Web mining Assignment 2

19BCE2311 Gaurav Singh

Question:

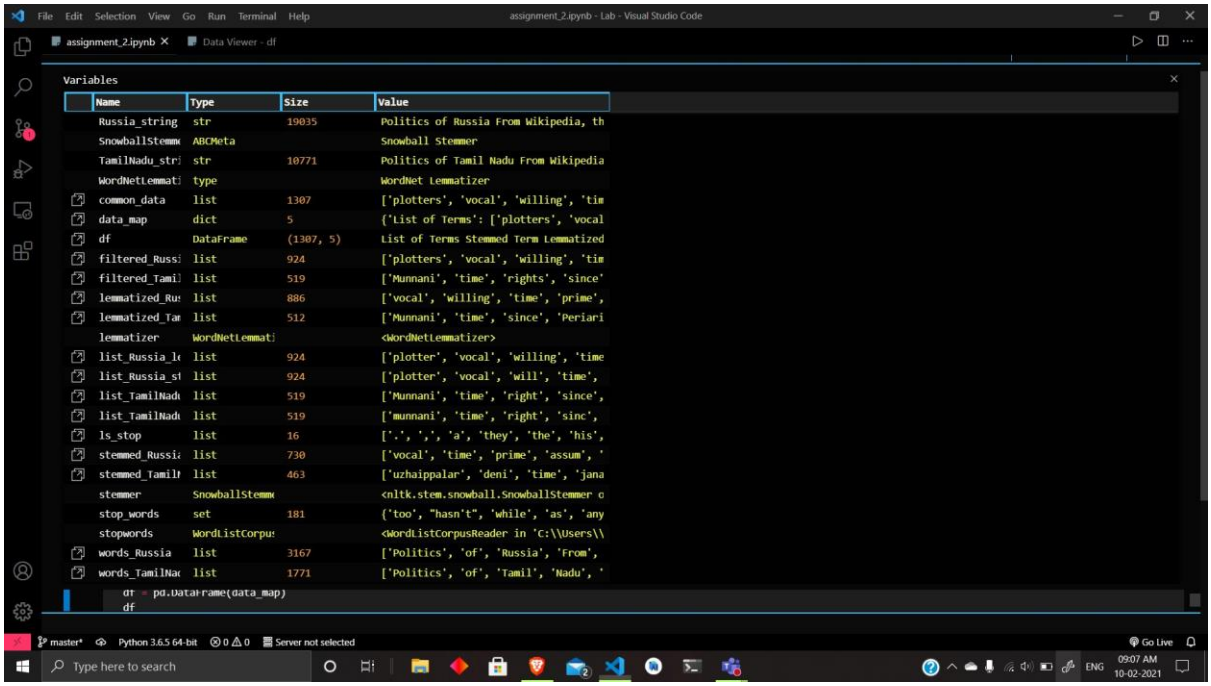
Practice Programming Exercise:
Write a python program to

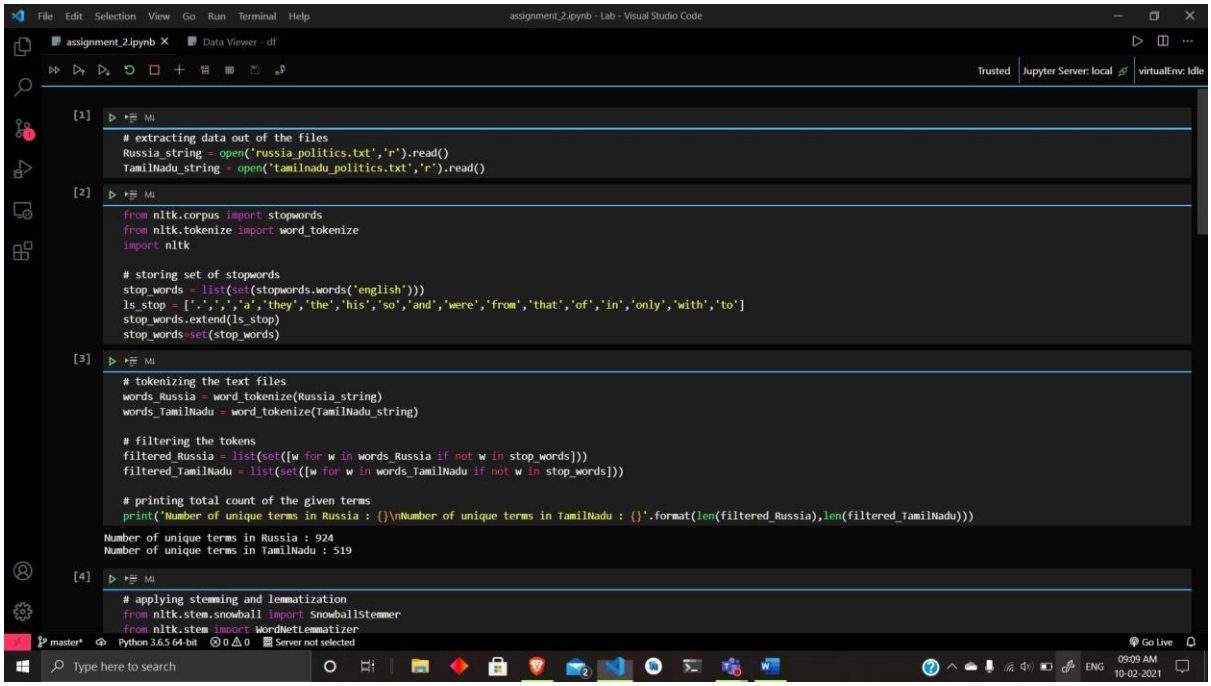
- Extract content from two text files attached
- Count the total number of unique terms in each text file (after removing stop words)
    - [List of additional Stop words to be considered = ['.',',','a','they','the','his','so','and','were','from','that','of','in','only','with','to']]

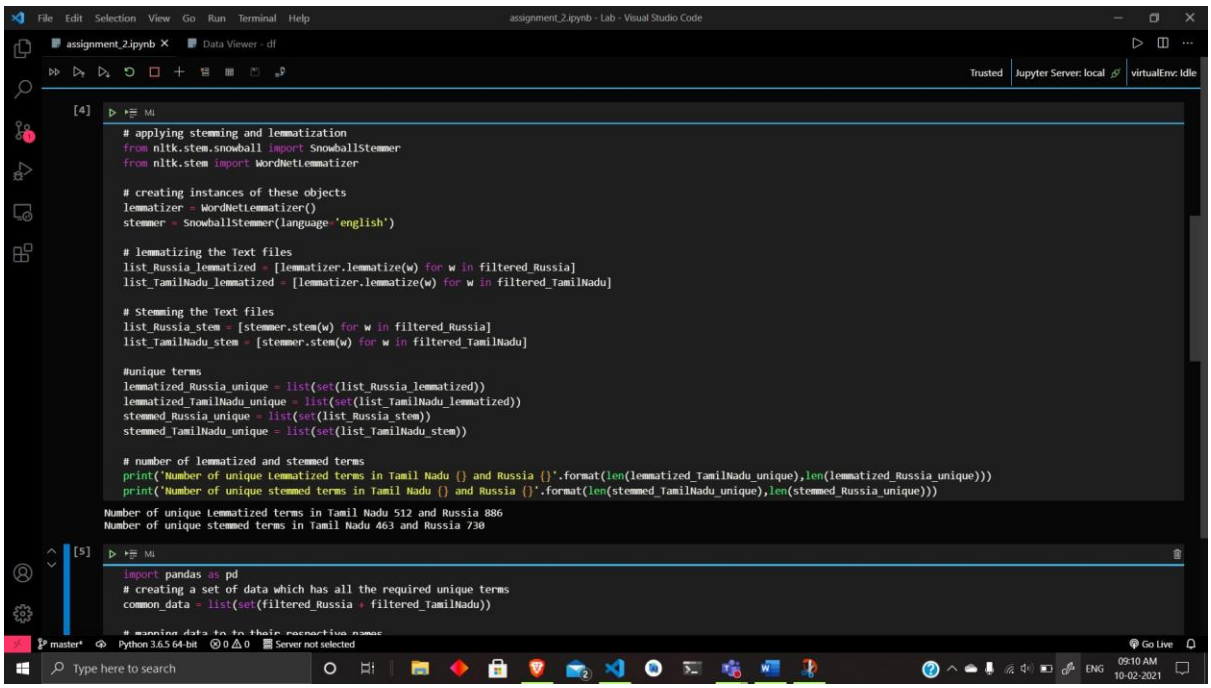
- Apply Stemming and lemmatization separately on the terms present in both files
- Print their number of unique terms after stemming and lemmatization separately.
- Display the result as Term-Document matrix representation using Pandas (use Bag-of-words model)

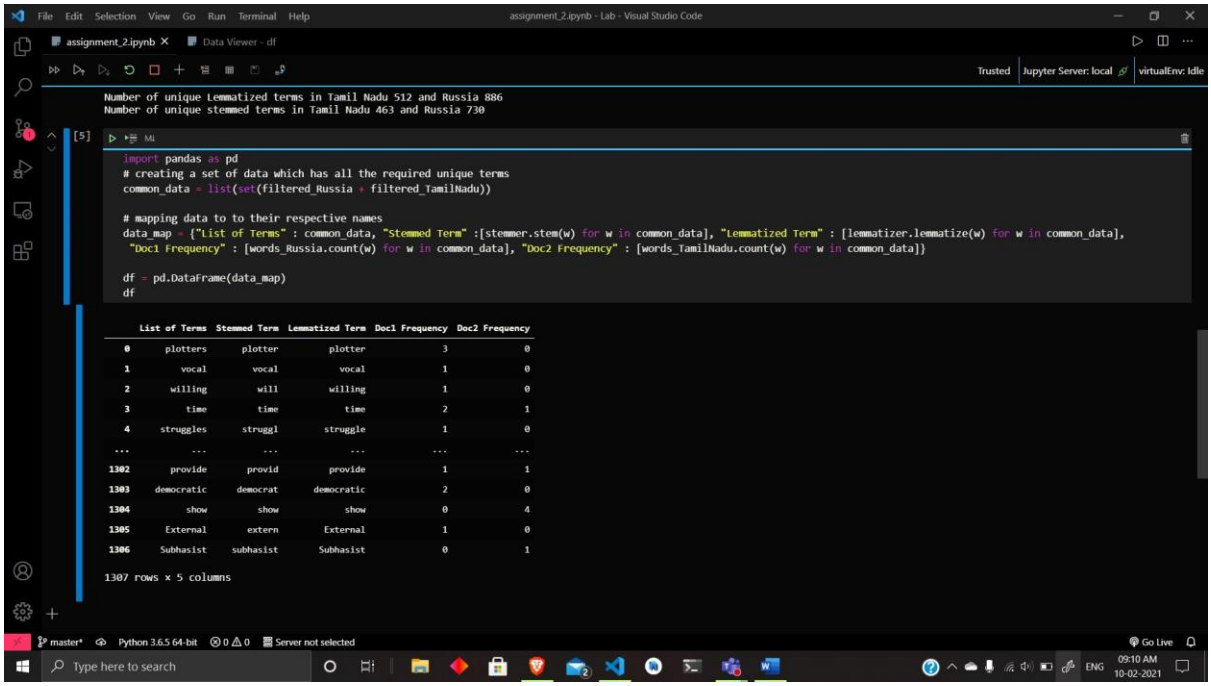
List of Terms	Stemmed Term	Lemmatized Term	DOC1 Frequency	DOC2 Frequency
Programming	Prog	Program	10	0

Screenshots




File Edit Selection View Go Run Terminal Help
assignment_2.ipynb - Lab - Visual Studio Code
assignment_2.ipynb
Data Viewer - df
Trusted
Jupyter Server: local
virtualEnv: Idle
[1]
# extracting data out of the files
Russia_string = open('russia_politics.txt','r').read()
TamilNadu_string = open('tamilnadu_politics.txt','r').read()
[2]
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
import nltk
# storing set of stopwords
stop_words = list(set(stopwords.words('english')))
ls_stop = ['.',',','a','they','the','his','so','and','were','from','that','of','in','only','with','to']
stop_words.extend(ls_stop)
stop_words=set(stop_words)
[3]
# tokenizing the text files
words_Russia = word_tokenize(Russia_string)
words_TamilNadu = word_tokenize(TamilNadu_string)
# filtering the tokens
filtered_Russia = list(set([w for w in words_Russia if not w in stop_words]))
filtered_TamilNadu = list(set([w for w in words_TamilNadu if not w in stop_words]))
# printing total count of the given terms
print('Number of unique terms in Russia : {}\nNumber of unique terms in TamilNadu : {}'.format(len(filtered_Russia),len(filtered_TamilNadu)))
Number of unique terms in Russia : 924
Number of unique terms in TamilNadu : 519
[4]
# applying stemming and lemmatization
from nltk.stem.snowball import SnowballStemmer
from nltk.stem import WordNetLemmatizer
master*
Python 3.6.5 64-bit
Server not selected
Go Live
Type here to search
ENG
09:09 AM
10-02-2021


File Edit Selection View Go Run Terminal Help
assignment_2.ipynb - Lab - Visual Studio Code
assignment_2.ipynb
Data Viewer - df
Trusted
Jupyter Server: local
virtualEnv: Idle
[4]
# applying stemming and lemmatization
from nltk.stem.snowball import SnowballStemmer
from nltk.stem import WordNetLemmatizer
# creating instances of these objects
lemmatizer = WordNetLemmatizer()
stemmer = SnowballStemmer(language='english')
# lemmatizing the Text files
list_Russia_lemmatized = [lemmatizer.lemmatize(w) for w in filtered_Russia]
list_TamilNadu_lemmatized = [lemmatizer.lemmatize(w) for w in filtered_TamilNadu]
# Stemming the Text files
list_Russia_stem = [stemmer.stem(w) for w in filtered_Russia]
list_TamilNadu_stem = [stemmer.stem(w) for w in filtered_TamilNadu]
#unique terms
lemmatized_Russia_unique = list(set(list_Russia_lemmatized))
lemmatized_TamilNadu_unique = list(set(list_TamilNadu_lemmatized))
stemmed_Russia_unique = list(set(list_Russia_stem))
stemmed_TamilNadu_unique = list(set(list_TamilNadu_stem))
# number of lemmatized and stemmed terms
print('Number of unique Lemmatized terms in Tamil Nadu {} and Russia {}'.format(len(lemmatized_TamilNadu_unique),len(lemmatized_Russia_unique)))
print('Number of unique stemmed terms in Tamil Nadu {} and Russia {}'.format(len(stemmed_TamilNadu_unique),len(stemmed_Russia_unique)))
Number of unique Lemmatized terms in Tamil Nadu 512 and Russia 886
Number of unique stemmed terms in Tamil Nadu 463 and Russia 730
[5]
import pandas as pd
# creating a set of data which has all the required unique terms
common_data = list(set(filtered_Russia + filtered_TamilNadu))
master*
Python 3.6.5 64-bit
Server not selected
Go Live
Type here to search
ENG
09:10 AM
10-02-2021



Number of unique Lemmatized terms in Tamil Nadu 512 and Russia 886
Number of unique stemmed terms in Tamil Nadu 463 and Russia 730

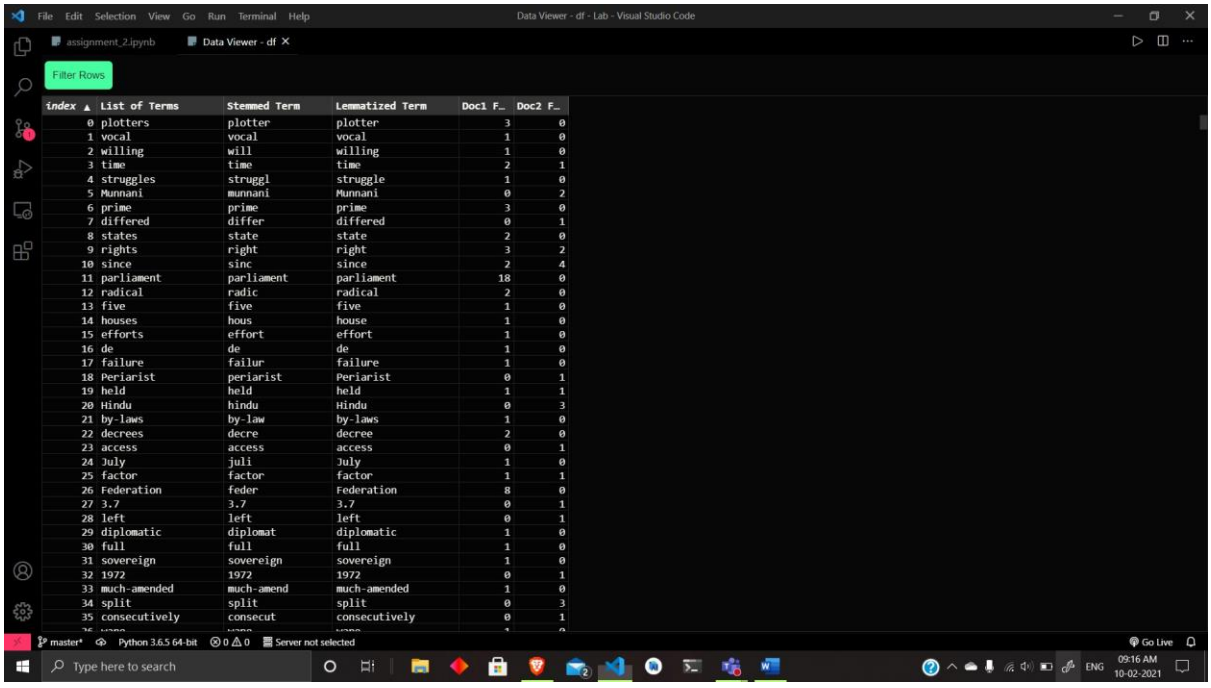
```
[5]
import pandas as pd
# creating a set of data which has all the required unique terms
common_data = list(set(filtered_Russia + filtered_TamilNadu))

# mapping data to to their respective names
data_map = {"List of Terms" : common_data, "Stemmed Term" :[stemmer.stem(w) for w in common_data], "Lemmatized Term" : [lemmatizer.lemmatize(w) for w in common_data],
 "Doc1 Frequency" : [words_Russia.count(w) for w in common_data], "Doc2 Frequency" : [words_TamilNadu.count(w) for w in common_data]}

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

	List of Terms	Stemmed Term	Lemmatized Term	Doc1 Frequency	Doc2 Frequency
0	plotters	plotter	plotter	3	0
1	vocal	vocal	vocal	1	0
2	willing	will	willing	1	0
3	time	time	time	2	1
4	struggles	struggl	struggle	1	0
...	...	...	...	...	...
1302	provide	provid	provide	1	1
1303	democratic	democrat	democratic	2	0
1304	show	show	show	0	4
1305	External	extern	External	1	0
1306	Subhasist	subhasist	Subhasist	0	1

1307 rows x 5 columns



Filter Rows

index ▲	List of Terms	Stemmed Term	Lemmatized Term	Doc1 F…	Doc2 F…
0	plotters	plotter	plotter	3	0
1	vocal	vocal	vocal	1	0
2	willing	will	willing	1	0
3	time	time	time	2	1
4	struggles	struggl	struggle	1	0
5	Munnani	munnani	Munnani	0	2
6	prime	prime	prime	3	0
7	differed	differ	differed	0	1
8	states	state	state	2	0
9	rights	right	right	3	2
10	since	sinc	since	2	4
11	parliament	parliament	parliament	18	0
12	radical	radic	radical	2	0
13	five	five	five	1	0
14	houses	hous	house	1	0
15	efforts	effort	effort	1	0
16	de	de	de	1	0
17	failure	failur	failure	1	0
18	Periarist	periarist	Periarist	0	1
19	held	held	held	1	1
20	Hindu	hindu	Hindu	0	3
21	by-laws	by-law	by-laws	1	0
22	decrees	decre	decree	2	0
23	access	access	access	0	1
24	July	juli	July	1	0
25	factor	factor	factor	1	1
26	Federation	feder	Federation	8	0
27	3.7	3.7	3.7	0	1
28	left	left	left	0	1
29	diplomatic	diplomat	diplomatic	1	0
30	full	full	full	1	0
31	sovereign	sovereign	sovereign	1	0
32	1972	1972	1972	0	1
33	much-amended	much-amend	much-amended	1	0
34	split	split	split	0	3
35	consecutively	consecut	consecutively	0	1

Code:

```
# extracting data out of the files
Russia_string = open('russia_politics.txt','r').read()
TamilNadu_string = open('tamilnadu_politics.txt','r').read()
```

```
from nltk.corpus import stopwords
```

```
from nltk.tokenize import word_tokenize
import nltk

# storing set of stopwords
stop_words = list(set(stopwords.words('english')))
ls_stop = ['.',',','a','they','the','his','so','and','were','from','that','of','in
','only','with','to']
stop_words.extend(ls_stop)
stop_words=set(stop_words)
```

```
# tokenizing the text files
words_Russia = word_tokenize(Russia_string)
words_TamilNadu = word_tokenize(TamilNadu_string)

# filtering the tokens
filtered_Russia = list(set([w for w in words_Russia if not w in stop_words]))
filtered_TamilNadu = list(set([w for w in words_TamilNadu if not w in stop_words])
)

# printing total count of the given terms
print('Number of unique terms in Russia : {}\nNumber of unique terms in TamilNadu
: {}'.format(len(filtered_Russia),len(filtered_TamilNadu)))
```

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

# creating instances of these objects
lemmatizer = WordNetLemmatizer()
stemmer = SnowballStemmer(language='english')

# lemmatizing the Text files
list_Russia_lemmatized = [lemmatizer.lemmatize(w) for w in filtered_Russia]
list_TamilNadu_lemmatized = [lemmatizer.lemmatize(w) for w in filtered_TamilNadu]

# Stemming the Text files
list_Russia_stem = [stemmer.stem(w) for w in filtered_Russia]
list_TamilNadu_stem = [stemmer.stem(w) for w in filtered_TamilNadu]

#unique terms
lemmatized_Russia_unique = list(set(list_Russia_lemmatized))
lemmatized_TamilNadu_unique = list(set(list_TamilNadu_lemmatized))
stemmed_Russia_unique = list(set(list_Russia_stem))
stemmed_TamilNadu_unique = list(set(list_TamilNadu_stem))

# number of lemmatized and stemmed terms
print('Number of unique Lemmatized terms in Tamil Nadu {} and Russia {}'.format(le
n(lemmatized_TamilNadu_unique),len(lemmatized_Russia_unique)))
```

```
print('Number of unique stemmed terms in Tamil Nadu {} and Russia {}'.format(len(s
temmed_TamilNadu_unique),len(stemmed_Russia_unique)))
```

```
import pandas as pd
# creating a set of data which has all the required unique terms
common_data = list(set(filtered_Russia + filtered_TamilNadu))

# mapping data to to their respective names
data_map = {"List of Terms" : common_data, "Stemmed Term" :[stemmer.stem(w) fo
r w in common_data], "Lemmatized Term" : [lemmatizer.lemmatize(w) for w in com
mon_data], "Doc1 Frequency" : [words_Russia.count(w) for w in common_data], "D
oc2 Frequency" : [words_TamilNadu.count(w) for w in common_data]}

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