19BCE2311 Gaurav Singh

Question:

Write a python program to

Find out the list of common and unique terms in the between the three text files attached (chess.txt, tennis.txt, cricket.txt) and print its count.

Apply stopword removal on those common and unique terms (using spacy), print its count and save the terms (after stopword removal) in index.txt file.

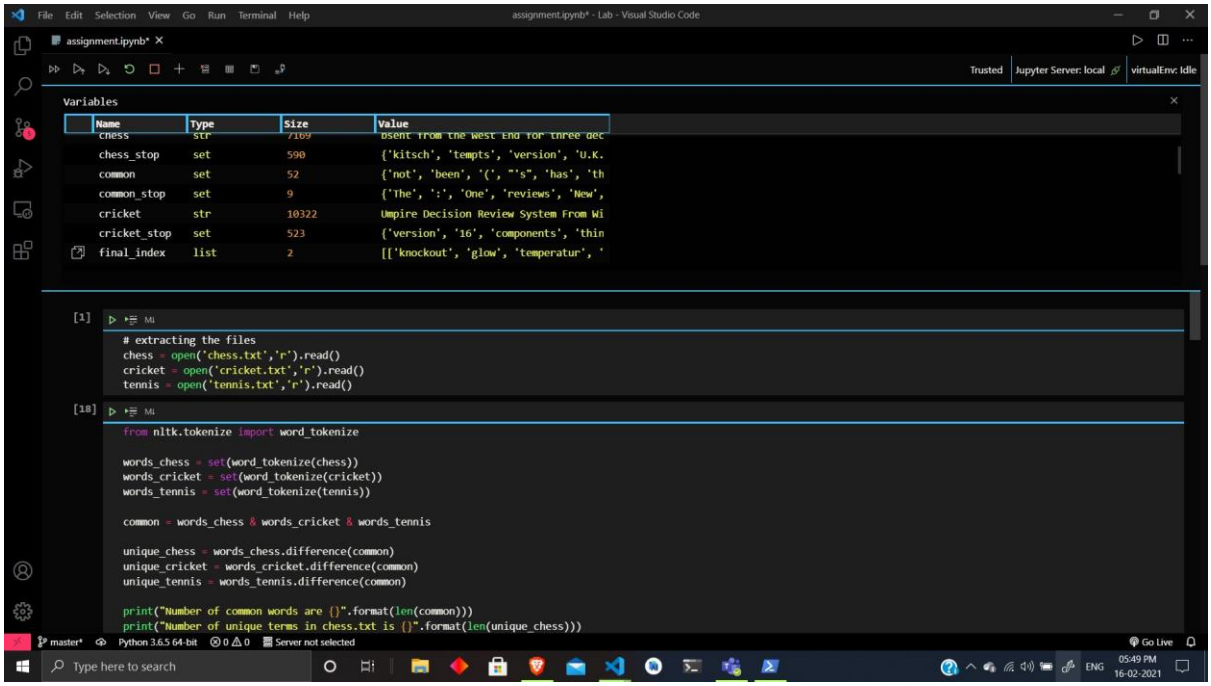
[List of additional Stop words to be considered = [dot, comma, single-quote, double quote, question mark, brackets [square, parentheses, curly, angle], exclamation mark]]

Apply Stemming and lemmatization on the terms present in index.txt file. Print the count of terms after applying stemming and lemmatization.

Replace the content of index.txt file by either stemmed terms or lemmatized term depending on its count.(lower count value should be considered for replacement) and rename the file name to final-index.txt

Print the POS tag of all the terms present in the final-index.txt file using pandas dataframe

Screenshots:



assignment.ipynb* - Lab - Visual Studio Code

[1]

```python
# extracting the files
chess = open('chess.txt','r').read()
cricket = open('cricket.txt','r').read()
tennis = open('tennis.txt','r').read()
```

[18]

```python
from nltk.tokenize import word_tokenize

words_chess = set(word_tokenize(chess))
words_cricket = set(word_tokenize(cricket))
words_tennis = set(word_tokenize(tennis))

common = words_chess & words_cricket & words_tennis

unique_chess = words_chess.difference(common)
unique_cricket = words_cricket.difference(common)
unique_tennis = words_tennis.difference(common)

print("Number of common words are {}".format(len(common)))
print("Number of unique terms in chess.txt is {}".format(len(unique_chess)))
print("Number of unique terms in cricket.txt is {}".format(len(unique_cricket)))
print("Number of unique terms in tennis.txt is {}".format(len(unique_tennis)))
```

```
Number of common words are 52
Number of unique terms in chess.txt is 634
Number of unique terms in cricket.txt is 592
Number of unique terms in tennis.txt is 483
```

[26]

```python
from spacy.lang.en.stop_words import STOP_WORDS
stop_words = set(STOP_WORDS)
stop_words = stop_words.union({'.',',','\'','\"','?','{','}','[',']','<','>','(',')','!'})

common_stop = common.difference(stop_words)
```

assignment.ipynb* - Lab - Visual Studio Code

[27]

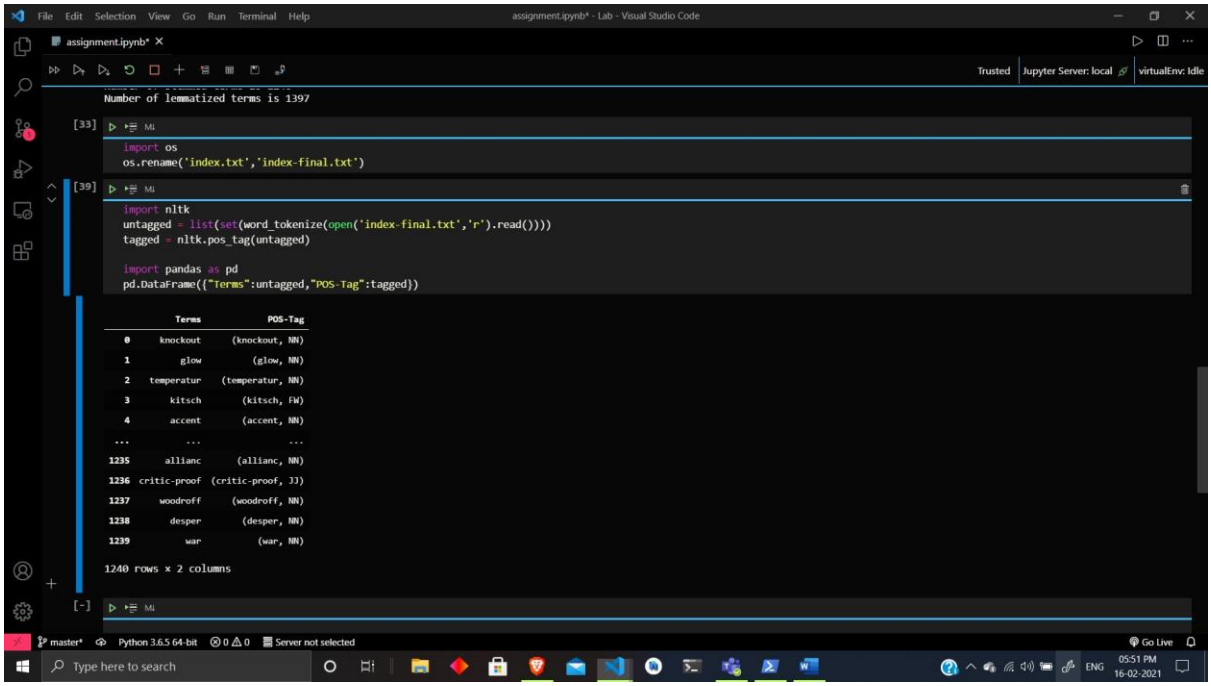
```python
values = [common_stop,chess_stop,cricket_stop,tennis_stop]
with open("index.txt", "w") as output:
    for row in values:
            s = " ".join(map(str, row))
            output.write(s+'\n')
```

[32]

```python
index = list(set(word_tokenize(open('index.txt','r').read())))

from nltk.stem.snowball import SnowballStemmer
from nltk.stem import WordNetLemmatizer

lemmatizer = WordNetLemmatizer()
stemmer = SnowballStemmer(language='english')

stemmed = list(set([stemmer.stem(w) for w in index]))
lemmatized = list(set([lemmatizer.lemmatize(w) for w in index]))

print("Number of stemmed terms is {}".format(len(stemmed)))
print("Number of lemmatized terms is {}".format(len(lemmatized)))

final_index = [stemmed,lemmatized]
if len(stemmed)<len(lemmatized):
    with open("index.txt", "w") as output:
        s = " ".join(map(str, stemmed))
        output.write(s+'\n')
else:
    with open("index.txt", "w") as output:
        s = " ".join(map(str, lemmatized))
        output.write(s+'\n')
```

```
Number of stemmed terms is 1240
Number of lemmatized terms is 1397
```

[33]



Code:

```
chess = open('chess.txt','r').read()
cricket = open('cricket.txt','r').read()
tennis = open('tennis.txt','r').read()

from nltk.tokenize import word_tokenize

words_chess = set(word_tokenize(chess))
words_cricket = set(word_tokenize(cricket))
words_tennis = set(word_tokenize(tennis))

common = words_chess & words_cricket & words_tennis

unique_chess = words_chess.difference(common)
unique_cricket = words_cricket.difference(common)
unique_tennis = words_tennis.difference(common)

print("Number of common words are {}".format(len(common)))
print("Number of unique terms in chess.txt is {}".format(len(unique_chess)))
print("Number of unique terms in cricket.txt is {}".format(len(unique_cricket)
))
print("Number of unique terms in tennis.txt is {}".format(len(unique_tennis)))
```

```
from spacy.lang.en.stop_words import STOP_WORDS
```

```
stop_words = set(STOP_WORDS)
stop_words = stop_words.union({'.',',','\'','\"','?','{','}','[',']','<','>','
(',')','!'})

common_stop = common.difference(stop_words)
chess_stop = unique_chess.difference(stop_words)
cricket_stop = unique_cricket.difference(stop_words)
tennis_stop = unique_tennis.difference(stop_words)
print("Number of terms common after removal of stop words = {}".format(len(com
mon_stop)))
print("Number of terms unique in chess.txt after removal of stop words = {}".f
ormat(len(chess_stop)))
print("Number of terms unique in cricket.txt after removal of stop words = {}"
.format(len(cricket_stop)))
print("Number of terms unique in tennis.txt after removal of stop words = {}".
format(len(tennis_stop)))
```

```
values = [common_stop,chess_stop,cricket_stop,tennis_stop]
with open("index.txt", "w") as output:
    for row in values:
            s = " ".join(map(str, row))
            output.write(s+'\n')


index = list(set(word_tokenize(open('index.txt','r').read())))

from nltk.stem.snowball import SnowballStemmer
from nltk.stem import WordNetLemmatizer

lemmatizer = WordNetLemmatizer()
stemmer = SnowballStemmer(language='english')

stemmed = list(set([stemmer.stem(w) for w in index]))
lemmatized = list(set([lemmatizer.lemmatize(w) for w in index]))

print("Number of stemmed terms is {}".format(len(stemmed)))
print("Number of lemmatized terms is {}".format(len(lemmatized)))

final_index = [stemmed,lemmatized]
if len(stemmed)<len(lemmatized):
    with open("index.txt", "w") as output:
        s = " ".join(map(str, stemmed))
        output.write(s+'\n')
else:
    with open("index.txt", "w") as output:
        s = " ".join(map(str, lemmatized))
```

```
        output.write(s+'\n')
```

```
import os
os.rename('index.txt','index-final.txt')
```

```
import nltk
untagged = list(set(word_tokenize(open('index-final.txt','r').read())))
tagged = nltk.pos_tag(untagged)

import pandas as pd
pd.DataFrame({"Terms":untagged,"POS-Tag":tagged})
```