In [1]:

```
import pandas as pd
import os
import glob
import shutil
import requests
from string import punctuation
from nltk.tokenize import word_tokenize,sent_tokenize
import re
from nltk.stem import LancasterStemmer,WordNetLemmatizer
```

In [2]:

```
df=pd.read_excel(r"input.xlsx")
```

In [3]:

```
url='https://insights.blackcoffer.com/in-future-or-in-upcoming-years-humans-and-machines-are-going-to-work-together-in-every

data=requests.get(url)
data.text  #Access has been prohibited so instead of beatuful soup we will use scrapy.

#Web scrapping using scrapy has been done on vs code, have shared code in vs code folder.
```

Out[3]:

```
'<head><title>Not Acceptable!</title></head><body><h1>Not Acceptable!</h1><p>An appropriate representation of the requested res
ource could not be found on this server. This error was generated by Mod_Security.</p></body></html>'
```

In [4]:

```
neDrive\Desktop\DataScience\Projects\Blackcoffer\20211030 Test Assignment\vscode\scraping_text\scraping_text\Extracted_data.csv')


```

Out[4]:

	text	title	url_id	URL
0	“If anything kills over 10 million people in t...	AI in healthcare to Improve Patient Outcomes	37	https://insights.blackcoffer.com/ai-in-healthc...
1	Where is this disruptive technology taking us?...	Will machine replace the human in the future o...	42	https://insights.blackcoffer.com/what-if-the-c...
2	Human minds, a fascination in itself carrying ...	What if the Creation is Taking Over the Creator?	38	https://insights.blackcoffer.com/what-jobs-wil...
3	“Anything that could give rise to smarter-than...	Will Machine Replace The Human in the Future o...	40	https://insights.blackcoffer.com/will-machine-...
4	“Machine intelligence is the last invention th...	Will AI Replace Us or Work With Us?	41	https://insights.blackcoffer.com/will-ai-repla...

In [5]:

```
#removind special characters from text data,so that can be stored in txt file.


def rm_char(data):

    res=re.sub("\u20b9","",data)
    return res

extracted_df["text"]= extracted_df["text"].apply(rm_char)
extracted_df.head()
```

Out[5]:

	text	title	url_id	URL
0	“If anything kills over 10 million people in t...	AI in healthcare to Improve Patient Outcomes	37	https://insights.blackcoffer.com/ai-in-healthc...
1	Where is this disruptive technology taking us?...	Will machine replace the human in the future o...	42	https://insights.blackcoffer.com/what-if-the-c...
2	Human minds, a fascination in itself carrying ...	What if the Creation is Taking Over the Creator?	38	https://insights.blackcoffer.com/what-jobs-wil...
3	“Anything that could give rise to smarter-than...	Will Machine Replace The Human in the Future o...	40	https://insights.blackcoffer.com/will-machine-...
4	“Machine intelligence is the last invention th...	Will AI Replace Us or Work With Us?	41	https://insights.blackcoffer.com/will-ai-repla...

In [6]:

```
shutil.rmtree("Gathered_texts")
os.mkdir('Gathered_texts')
```

Putting all stopwords in a single place, inside a list.

In [7]:

```
#Putting all stopwords in a single place inside list.

stop_word_paths=glob.glob('StopWords\*.txt')     #getting stopwords paths
for path in stop_word_paths:
    with open(path,'r') as file:
        data=file.read()
        with open('stop_words.txt','a') as file1:
            file1.write(data)

with open('stop_words.txt','r') as file:
    all_stopwords=file.read()

cleaned=re.sub("\W+"," ",all_stopwords).lower()
stopwords_list=[word.lower() for word in re.sub("\d+","",cleaned).split() if word not in punctuation]
```

Preprocessing negative words

In [8]:

```
with open(r"MasterDictionary/negative-words.txt") as file:
    negative_data=file.read()

cleaned=re.sub("\W+"," ",negative_data)
cleaned_neg_text=[word.lower() for word in word_tokenize(re.sub("\d+","",cleaned)) if (word not in punctuation) and (word.lo

#Performing lemmatization to convert each word into its root word.
final_neg_words=[]
lemma=WordNetLemmatizer()

for word in cleaned_neg_text:
    final_neg_words.append(lemma.lemmatize(word,'v'))
```

Preprocessing positive words

In [9]:

```
with open(r"MasterDictionary/positive-words.txt") as file:
    positive_data=file.read()

cleaned=re.sub("\W+"," ",positive_data)
cleaned_pos_text=[word.lower() for word in word_tokenize (re.sub("\d+","",cleaned)) if (word not in punctuation) and (word.l

#Performing lemmatization to convert each word into its root word.
final_pos_words=[]
lemma=WordNetLemmatizer()

for word in cleaned_pos_text:
    final_pos_words.append(lemma.lemmatize(word,'v'))
```

Saving Extracted data in .txt file having name as url_id and Analysing text

In [10]:

```
url_no=[]
url=[]
pos_score=[]
neg_score=[]
pol_score=[]
sub_score=[]
avg_sent_length=[]
p_of_complex=[]
fog_index=[]
avg_word_per_sent=[]
word_counts=[]
syllable_per_word=[]
personal_pronounce=[]
avg_word_length=[]
complex_word_count_data=[]


for text,title,url_id,link in extracted_df.values:
    loc=str(url_id)+".txt"
    with open(os.path.join('Gathered_texts',loc),'a') as file:
        file.write(title)
        file.write('\n')
        file.write(text)


    output_text=title+' '+text

    cleaned=re.sub("\W+"," ",output_text)
    word_list=[word.lower() for word in word_tokenize (re.sub("\d+","",cleaned)) if (word not in punctuation) and (word.lowe

    final_lemmatized_words=[]
    lemma=WordNetLemmatizer()

    for word in word_list:
        final_lemmatized_words.append(lemma.lemmatize(word,'v'))

    pos=0
    neg=0

    for words in  final_lemmatized_words:
        if words in final_pos_words:
            pos+=1
        elif words in final_neg_words:
            neg-=1

    neg*=-1

    polarity=(pos-neg)/((pos+neg)+0.000001)
    sub=(pos + neg)/ ((len(final_lemmatized_words)) + 0.000001)

    #Performing Sentence tokenization to find no of sentences.
    no_of_sent=sent_tokenize(output_text)

    #Performing word tokenization to find no of words
    data_word=re.sub('[.")()]'," ",output_text)
    word_no=word_tokenize(data_word)

    avg_len=len(word_no)/len(no_of_sent)

    #Finding complex words from sentence

    complex_words=[]

    for word in word_no:
        count=0
        for char in word:
            if char.lower() in ["a","e","i","o","u"] and not(word.endswith("es"))and not(word.endswith("ed")):
                count+=1

        if count>2:
            complex_words.append(word)


    per_compl_score=len(complex_words)/len(word_no)
    fog=(0.4*avg_len)+per_compl_score
    avg_no_of_words_per_sent=avg_len
    complex_words_count=len(complex_words)
    word_count=len(final_lemmatized_words)

    count=0
    for word in word_no:
        for char in word:
            if char.lower() in ["a","e","i","o","u"] and not(word.endswith("es"))and not(word.endswith("ed")):
                count+=1

    syll_per_word=count/len(word_no)
```

```
personal_pronouns=0

text='i we my ours us'
for word in word_tokenize(output_text):

    if word !="US":
        if word.lower() in text.split():
            personal_pronouns +=1
total_char=0
for word in word_no:
    for char in word:
        total_char+=1

avr_word_len=total_char/len(word_no)




url_no.append(url_id)
url.append(link)
pos_score.append(pos)
neg_score.append(neg)
pol_score.append(polarity)
sub_score.append(sub)
avg_sent_length.append(avg_len)
p_of_complex.append(per_compl_score)
fog_index.append(fog)
avg_word_per_sent.append(avg_no_of_words_per_sent)
complex_word_count_data.append(complex_words_count)
word_counts.append(word_count)
syllable_per_word.append(syll_per_word)
personal_pronounce.append(personal_pronouns)
avg_word_length.append(avr_word_len)
```

In [11]:

```
output=pd.DataFrame({"URL_ID":url_no,"URL":url,"POSITIVE SCORE":pos_score,"NEGATIVE SCORE":neg_score,"POLARITY SCORE":pol_sc
```

In [12]:

```
output.head()
```

Out[12]:

URL	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	SYLLAE P WO
ttps://insights.blackcoffer.com/ai-in-healthc...	102	54	0.307692	0.165605	25.712329	0.272243	10.557174	25.712329	511	942	1.788
tps://insights.blackcoffer.com/what-if-the-c...	67	43	0.218182	0.216963	24.232143	0.187915	9.880772	24.232143	255	507	1.571
tps://insights.blackcoffer.com/what-jobs-wil...	86	57	0.202797	0.269303	19.637500	0.161044	8.016044	19.637500	253	531	1.457
ıttps://insights.blackcoffer.com/will-machine-...	81	39	0.350000	0.215827	18.269663	0.184502	7.492367	18.269663	300	556	1.635
ıttps://insights.blackcoffer.com/will-ai-repla...	81	50	0.236641	0.185816	23.974026	0.194475	9.784085	23.974026	359	705	1.575

In [13]:

```
output.to_excel("Result.xlsx",index=False)
```

In [ ]: