the DataFrame to csv
merged_df.to_csv('output_ishan.csv', index=False)
```