### Text_data (Quiz-4)

In [1]:

```
import pandas as pd
import os
```

In [2]:

```
os.getcwd()
```

Out[2]:

```
'C:\\Users\\Irfan\\Downloads'
```

In [111]:

```
df = os.path.join(os.getcwd(),r"C:\Users\Irfan\Downloads\text_data\text")
```

In [112]:

```
len(list(os.walk(df)))
```

Out[112]:

```
1
```

In [113]:

```
d = []
for root, folders, files in os.walk(df):
    for file in files:
        a = os.path.join(root,file)
        with open(a) as inf:
            d.append(inf.read())
```

In [114]:

```
d
```

```
 'Premise\ngood exercises can often successfully alleviating one's heavy burdens\n',
 "Claim\nrigorous exercise can also hone one's will, which enable people to be able to tackle stringent problems and ard
uous tasks\n",
 "Premise\nThis hormone can delight one's mind, creating positive feelings\n",
 'Premise\nWhen people exercise, the neuron cells in the brain release a chemical compound named Dopamine\n',
 'Claim\nSports activities can improve mental health\n',
 "MajorClaim\nI agree only to certain degree that in today's world, image serves as a more effective means of communicat
ion\n",
 'MajorClaim\nConsideration and communication, in my personal opinion, are the most important quality\n',
 'MajorClaim\nboth images and words go hand in hand and one cannot wholly emphasise on only one aspect, either images or
words\n',
 'Claim\nit is undeniable that images in the absence of words can obviously claim the attraction of many\n',
 'Claim\npictures can influence the way people think\n',
 'Premise\nnowadays horrendous images are displayed on the cigarette boxes to illustrate the consequences of smoking\n',
 'Premise\nstatistics show a slight reduction in the number of smokers, indicating that they realize the effects of the
negative habit\n',
 'Premise\nthe magnificent photograph captured by Kevin Carter, which portrayed a starving Sudanese child struck by extr
eme poverty has successfully highlighted the plight faced by the citizens in Sudan\n',
 'Premise\nimages are also widely used in newspapers, magazines and advertisements\n',
 'Claim\nwritten words are also vital in order to spread across certain messages\n',
 'Premise\nWith only pictures, everyone is left to their own interpretation on how they perceive the images\n'
```

In [115]:

```
df = pd.DataFrame(d)
df
```

Out[115]:

| | 0 |
|---|---|
| **0** | MajorClaim\nwe should attach more importance t... |
| **1** | Premise\nTake Olympic games which is a form of... |
| **2** | Premise\nThe high technology and new ideas app... |
| **3** | Premise\npollutions are not just caused by the... |
| **4** | Premise\nthe improvements of work efficiency a... |
| **...** | ... |
| **6084** | MajorClaim\naddressing pollution and traffic i... |
| **6085** | Premise\nwhether it can work out for alleviati... |
| **6086** | Premise\nprice control institution has been us... |
| **6087** | Claim\nit seems not easy to increase petrol pr... |
| **6088** | Premise\ngovernments have a macro-economic per... |

6089 rows × 1 columns

In [117]:

```
a=[]
b= []
for i in df[0]:
    a.append(i.split('\n')[0])
    b.append(i.split('\n')[1])
df['label'] = a
df['text'] = b
```

In [118]:

```
df['label']
```

Out[118]:

```
0       MajorClaim
1          Premise
2          Premise
3          Premise
4          Premise
           ...
6084    MajorClaim
6085       Premise
6086       Premise
6087         Claim
6088       Premise
Name: label, Length: 6089, dtype: object
```

In [119]:

```
df['text']
```

Out[119]:

```
0       we should attach more importance to cooperatio...
1       Take Olympic games which is a form of competit...
2       The high technology and new ideas applied into...
3       pollutions are not just caused by the burning ...
4       the improvements of work efficiency also attri...
                              ...
6084    addressing pollution and traffic issues only b...
6085    whether it can work out for alleviating traffi...
6086    price control institution has been used in ple...
6087    it seems not easy to increase petrol price ins...
6088    governments have a macro-economic perspective ...
Name: text, Length: 6089, dtype: object
```

In [120]:

```
df = df.drop(0,axis=1)
```

In [121]:

```
df
```

Out[121]:

| | label | text |
|---|---|---|
| **0** | MajorClaim | we should attach more importance to cooperatio... |
| **1** | Premise | Take Olympic games which is a form of competit... |
| **2** | Premise | The high technology and new ideas applied into... |
| **3** | Premise | pollutions are not just caused by the burning ... |
| **4** | Premise | the improvements of work efficiency also attri... |
| **...** | ... | ... |
| **6084** | MajorClaim | addressing pollution and traffic issues only b... |
| **6085** | Premise | whether it can work out for alleviating traffi... |
| **6086** | Premise | price control institution has been used in ple... |
| **6087** | Claim | it seems not easy to increase petrol price ins... |
| **6088** | Premise | governments have a macro-economic perspective ... |

6089 rows × 2 columns

In [124]:

```
df.shape
```

Out[124]:

```
(6089, 2)
```

In [125]:

```
df['label'].value_counts()
```

Out[125]:

```
Premise       3832
Claim         1506
MajorClaim     751
Name: label, dtype: int64
```

In [126]:

```
len(df['text'])
```

Out[126]:

```
6089
```

In [127]:

```
import nltk
nltk.download('stopwords')
# Downloading wordnet before applying Lemmatizer
nltk.download('wordnet')
nltk.download('omw-1.4')
```

```
[nltk_data] Downloading package stopwords to
[nltk_data]     C:\Users\Irfan\AppData\Roaming\nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to
[nltk_data]     C:\Users\Irfan\AppData\Roaming\nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
[nltk_data] Downloading package omw-1.4 to
[nltk_data]     C:\Users\Irfan\AppData\Roaming\nltk_data...
[nltk_data]   Package omw-1.4 is already up-to-date!
```

Out[127]:

```
True
```

In [128]:

```
import re
import nltk
from nltk.tokenize import word_tokenize
from nltk.corpus import stopwords
from nltk.stem.porter import PorterStemmer
from nltk.stem import WordNetLemmatizer
```

In [129]:

```
stemmer = PorterStemmer()
lemmatizer = WordNetLemmatizer()
```

In [130]:

```
def preprocess(df, flag):
    # Removing special characters and digits
    sentence = re.sub("[^a-zA-Z]", " ", df)

    # change sentence to lower case
    sentence = sentence.lower()

    # tokenize into words
    tokens = sentence.split()

    # remove stop words
    clean_tokens = [t for t in tokens if t not in stopwords.words("english")]

    # Stemming/Lemmatization
    if(flag == 'stem'):
        clean_tokens = [stemmer.stem(word) for word in clean_tokens]
    else:
        clean_tokens = [lemmatizer.lemmatize(word) for word in clean_tokens]

    return pd.Series([" ".join(clean_tokens), len(clean_tokens)])
```

In [81]:

```
def preprocess(raw_text, flag):
    # Removing special characters and digits
    sentence = re.sub("[^a-zA-Z]", " ", raw_text)
    print(len(sentence))
    # change sentence to lower case
    sentence = sentence.lower()

    # tokenize into words
    tokens = sentence.split()

    # remove stop words
    clean_tokens = [t for t in tokens if not t in stopwords.words("english")]

    # Stemming/Lemmatization
    if(flag == 'stem'):
        clean_tokens = [stemmer.stem(word) for word in clean_tokens]
    else:
        clean_tokens = [lemmatizer.lemmatize(word) for word in clean_tokens]

    return pd.Series( len(sentence))
```

In [82]:

```
temp_df = df['text'].progress_apply(lambda x: preprocess(x, 'stem'))

temp_df.max()
```

```
145

100%|██████████████████████████████████████████████████████████████████████████▉| 6086/6089 [01:11<00:00, 71.25it/s]

143
142
82
85
81
70
84
92
131
52
107

100%|███████████████████████████████████████████████████████████████████████████| 6089/6089 [01:12<00:00, 83.55it/s]
```

Out[82]:

```
0    344
```

In [83]:

```
def preprocess(raw_text, flag):
    # Removing special characters and digits
    sentence = re.sub("[^a-zA-Z]", " ", raw_text)
    print(len(sentence))
    # change sentence to lower case
    sentence = sentence.lower()

    # tokenize into words
    tokens = sentence.split()

    # remove stop words
    clean_tokens = [t for t in tokens if not t in stopwords.words("english")]

    # Stemming/Lemmatization
    if(flag == 'stem'):
        clean_tokens = [stemmer.stem(word) for word in clean_tokens]
    else:
        clean_tokens = [lemmatizer.lemmatize(word) for word in clean_tokens]

    return pd.Series([" ".join(tokens), len(tokens)])
```

In [85]:

```
temp_df = df['text'].progress_apply(lambda x: preprocess(x, 'stem'))

temp_df.max()
```

```
134
145
143
142
82
85
81
70
84
92
131
52
107

100%|████████████████████████████████████████████████████████████████████████████████| 6089/6089 [01:16<00:00, 80.03it/s]
```

Out[85]:

```
0    zoos which are equipped with modern facilities...
1                                                   67
```

In [89]:

```
def preprocess(raw_text, flag):
    # Removing special characters and digits
    sentence = re.sub("[^a-zA-Z]", " ", raw_text)
    print(len(sentence))
    # change sentence to lower case
    sentence = sentence.lower()

    # tokenize into words
    tokens = sentence.split()

    # remove stop words
    clean_tokens = [t for t in tokens if not t in stopwords.words("english")]

    # Stemming/Lemmatization
    if(flag == 'stem'):
        clean_tokens = [stemmer.stem(word) for word in clean_tokens]
    else:
        clean_tokens = [lemmatizer.lemmatize(word) for word in clean_tokens]

    return pd.Series([" ".join(clean_tokens), len(clean_tokens)])
```

In [91]:

```
temp_df = df['text'].progress_apply(lambda x: preprocess(x, 'lemma'))

temp_df.max()
```

```
134
145
143
142
82
85
81
70
84
92
131
52
107

100%|████████████████████████████████████████████████████████████████████████████████| 6089/6089 [01:11<00:00, 85.64it/s]
```

Out[91]:

```
0    zoo useful earth
1                  31
```

In [ ]:

In [92]:

```
from tqdm import tqdm, tqdm_notebook
tqdm.pandas()
```

In [93]:

```
preprocess('text','clean_tokens')
```

```
4
```

Out[93]:

```
0    text
1       1
dtype: object
```

In [94]:

```
preprocess('text','tokens')
```

```
4
```

Out[94]:

```
0    text
1       1
dtype: object
```

In [95]:

```
temp_df = df['text'].apply(lambda x : preprocess(x, 'stem'))

temp_df.head()
```

| | 0 | 1 |
|---|---|---|
| **0** | attach import cooper primari educ | 5 |
| **1** | take olymp game form competit instanc hard ima... | 25 |
| **2** | high technolog new idea appli practic may lead... | 15 |
| **3** | pollut caus burn oil chemic pollut extra light... | 13 |
| **4** | improv work effici also attribut speed work pa... | 13 |
| **...** | ... | ... |
| **6084** | address pollut traffic issu increas oil price ... | 8 |
| **6085** | whether work allevi traffic pollut pressur oug... | 8 |
| **6086** | price control institut use plenti social area ... | 13 |
| **6087** | seem easi increas petrol price instantli | 6 |
| **6088** | govern macro econom perspect control price var... | 11 |

In [96]:

```
temp_df.columns = ['clean_text_stem', 'text_length_stem']

temp_df
```

Out[96]:

| | clean_text_stem | text_length_stem |
|---|---|---|
| **0** | attach import cooper primari educ | 5 |
| **1** | take olymp game form competit instanc hard ima... | 25 |
| **2** | high technolog new idea appli practic may lead... | 15 |
| **3** | pollut caus burn oil chemic pollut extra light... | 13 |
| **4** | improv work effici also attribut speed work pa... | 13 |
| **...** | ... | ... |
| **6084** | address pollut traffic issu increas oil price ... | 8 |
| **6085** | whether work allevi traffic pollut pressur oug... | 8 |
| **6086** | price control institut use plenti social area ... | 13 |
| **6087** | seem easi increas petrol price instantli | 6 |
| **6088** | govern macro econom perspect control price var... | 11 |

6089 rows × 2 columns

In [97]:

```
df = pd.concat([df, temp_df], axis=1)

df
```

Out[97]:

| | label | text | clean_text_stem | text_length_stem | clean_text_lemma | text_length_lemma | clean_text_stem | text_length_stem | clean_text_lem |
|---|---|---|---|---|---|---|---|---|---|
| **0** | MajorClaim | we should attach more importance to cooperatio... | attach import cooper primari educ | 5 | attach importance cooperation primary education | 5 | attach import cooper primari educ | 5 | attach importa coopera primary educa |
| **1** | Premise | Take Olympic games which is a form of competit... | take olymp game form competit instanc hard ima... | 25 | take olympic game form competition instance ha... | 25 | take olymp game form competit instanc hard ima... | 25 | take olympic ga form competi instance I |
| **2** | Premise | The high technology and new ideas applied into... | high technolog new idea appli practic may lead... | 15 | high technology new idea applied practice may ... | 15 | high technolog new idea appli practic may lead... | 15 | high technol new idea app practice ma |
| **3** | Premise | pollutions are not just caused by the burning ... | pollut caus burn oil chemic pollut extra light... | 13 | pollution caused burning oil chemical pollutan... | 13 | pollut caus burn oil chemic pollut extra light... | 13 | pollution cau burnin chemical pollut |
| **4** | Premise | the improvements of work efficiency also attri... | improv work effici also attribut speed work pa... | 13 | improvement work efficiency also attribute spe... | 13 | improv work effici also attribut speed work pa... | 13 | improvement v efficiency attribute sp |
| **...** | ... | ... | ... | ... | ... | ... | ... | ... | |
| **6084** | MajorClaim | addressing pollution and traffic issues only b... | address pollut traffic issu increas oil price ... | 8 | addressing pollution traffic issue increasing ... | 8 | address pollut traffic issu increas oil price ... | 8 | address pollution tr issue increasin |
| **6085** | Premise | whether it can work out for alleviating traffi... | whether work allevi traffic pollut pressur oug... | 8 | whether work alleviating traffic pollution pre... | 8 | whether work allevi traffic pollut pressur oug... | 8 | whether v alleviating tr pollution p |
| **6086** | Premise | price control institution has been used in ple... | price control institut use plenti social area ... | 13 | price control institution used plenty social a... | 13 | price control institut use plenti social area ... | 13 | price cor institution u plenty social |
| **6087** | Claim | it seems not easy to increase petrol price ins... | seem easi increas petrol price instantli | 6 | seems easy increase petrol price instantly | 6 | seem easi increas petrol price instantli | 6 | seems e increase pe price insta |
| **6088** | Premise | governments have a macro-economic perspective ... | govern macro econom perspect control price var... | 11 | government macro economic perspective control ... | 11 | govern macro econom perspect control price var... | 11 | government ma econc perspective cor |

6089 rows × 12 columns

In [98]:

```
temp_df = df['text'].apply(lambda x: preprocess(x, 'lemma'))

temp_df
```

| | 0 | 1 |
|---|---|---|
| **0** | attach importance cooperation primary education | 5 |
| **1** | take olympic game form competition instance ha... | 25 |
| **2** | high technology new idea applied practice may ... | 15 |
| **3** | pollution caused burning oil chemical pollutan... | 13 |
| **4** | improvement work efficiency also attribute spe... | 13 |
| **...** | ... | ... |
| **6084** | addressing pollution traffic issue increasing ... | 8 |
| **6085** | whether work alleviating traffic pollution pre... | 8 |
| **6086** | price control institution used plenty social a... | 13 |
| **6087** | seems easy increase petrol price instantly | 6 |
| **6088** | government macro economic perspective control ... | 11 |

In [101]:

```
temp_df.columns = ['clean_text_lemma', 'text_length_lemma']

temp_df.head()
```

Out[101]:

| | clean_text_lemma | text_length_lemma |
|---|---|---|
| **0** | attach importance cooperation primary education | 5 |
| **1** | take olympic game form competition instance ha... | 25 |
| **2** | high technology new idea applied practice may ... | 15 |
| **3** | pollution caused burning oil chemical pollutan... | 13 |
| **4** | improvement work efficiency also attribute spe... | 13 |

In [131]:

```
df = pd.concat([df, temp_df], axis=1)

df.head()
```

Out[131]:

| | label | text | clean_text_lemma | text_length_lemma |
|---|---|---|---|---|
| **0** | MajorClaim | we should attach more importance to cooperatio... | attach importance cooperation primary education | 5 |
| **1** | Premise | Take Olympic games which is a form of competit... | take olympic game form competition instance ha... | 25 |
| **2** | Premise | The high technology and new ideas applied into... | high technology new idea applied practice may ... | 15 |
| **3** | Premise | pollutions are not just caused by the burning ... | pollution caused burning oil chemical pollutan... | 13 |
| **4** | Premise | the improvements of work efficiency also attri... | improvement work efficiency also attribute spe... | 13 |

In [147]:

```
from sklearn.feature_extraction.text import CountVectorizer
vocab = CountVectorizer(ngram_range=[0,1])
dtm = vocab.fit_transform(df['clean_text_lemma'])
```

In [148]:

```
len(vocab.vocabulary_)
```

Out[148]:

```
5974
```

In [134]:

```
vocab = CountVectorizer(ngram_range=[1,2])

dtm = vocab.fit_transform(df['clean_text_lemma'])
```

In [135]:

```
#print(vocab.vocabulary_)
```

In [136]:

```
len(vocab.vocabulary_)
```

Out[136]:

```
41969
```

In [137]:

```
vocab = CountVectorizer(ngram_range=[1,3])

dtm = vocab.fit_transform(df['clean_text_lemma'])
```

In [138]:

```
len(vocab.vocabulary_)
```

Out[138]:

```
79726
```

In [ ]:

In [ ]:

In [ ]:

```
```