\*\* First to execute the program install all the required packages mentioned in "requirement.txt" \*\*

\*\* Run “main.py” to execute the automation script \*\*

1. "text\_extraction.py" scraps the urls to collect the header and content of the articles using BeautifulSoup.

response = requests.get(url, headers={'User-Agent': 'XY'})

soup = BeautifulSoup(response.text, 'lxml')

2. The articles data is stored in "files/articles/URL\_ID".

3. Create a master dictionary from "LoughranMcDonald\_MasterDictionary\_2020.xlsx" to create further `pos\_dic` and `neg\_dic`.

master\_dic = pd.read\_excel('files/LoughranMcDonald\_MasterDictionary\_2020.xlsx')

pos\_dict = [x for x in master\_dic.loc[master\_dic['Positive'] != 0]['Word']]

neg\_dict = [x for x in master\_dic.loc[master\_dic['Negative'] != 0]['Word']]

4. Create a list of stop words from "StopWords\_Generic.txt" to further use in text analysis.

with open('files/StopWords\_Generic.txt', 'r') as file:

stop\_words = file.read()

5. Use custom stop words list to clean the data in the articles and further compute variables.

cleaned\_data = [word for word in words if word not in stop\_words]

6. Use `pos\_dic` and `neg\_dic` to calculate `positive\_score` and `negative\_score`, calculate other variables as mentioned in "Text Analysis.docx".

pos\_score = 0

neg\_score = 0

for word in cleaned\_data:

if word in pos\_dict:

pos\_score += 1

elif word in neg\_dict:

neg\_score += 1

polarity\_score = (pos\_score - neg\_score) / ((pos\_score + neg\_score) + 0.000001)

subjectivity\_score = (pos\_score + neg\_score)/ ((len(cleaned\_data)) + 0.000001)

see the next page

7. Clean the data for sentiment analysis by removing stop words using `nltk.corpus` stopwords and punctuations from the content.

# removing stopwords using nltk stopwords

stop\_words\_nltk = stopwords.words('english')

cleaned\_data\_nltk = [word for word in words if word not in stop\_words\_nltk]

# removing punctuations

punctuations = '''!()-[]{};:'"\,<>./?@#$%^&\*\_~'''

cleaned\_data\_nltk = [word for word in cleaned\_data\_nltk if word not in punctuations]

8. After calculating all variables simple write all of the variables to "Output\_Data\_Structures.xlsx" using Pandas `to\_excel(filename)`

output.to\_excel('Output\_Data\_Structure.xlsx')

**Caution:** If shows any ERROR please verify if all requirements are satisfied as mentioned above.