

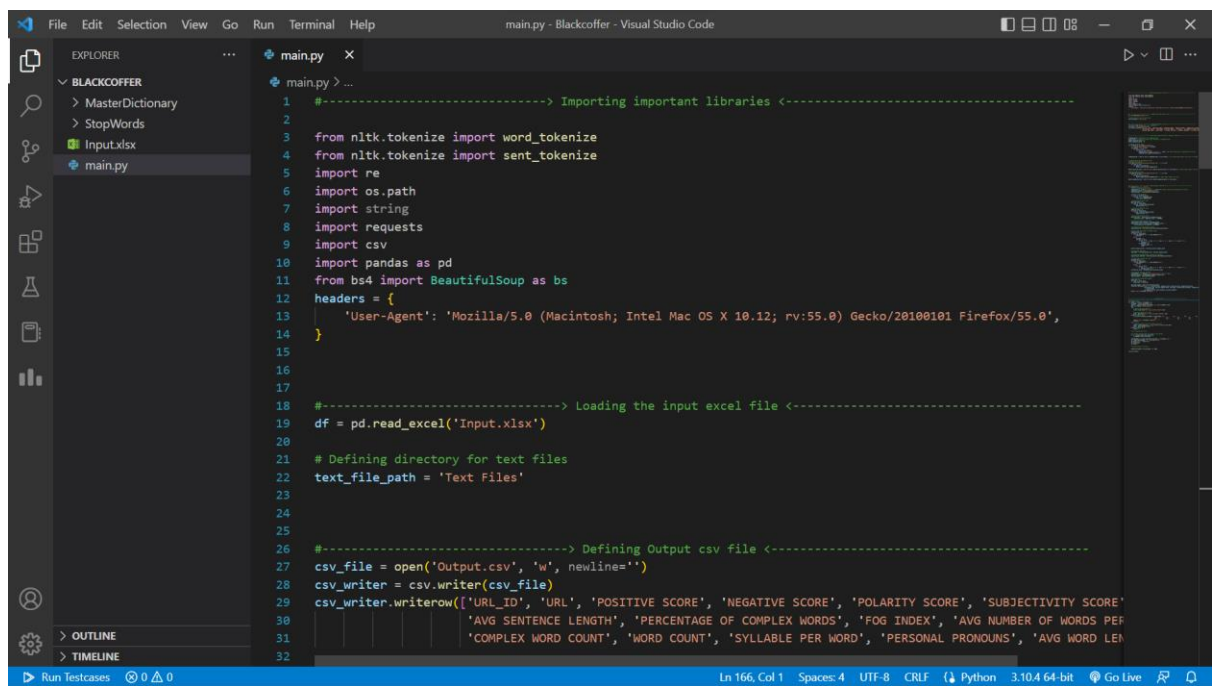
Instructions to use the code

1. Download the all files from the drive link.
2. Folder should contain the following files and folders in same directory or location:
 - a) MasterDictionary folder
 - b) StopWords folder
 - c) "Input.xlsx" file
 - d) "main.py" file

See Below:

Name	Date modified	Type	Size
MasterDictionary	30-10-2022 02:42 AM	File folder	
StopWords	30-10-2022 01:50 AM	File folder	
Input	29-10-2022 01:33 AM	Microsoft Excel Work...	15 KB
main	30-10-2022 05:58 PM	Python Source File	8 KB

3. Now open this directory in a code editor like: Vs code



```
1 #-----> Importing important libraries <-----
2
3 from nltk.tokenize import word_tokenize
4 from nltk.tokenize import sent_tokenize
5 import re
6 import os.path
7 import string
8 import requests
9 import csv
10 import pandas as pd
11 from bs4 import BeautifulSoup as bs
12 headers = {
13     'User-Agent': 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10.12; rv:55.0) Gecko/20100101 Firefox/55.0',
14 }
15
16
17
18 #-----> Loading the input excel file <-----
19 df = pd.read_excel('Input.xlsx')
20
21 # Defining directory for text files
22 text_file_path = 'Text Files'
23
24
25
26 #-----> Defining Output csv file <-----
27 csv_file = open('Output.csv', 'w', newline='')
28 csv_writer = csv.writer(csv_file)
29 csv_writer.writerow(['URL_ID', 'URL', 'POSITIVE SCORE', 'NEGATIVE SCORE', 'POLARITY SCORE', 'SUBJECTIVITY SCORE',
30                     'AVG SENTENCE LENGTH', 'PERCENTAGE OF COMPLEX WORDS', 'FOG INDEX', 'AVG NUMBER OF WORDS PER SENTENCE',
31                     'COMPLEX WORD COUNT', 'WORD COUNT', 'SYLLABLE PER WORD', 'PERSONAL PRONOUNS', 'AVG WORD LENGTH'])
32
```

4. Open “main.py” file.
5. Make sure your system has all the required libraries installed and your system is connected to the internet.
6. Now run the “main.py” file.
7. The “main.py” file will execute and the status will be shown in output terminal as below:

```
main.py x
main.py > ...
1  #-----> Importing important libraries <-----
2
3  from nltk.tokenize import word_tokenize
4  from nltk.tokenize import sent_tokenize
5  import re
6  import os.path
7  import string
8  import requests
9  import csv
10 import pandas as pd
11 from bs4 import BeautifulSoup as bs
12 headers = {
13     'User-Agent': 'Mozilla/5.0 (Macintosh; Intel Mac OS X 10_12; rv:55.0) Gecko/20100101 Firefox/55.0',
14 }
15
16
17
18 #-----> Loading the input excel file <-----
19 df = pd.read_excel('Input.xlsx')
```

PROBLEMS OUTPUT DEBUG CONSOLE **TERMINAL** JUPYTER

Windows PowerShell
Copyright (C) Microsoft Corporation. All rights reserved.

Install the latest PowerShell for new features and improvements! <https://aka.ms/PSWindows>

PS D:\Study\Others\Blackcoffer> python -u "d:\Study\Others\Blackcoffer\main.py"
file: 37 Completed....
file: 38 Completed....

8. As you run the script, a folder with name “Text Files” will be created in your working directory. This folder will contains all the contents of articles with their title in individual text files having URL ID as the name of text file.

a) “Text Files” Folder Created

MasterDictionary	30-10-2022 02:42 AM	File folder	
StopWords	30-10-2022 01:50 AM	File folder	
Text Files	30-10-2022 06:08 PM	File folder	
Input	29-10-2022 01:33 AM	Microsoft Excel Work...	15 KB
main	30-10-2022 05:58 PM	Python Source File	8 KB

b) Inside “Text Files” Folder

Name	Date modified	Type	Size
37	30-10-2022 06:08 PM	Text Document	12 KB
38	30-10-2022 06:08 PM	Text Document	9 KB
39	30-10-2022 06:08 PM	Text Document	11 KB
40	30-10-2022 06:08 PM	Text Document	10 KB
41	30-10-2022 06:08 PM	Text Document	11 KB
42	30-10-2022 06:08 PM	Text Document	8 KB
43	30-10-2022 06:08 PM	Text Document	5 KB

- On running the “main.py” file, an “Output.csv” file will also be created in the same directory. This file will contain all the analysis data of each article as required.

Directory

MasterDictionary	30-10-2022 02:42 AM	File folder	
StopWords	30-10-2022 01:50 AM	File folder	
Text Files	30-10-2022 06:08 PM	File folder	
Input	29-10-2022 01:33 AM	Microsoft Excel Work...	15 KB
main	30-10-2022 05:58 PM	Python Source File	8 KB
Output	30-10-2022 06:08 PM	Microsoft Excel Com...	0 KB

Preview of “Output.csv” file

A1	URL_ID	URL	POSITIVE	NEGATIVE	POLARITY	SUBJECTIV	AVG SENT	PERCENTA	FOG INDE	AVG NUM	COMPLEX	WORD CO	SYLLABLE	PERSONAL	AVG WOR
2	37	https://ins	61	33	0.297872	0.090559	13.65789	2.105477	6.305349	13.65789	493	1038	2.385356	1	7.44316
3	38	https://ins	117	70	0.251337	0.111842	21.16456	2.331939	9.398598	21.16456	717	1672	2.242225	6	7.182416
4	39	https://ins	183	103	0.27972	0.110853	30.71429	2.275132	13.19577	30.71429	1134	2580	2.269767	2	7.25
5	40	https://ins	240	128	0.304348	0.110743	35.35106	2.343441	15.0778	35.35106	1418	3323	2.242552	17	7.121878
6	41	https://ins	290	153	0.309255	0.105526	53.82051	2.398857	22.48775	53.82051	1750	4198	2.224869	14	7.078847
7	42	https://ins	333	177	0.305882	0.105941	78.91803	2.432542	32.54023	78.91803	1979	4814	2.218114	18	7.058995
8	43	https://ins	358	189	0.308958	0.105497	117.8409	2.476122	48.12681	117.8409	2094	5185	2.199036	7	7.039923
9	44	https://ins	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
10	45	https://ins	394	202	0.322148	0.107329	158.6571	2.49573	64.46115	158.6571	2225	5553	2.182424	0	6.993337
11	46	https://ins	458	242	0.308571	0.106691	82.0125	2.491834	33.80173	82.0125	2633	6561	2.193263	9	6.980034
12	47	https://ins	503	307	0.241975	0.106411	73.19231	2.507246	30.27982	73.19231	3036	7612	2.181818	1	6.949422
13	48	https://ins	537	326	0.244496	0.104555	144.807	2.460942	58.90718	144.807	3354	8254	2.206324	10	6.98849
14	49	https://ins	564	350	0.234136	0.10265	109.9259	2.477462	44.96136	109.9259	3594	8904	2.199124	6	6.96204
15	50	https://ins	624	374	0.250501	0.103764	155.129	2.478227	63.0429	155.129	3881	9618	2.200665	21	6.980245
16	51	https://ins	690	397	0.269549	0.104299	127.0976	2.482611	51.83207	127.0976	4198	10422	2.199098	9	6.985224
17	52	https://ins	716	397	0.286613	0.104106	509.0952	2.470765	204.6264	509.0952	4327	10691	2.2055	0	6.99841
18	53	https://ins	795	436	0.291633	0.107878	1426.375	2.455034	571.532	1426.375	4648	11411	2.212514	14	6.999562
19	54	https://ins	820	436	0.305732	0.105281	213.0357	2.445173	86.19235	213.0357	4879	11930	2.213747	0	6.99321
20	55	https://ins	834	442	0.30721	0.103529	300.6098	2.436252	121.2184	300.6098	5059	12325	2.216714	0	7.011521
21	56	https://ins	837	446	0.304754	0.102829	1134.273	2.44025	454.6852	1134.273	5113	12477	2.214715	2	7.013064
22	57	https://ins	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA	NA
23	58	https://ins	861	513	0.253275	0.104099	239.9818	2.44834	96.97206	239.9818	5391	13199	2.206455	2	7.002046
24	59	https://ins	878	537	0.240877	0.10266	482.0386	2.453871	104.553	482.0386	5537	12550	2.207731	2	6.996326

Note:

- Let the “main.py” file run completely and do not interrupt it in middle. It will run till “file : 150”. Only after that you will see the data in “Output.csv” file.
- Some links in “Input.xlsx” file were not functional. So, for those links, the data of columns in “Output.csv” file is filled with “NA” value.

Example: URL_ID: 44, URL_ID: 57