Blackcoffer Consulting
Data Extraction and Text Analysis

1. Import the required python libraries and packages –

```
import pandas as pd
import requests
import bs4 as bfs
import nltk
from nltk.tokenize import sent_tokenize
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
from nltk.sentiment.vader import SentimentIntensityAnalyzer
import string
from textblob import TextBlob
import csv
import numpy as np
```

```
!pip install openpyxl
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/public/simple/
Requirement already satisfied: openpyxl in /usr/local/lib/python3.7/dist-packages (3.0.10)
Requirement already satisfied: et-xmlfile in /usr/local/lib/python3.7/dist-packages (from openpyxl) (1.1.0)
```

2. To remove all the words that does no add much meaning to the sentence use nltk library (Natural Language Toolkit) –

```
nltk.download('stopwords')
nltk.download('punkt')
```

```
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Unzipping corpora/stopwords.zip.
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Unzipping tokenizers/punkt.zip.
True
```

3. Import the required files into python environment for further processing. In Google Colab the below given command is used to import files –

```
import io
import pandas as pd
from google.colab import files
```

```
uploaded = files.upload()
```

```
Choose Files  Input.xlsx
• Input.xlsx(application/vnd.openxmlformats-officedocument.spreadsheetml.sheet) - 20591 bytes, last modified: 6/14/2022 - 100% done
Saving Input.xlsx to Input (1).xlsx
```

4. Convert the imported files into a DataFrame for easy accessing –

```
df = pd.read_excel(io.BytesIO(uploaded.get('Input.xlsx')))
```

```
df
```

	URL_ID	URL
0	1.0	https://insights.blackcoffer.com/how-is-login-...
1	2.0	https://insights.blackcoffer.com/how-does-ai-h...
2	3.0	https://insights.blackcoffer.com/ai-and-its-im...
3	4.0	https://insights.blackcoffer.com/how-do-deep-l...
4	5.0	https://insights.blackcoffer.com/how-artificia...
...	...	...
165	167.0	https://insights.blackcoffer.com/role-big-data...
166	168.0	https://insights.blackcoffer.com/sales-forecas...
167	169.0	https://insights.blackcoffer.com/detect-data-e...
168	170.0	https://insights.blackcoffer.com/data-exfiltra...
169	171.0	https://insights.blackcoffer.com/impacts-of-co...

170 rows × 2 columns

5. Extract all the URLs from the Excel file –

```
li = [url for url in df['URL']]
li
```

```
'https://insights.blackcoffer.com/continued-demand-for-sustainability/',
'https://insights.blackcoffer.com/coronavirus-disease-covid-19-effect-the-impact-and-role-of-mass-media-during-the
'https://insights.blackcoffer.com/should-people-wear-fabric-gloves-seeking-evidence-regarding-the-differential-tra
'https://insights.blackcoffer.com/why-is-there-a-severe-immunological-and-inflammatory-explosion-in-those-affected
'https://insights.blackcoffer.com/what-do-you-think-is-the-lesson-or-lessons-to-be-learned-with-covid-19/',
'https://insights.blackcoffer.com/coronavirus-the-unexpected-challenge-for-the-european-union/',
'https://insights.blackcoffer.com/industrial-revolution-4-0-pros-and-cons/',
'https://insights.blackcoffer.com/impact-of-covid-19-coronavirus-on-the-indian-economy/',
'https://insights.blackcoffer.com/impact-of-covid-19-coronavirus-on-the-indian-economy-2/',
'https://insights.blackcoffer.com/impact-of-covid-19coronavirus-on-the-indian-economy/',
'https://insights.blackcoffer.com/impact-of-covid-19-coronavirus-on-the-global-economy/',
'https://insights.blackcoffer.com/ensuring-growth-through-insurance-technology/',
'https://insights.blackcoffer.com/blockchain-in-fintech/',
'https://insights.blackcoffer.com/blockchain-for-payments/',
'https://insights.blackcoffer.com/the-future-of-investing/',
'https://insights.blackcoffer.com/big-data-analytics-in-healthcare/',
'https://insights.blackcoffer.com/business-analytics-in-the-healthcare-industry/',
'https://insights.blackcoffer.com/challenges-and-opportunities-of-big-data-in-healthcare/',
'https://insights.blackcoffer.com/obstacles-to-data-driven-healthcare/',
'https://insights.blackcoffer.com/monetization-of-data-innovate-to-harvest-the-full-value-of-data/',
'https://insights.blackcoffer.com/traceability-of-information-master-your-data-capital/',
```

6. Perform Data Extraction from these URLs –

```
text = []
for url in li:
  text.append(requests.get(url,headers={"User-Agent": "XY"}))

for i in range(len(text)):
  text[i] = bfs.BeautifulSoup(text[i].content,'html.parser')

articles = []
for text in text:
  articles.append(text.find(attrs= {"class":"td-post-content"}).text)

for i in range(len(articles)):
  articles[i]= articles[i].replace('\n','')

stop_words = list(set(stopwords.words('english')))

sentences = []
for article in articles:
  sentences.append(len(sent_tokenize(article)))

cleaned_articles = [' ']*len(articles)
```

```
for i in range(len(articles)):
  for w in stop_words:
    cleaned_articles[i]= articles[i].replace(' '+w+' ',' ').replace('?','').replace('.','').replace(',','').replace(
    #? ! , .

words = []
for article in articles:
  words.append(len(word_tokenize(article)))

words_cleaned = []
for article in cleaned_articles:
  words_cleaned.append(len(word_tokenize(article)))
```

7. To Perform Analysis over the data, segregate them into Positive and Negative Words –

```
uploaded = files.upload()
```

```
Choose Files negative-words.txt
• negative-words.txt(text/plain) - 44758 bytes, last modified: 6/14/2022 - 100% done
Saving negative-words.txt to negative-words.txt
```

```
from google.colab import files

with open('negative-words.txt', 'w') as f:
  f.write('some content')

files.download('negative-words.txt')
```

```
uploaded = files.upload()
```

```
Choose Files positive-words.txt
• positive-words.txt(text/plain) - 19093 bytes, last modified: 6/14/2022 - 100% done
Saving positive-words.txt to positive-words (1).txt
```

```
from google.colab import files

with open('positive-words.txt', 'w') as f:
  f.write('some content')

files.download('positive-words.txt')
```

8. Next comes the text analysis –

```
positive_score = [0]*len(articles)
negative_score = [0]*len(articles)
```

```
words_cleaned = np.array(words_cleaned)
sentences = np.array(sentences)
```

```
df['POSITIVE SCORE'] = positive_score
df['NEGATIVE SCORE'] = negative_score
```

```
df['POLARITY SCORE'] = (df['POSITIVE SCORE']-df['NEGATIVE SCORE'])/ ((df['POSITIVE SCORE'] +df['NEGATIVE SCORE']) + 0.000001)
```

```
df['SUBJECTIVITY SCORE'] = (df['POSITIVE SCORE'] + df['NEGATIVE SCORE'])/( (words_cleaned) + 0.000001)
```

```
df['AVG SENTENCE LENGTH'] = np.array(words)/np.array(sentences)
```

```
complex_words = []
sylabble_counts = []
```

9. All the required fields are calculated –

- 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

```
for article in articles:
  sylabble_count=0
  d=article.split()
  ans=0
  for word in d:
    count=0
    for i in range(len(word)):
      if(word[i]=='a' or word[i]=='e' or word[i] =='i' or word[i] == 'o' or word[i] == 'u'):
           count+=1
#            print(words[i])
      if(i==len(word)-2 and (word[i]=='e' and word[i+1]=='d')):
        count-=1;
      if(i==len(word)-2 and (word[i]=='e' and word[i]=='s')):
        count-=1;
    sylabble_count+=count
    if(count>2):
        ans+=1
  sylabble_counts.append(sylabble_count)
  complex_words.append(ans)
```

```
df['PERCENTAGE OF COMPLEX WORDS'] = np.array(complex_words)/np.array(words)
```

```
df['FOG INDEX'] = 0.4 * (df['AVG SENTENCE LENGTH'] + df['PERCENTAGE OF COMPLEX WORDS'])
```

```
df['AVG NUMBER OF WORDS PER SENTENCES'] = df['AVG SENTENCE LENGTH']
```

```
df['COMPLEX WORD COUNT'] = complex_words
```

```
df['WORD COUNT'] = words
```

```
df['SYLLABLE PER WORD'] = np.array(sylabble_counts)/np.array(words)
```

```
total_characters = []
for article in articles:
  characters = 0
  for word in article.split():
    characters+=len(word)
  total_characters.append(characters)
```

```
personal_nouns = []
personal_noun =['I', 'we','my', 'ours','and' 'us','My','We','Ours','Us','And']
for article in articles:
  ans=0
  for word in article:
    if word in personal_noun:
      ans+=1
  personal_nouns.append(ans)
```

```
df['PERSONAL PRONOUN'] = personal_nouns
```

```
df['AVG WORD LENGTH'] = np.array(total_characters)/np.array(words)
```

10. Store the final output in the updated DataFrame –

```
df
```

	URL_ID	URL	POSITIVE SCORE	NEGATIVE SCORE	POLARITY SCORE	SUBJECTIVITY SCORE	AVG SENTENCE LENGTH	PERCENTAGE OF COMPLEX WORDS	FOG INDEX	AVG NUMBER OF WORDS PER SENTENCES	COMPLEX WORD COUNT	WORD COUNT	S
0	1.0	https://insights.blackcoffer.com/how-is-login-...	0	0	0.0	0.0	43.500000	0.227331	17.490932	43.500000	178	783	
1	2.0	https://insights.blackcoffer.com/how-does-ai-h...	0	0	0.0	0.0	34.300000	0.265306	13.826122	34.300000	182	686	
2	3.0	https://insights.blackcoffer.com/ai-and-its-im...	0	0	0.0	0.0	49.538462	0.274845	19.925323	49.538462	531	1932	
3	4.0	https://insights.blackcoffer.com/how-do-deep-l...	0	0	0.0	0.0	33.571429	0.276596	13.539210	33.571429	130	470	
4	5.0	https://insights.blackcoffer.com/how-artificia...	0	0	0.0	0.0	28.666667	0.239099	11.562306	28.666667	329	1376	
...	...	...	...	...	...	...	...	...	...	...	...	...	
165	167.0	https://insights.blackcoffer.com/role-big-data...	0	0	0.0	0.0	35.787234	0.237812	14.410018	35.787234	400	1682	
166	168.0	https://insights.blackcoffer.com/sales-forecas...	0	0	0.0	0.0	36.758621	0.268293	14.810765	36.758621	286	1066	
167	169.0	https://insights.blackcoffer.com/detect-data-e...	0	0	0.0	0.0	21.780000	0.263545	8.817418	21.780000	287	1089	

(The data frame contains all the above mentioned attributes)

11. Display the Articles which was used to performing text analysis –

```
articles
```

```
['When people hear AI they often think about sentient robots and magic boxes. AI today is much more mundane and sim
 'With increasing computing power and more data, the potential value of algorithms became higher. People and compan
 'If you were a fan of the 90's film Clueless back in the day, then you'll remember the protagonist, Cher Horowitz'
 'Understanding exactly how data is ingested, analyzed, and returned to the end-user can have a big impact on expec
 'From the stone age to the modern world, from hunting and gathering to cultivating crops, from living in caves to
 'Artificial intelligence (AI) is the development of computer systems that can perform tasks that normally require
 'Artificial intelligence (AI) is the most important branch of computer science in this era of big data. AI was bor
 '"In God we trust, all others must bring data." – W. Edwards Deming.2,000,000,000,000,000,000 bytes. The figure is
 'Big data refers to large sets of unstructured, semi-structured, or structured data obtained from numerous sources
 'Artificial intelligence (AI) is intelligence demonstrated by machines, as opposed to the natural intelligence dis
 'The big data industry has grown at an incredible rate as businesses realize the importance of insightful data ana
 'If I asked what is the one thing that this pandemic taught us? I believe in one word it will be: Adaptability, th
 'Big Data as the word suggests means a large amount of data. Data can be both in structured as well as unstructure
 'Will my Uber driver be a robot? Could the next medicine prescription be from an app? Do robots make calls for my
 'Defense is crucial to a nation's well-being and progress. For this reason, every country assigns tantamount impor
 'Have you heard the movie Big Hero 6? It's an animated movie in which this guy named Tadashi built AI machine Baym
 'Introduction"If anything kills over 10 million people in the next few decades, it will be a highly infectious vir
 'Human minds, a fascination in itself carrying the potential of tinkering nature with the pixie dust intelligence,
 'IntroductionAI is rapidly evolving in the employment sector, particularly in matters involving business and finan
 '"Anything that could give rise to smarter-than-human intelligence – in the form of Artificial Intelligence, brain
 '"Machine intelligence is the last invention that humanity will ever need to make"Nick BostromTo put it frankly, A
 'IntroductionWhere is this disruptive technology taking us? Take it or leave it, disruptive technology always crea
```

12. Finally convert the DataFrame into an Excel file and download it –

```
df.to_excel('Output Data Structure.xlsx')
```

```
files.download("Output Data Structure.xlsx")
```

```
Downloading "Output Data Structure.xlsx":
```

Output Excel File:

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		URL_ID	URL	SITIVE SCO	ATIVE SC	LARITY SC	ECTIVITY S	NTENCE LE	OF COMP	FOG INDEX	F WORDS	EX WORD	ORD COUN	BLE PER W	ONAL PRO	WORD LENGTH	
2	0	1	https://ir	0	0	0	0	43.5	0.22733	17.4909	43.5	178	783	1.73819	18	4.79183	
3	1	2	https://ir	0	0	0	0	34.3	0.26531	13.8261	34.3	182	686	1.81778	13	5	
4	2	3	https://ir	0	0	0	0	49.5385	0.27484	19.9253	49.5385	531	1932	1.83178	59	5.10041	
5	3	4	https://ir	0	0	0	0	33.5714	0.2766	13.5392	33.5714	130	470	1.85957	4	4.97872	
6	4	5	https://ir	0	0	0	0	28.6667	0.2391	11.5623	28.6667	329	1376	1.70349	49	4.6061	
7	5	6	https://ir	0	0	0	0	33.4054	0.30097	13.4826	33.4054	372	1236	1.839	28	5.00162	
8	6	7	https://ir	0	0	0	0	30	0.27333	12.1093	30	410	1500	1.79667	67	5.166	
9	7	8	https://ir	0	0	0	0	40.8542	0.28404	16.4553	40.8542	557	1961	1.85059	22	5.00561	
10	8	9	https://ir	0	0	0	0	32.4074	0.25371	13.0644	32.4074	444	1750	1.83371	25	4.976	
11	9	10	https://ir	0	0	0	0	25.0385	0.22734	10.1063	25.0385	148	651	1.65438	27	4.61137	
12	10	11	https://ir	0	0	0	0	31.1739	0.25244	12.5705	31.1739	181	717	1.84379	19	4.87169	
13	11	12	https://ir	0	0	0	0	27.8214	0.21438	11.2143	27.8214	334	1558	1.66624	30	4.64185	
14	12	13	https://ir	0	0	0	0	26.3958	0.23125	10.6508	26.3958	293	1267	1.71192	12	4.67482	
15	13	14	https://ir	0	0	0	0	27.4783	0.25	11.0913	27.4783	158	632	1.81329	7	4.87184	
16	14	15	https://ir	0	0	0	0	19.9487	0.21401	8.06509	19.9487	333	1556	1.65746	21	4.59833	
17	15	16	https://ir	0	0	0	0	39.6154	0.20777	15.9293	39.6154	214	1030	1.62039	27	4.62621	
18	16	17	https://ir	0	0	0	0	37.0377	0.30922	14.9388	37.0377	607	1963	1.90627	39	5.28884	
19	17	18	https://ir	0	0	0	0	23.8529	0.19852	9.62058	23.8529	322	1622	1.56165	37	4.35573	
20	18	19	https://ir	0	0	0	0	27.9701	0.28495	11.302	27.9701	534	1874	1.85272	31	5.02508	
21	19	20	https://ir	0	0	0	0	22.9487	0.2067	9.26217	22.9487	370	1790	1.61844	74	4.5324	
22	20	21	https://ir	0	0	0	0	28.75	0.21535	11.5861	28.75	421	1955	1.66394	36	4.66957	
23	21	22	https://ir	0	0	0	0	30.4792	0.21941	12.2794	30.4792	321	1463	1.65619	17	4.5851	
24	22	23	https://ir	0	0	0	0	20.3171	0.21729	8.21374	20.3171	181	833	1.64586	14	4.60864	
25	23	24	https://ir	0	0	0	0	30.9636	0.22372	12.4749	30.9636	381	1703	1.67998	12	4.58191	
26	24	25	https://ir	0	0	0	0	27.7586	0.19876	11.183	27.7586	160	805	1.60248	11	4.46832	
27	25	26	https://ir	0	0	0	0	35.0441	0.22241	14.1066	35.0441	530	2383	1.67478	87	4.57071	
28	26	27	https://ir	0	0	0	0	26.908	0.21956	10.851	26.908	514	2341	1.61897	32	4.51388	
29	27	28	https://ir	0	0	0	0	31.65	0.29226	12.7769	31.65	370	1266	1.95024	15	5.27883	
30	28	29	https://ir	0	0	0	0	21.7969	0.22366	8.80821	21.7969	312	1395	1.62509	38	4.56631	
31	29	30	https://ir	0	0	0	0	53.9091	0.21585	21.65	53.9091	384	1779	1.71501	55	4.7448	
32	30	31	https://ir	0	0	0	0	34.4423	0.23897	13.8725	34.4423	428	1791	1.72585	63	4.80625	

Sheet1