GROUP MEMBERS

KUMAR SANATAN

ROLL NO – 2211AI24

CHANDERA RAVI

ROLL NO – 2211AI16

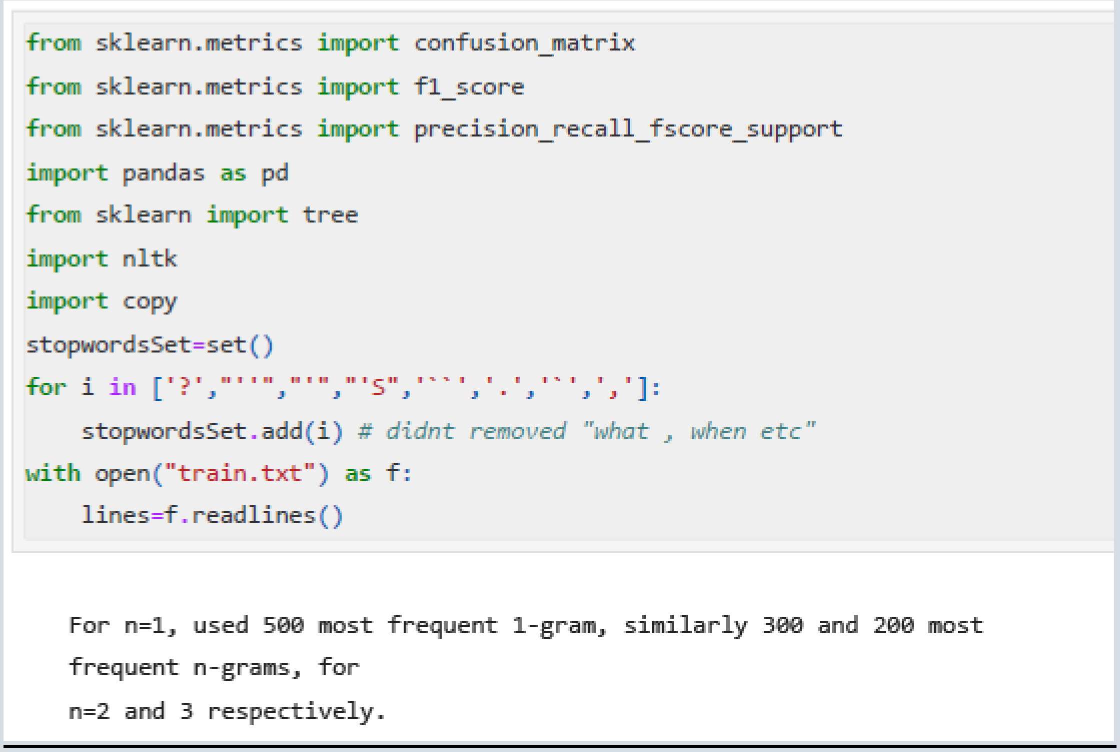
CS 571
ARTIFICIAL INTELLIGENC LAB
ASSIGMENT 9: Decision Tree

INDIAN INSTITUTE OF TECHNOLOGY
PATNA

Date: 15th Nov. 2022 **Deadline:** 29th Nov. 2022

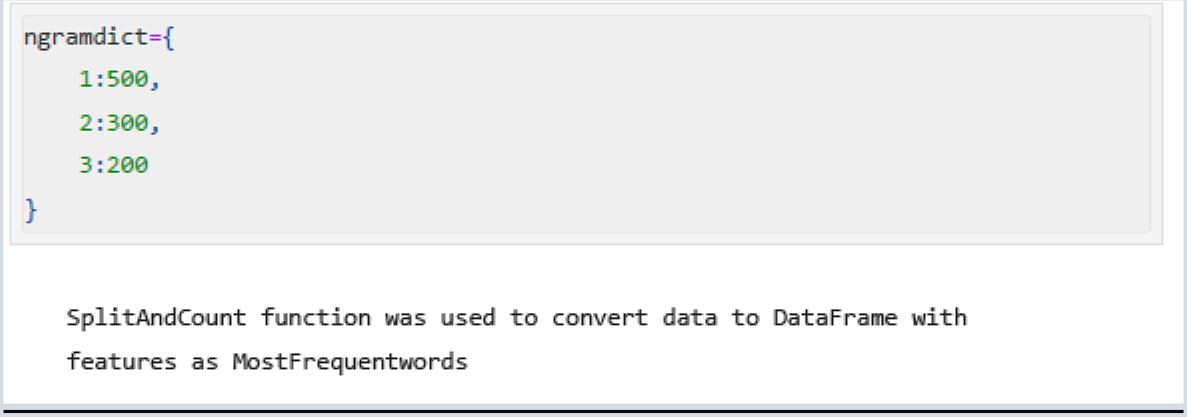
OBJECTIVE

Write a Python program that implements Question classification using Decision Tree classifier.



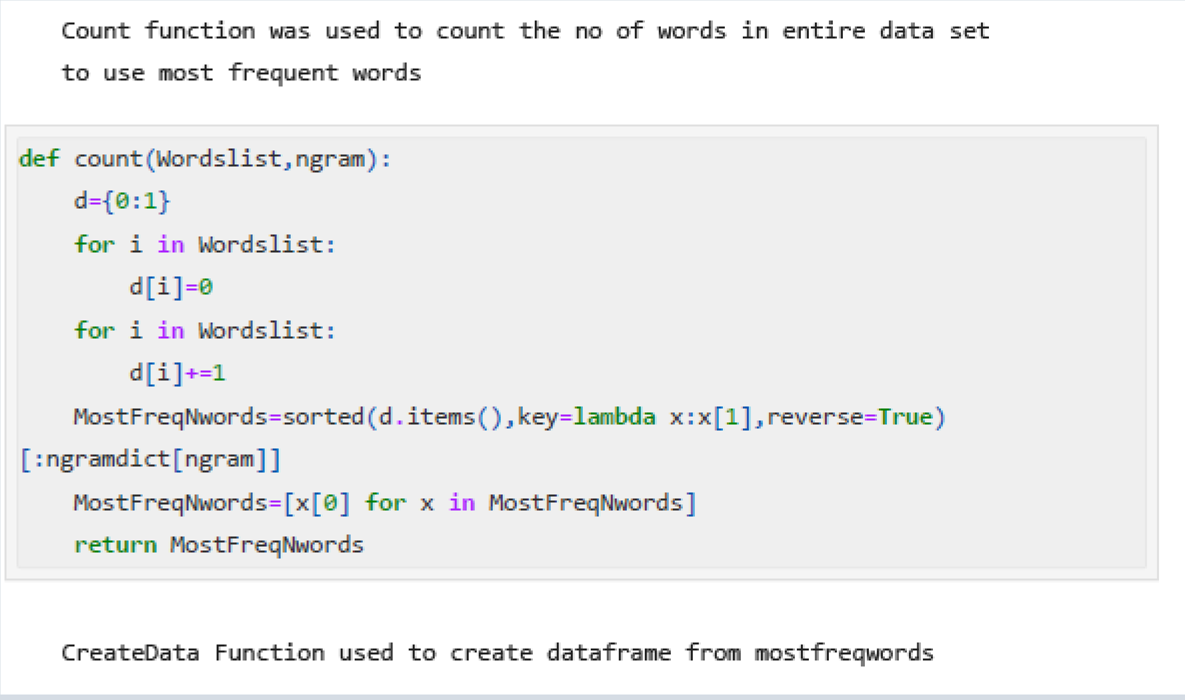
```python
from sklearn.metrics import confusion_matrix
from sklearn.metrics import f1_score
from sklearn.metrics import precision_recall_fscore_support
import pandas as pd
from sklearn import tree
import nltk
import copy
stopwordsSet=set()
for i in ['?',"''","'","'S",'``','.','`',',']:
    stopwordsSet.add(i) # didnt removed "what , when etc"
with open("train.txt") as f:
    lines=f.readlines()
```

For n=1, used 500 most frequent 1-gram, similarly 300 and 200 most frequent n-grams, for
n=2 and 3 respectively.



```python
ngramdict={
    1:500,
    2:300,
    3:200
}
```

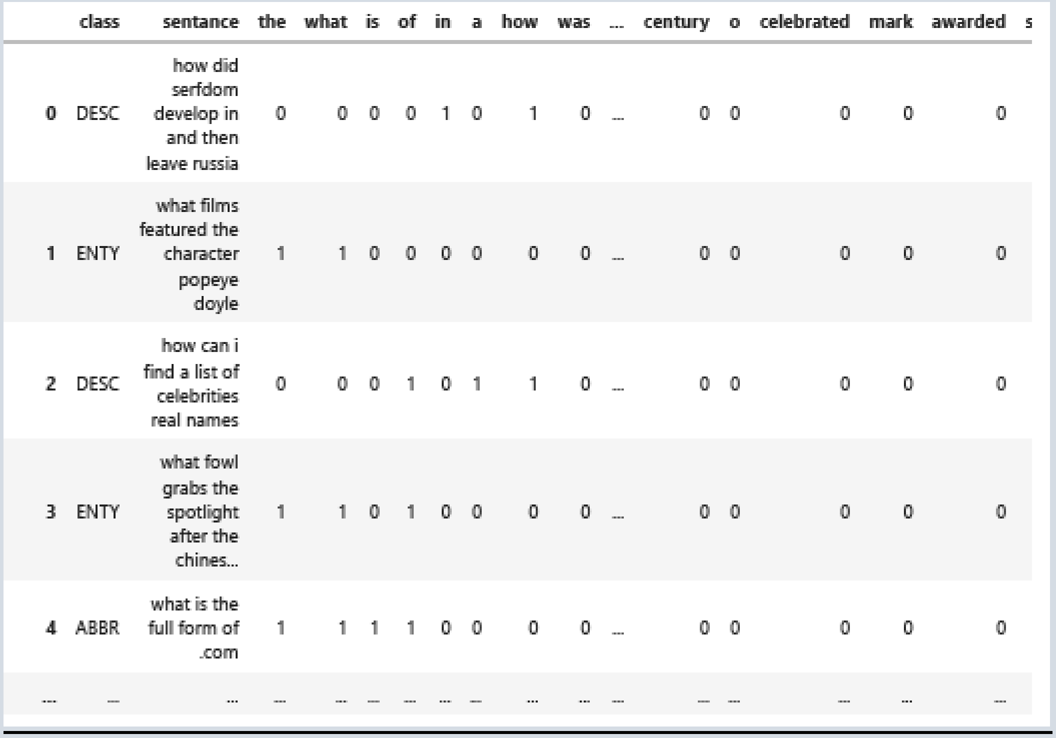
SplitAndCount function was used to convert data to DataFrame with features as MostFrequentwords



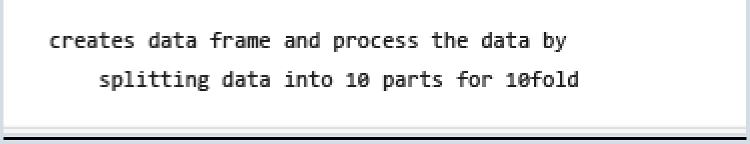
Count function was used to count the no of words in entire data set to use most frequent words

```
def count(Wordslist,ngram):
    d={0:1}
    for i in Wordslist:
        d[i]=0
    for i in Wordslist:
        d[i]+=1
    MostFreqNwords=sorted(d.items(),key=lambda x:x[1],reverse=True)
[:ngramdict[ngram]]
    MostFreqNwords=[x[0] for x in MostFreqNwords]
    return MostFreqNwords
```

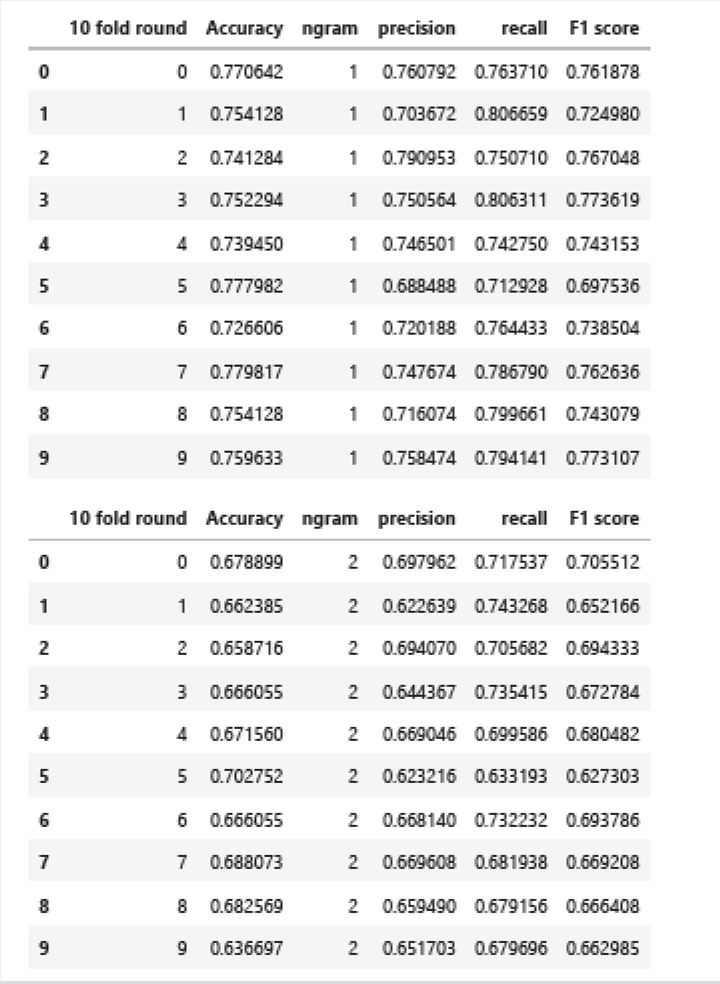
CreateData Function used to create dataframe from mostfreqwords



	class	sentance	the	what	is	of	in	a	how	was	...	century	o	celebrated	mark	awarded	s
0	DESC	how did serfdom develop in and then leave russia	0	0	0	0	1	0	1	0	...	0	0	0	0	0	
1	ENTY	what films featured the character popeye doyle	1	1	0	0	0	0	0	0	...	0	0	0	0	0	
2	DESC	how can i find a list of celebrities real names	0	0	0	1	0	1	1	0	...	0	0	0	0	0	
3	ENTY	what fowl grabs the spotlight after the chines...	1	1	0	1	0	0	0	0	...	0	0	0	0	0	
4	ABBR	what is the full form of .com	1	1	1	1	0	0	0	0	...	0	0	0	0	0	
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	

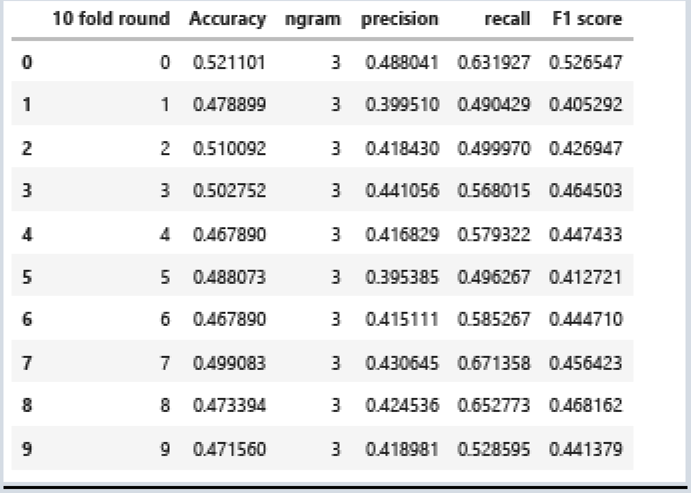


creates data frame and process the data by
    splitting data into 10 parts for 10fold

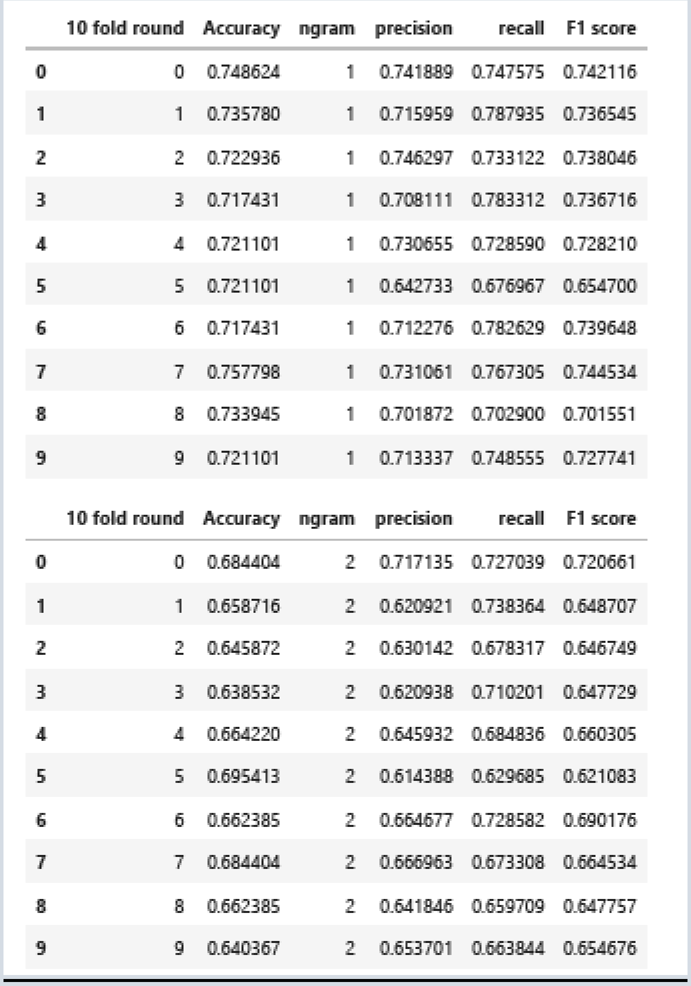


	10 fold round	Accuracy	ngram	precision	recall	F1 score
0	0	0.770642	1	0.760792	0.763710	0.761878
1	1	0.754128	1	0.703672	0.806659	0.724980
2	2	0.741284	1	0.790953	0.750710	0.767048
3	3	0.752294	1	0.750564	0.806311	0.773619
4	4	0.739450	1	0.746501	0.742750	0.743153
5	5	0.777982	1	0.688488	0.712928	0.697536
6	6	0.726606	1	0.720188	0.764433	0.738504
7	7	0.779817	1	0.747674	0.786790	0.762636
8	8	0.754128	1	0.716074	0.799661	0.743079
9	9	0.759633	1	0.758474	0.794141	0.773107

	10 fold round	Accuracy	ngram	precision	recall	F1 score
0	0	0.678899	2	0.697962	0.717537	0.705512
1	1	0.662385	2	0.622639	0.743268	0.652166
2	2	0.658716	2	0.694070	0.705682	0.694333
3	3	0.666055	2	0.644367	0.735415	0.672784
4	4	0.671560	2	0.669046	0.699586	0.680482
5	5	0.702752	2	0.623216	0.633193	0.627303
6	6	0.666055	2	0.668140	0.732232	0.693786
7	7	0.688073	2	0.669608	0.681938	0.669208
8	8	0.682569	2	0.659490	0.679156	0.666408
9	9	0.636697	2	0.651703	0.679696	0.662985

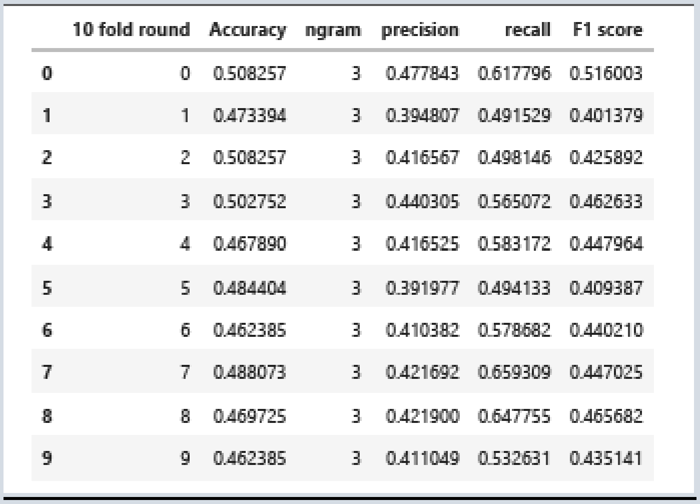


	10 fold round	Accuracy	ngram	precision	recall	F1 score
0	0	0.521101	3	0.488041	0.631927	0.526547
1	1	0.478899	3	0.399510	0.490429	0.405292
2	2	0.510092	3	0.418430	0.499970	0.426947
3	3	0.502752	3	0.441056	0.568015	0.464503
4	4	0.467890	3	0.416829	0.579322	0.447433
5	5	0.488073	3	0.395385	0.496267	0.412721
6	6	0.467890	3	0.415111	0.585267	0.444710
7	7	0.499083	3	0.430645	0.671358	0.456423
8	8	0.473394	3	0.424536	0.652773	0.468162
9	9	0.471560	3	0.418981	0.528595	0.441379

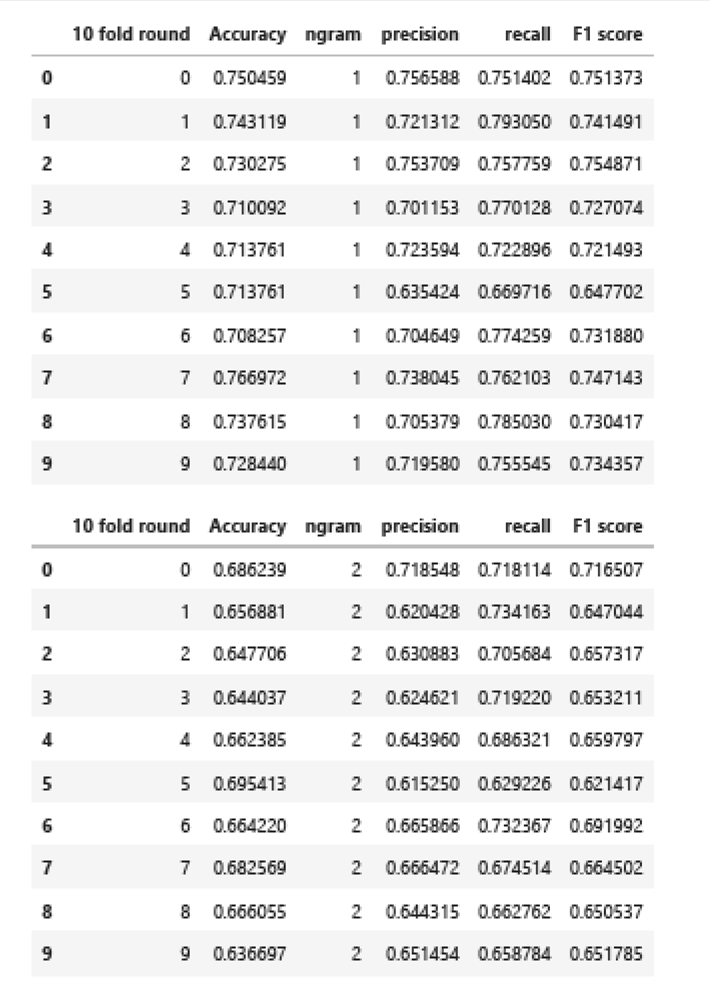


	10 fold round	Accuracy	ngram	precision	recall	F1 score
0	0	0.748624	1	0.741889	0.747575	0.742116
1	1	0.735780	1	0.715959	0.787935	0.736545
2	2	0.722936	1	0.746297	0.733122	0.738046
3	3	0.717431	1	0.708111	0.783312	0.736716
4	4	0.721101	1	0.730655	0.728590	0.728210
5	5	0.721101	1	0.642733	0.676967	0.654700
6	6	0.717431	1	0.712276	0.782629	0.739648
7	7	0.757798	1	0.731061	0.767305	0.744534
8	8	0.733945	1	0.701872	0.702900	0.701551
9	9	0.721101	1	0.713337	0.748555	0.727741

	10 fold round	Accuracy	ngram	precision	recall	F1 score
0	0	0.684404	2	0.717135	0.727039	0.720661
1	1	0.658716	2	0.620921	0.738364	0.648707
2	2	0.645872	2	0.630142	0.678317	0.646749
3	3	0.638532	2	0.620938	0.710201	0.647729
4	4	0.664220	2	0.645932	0.684836	0.660305
5	5	0.695413	2	0.614388	0.629685	0.621083
6	6	0.662385	2	0.664677	0.728582	0.690176
7	7	0.684404	2	0.666963	0.673308	0.664534
8	8	0.662385	2	0.641846	0.659709	0.647757
9	9	0.640367	2	0.653701	0.663844	0.654676

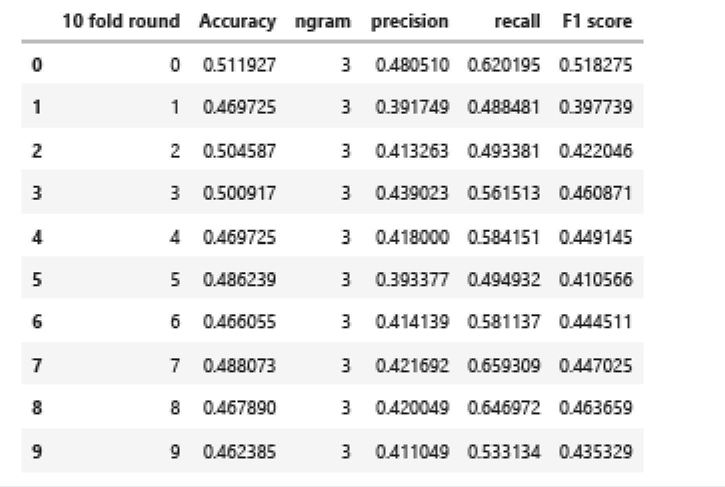


	10 fold round	Accuracy	ngram	precision	recall	F1 score
0	0	0.508257	3	0.477843	0.617796	0.516003
1	1	0.473394	3	0.394807	0.491529	0.401379
2	2	0.508257	3	0.416567	0.498146	0.425892
3	3	0.502752	3	0.440305	0.565072	0.462633
4	4	0.467890	3	0.416525	0.583172	0.447964
5	5	0.484404	3	0.391977	0.494133	0.409387
6	6	0.462385	3	0.410382	0.578682	0.440210
7	7	0.488073	3	0.421692	0.659309	0.447025
8	8	0.469725	3	0.421900	0.647755	0.465682
9	9	0.462385	3	0.411049	0.532631	0.435141

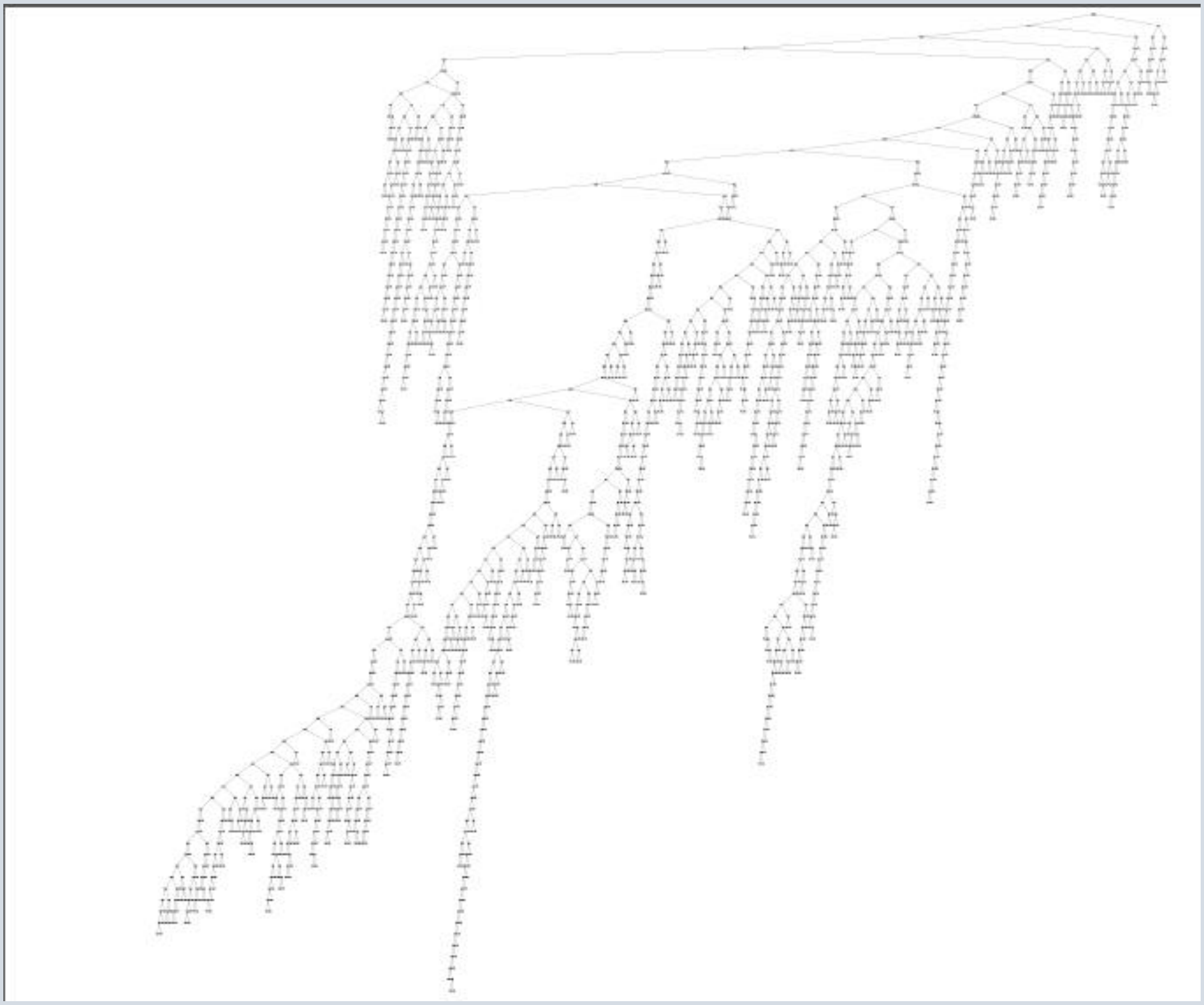


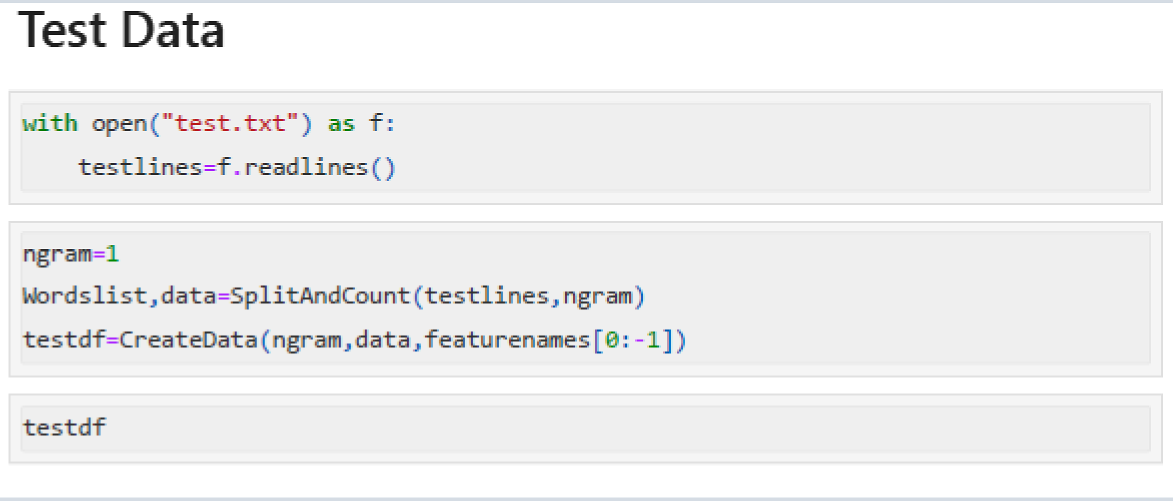
	10 fold round	Accuracy	ngram	precision	recall	F1 score
0	0	0.750459	1	0.756588	0.751402	0.751373
1	1	0.743119	1	0.721312	0.793050	0.741491
2	2	0.730275	1	0.753709	0.757759	0.754871
3	3	0.710092	1	0.701153	0.770128	0.727074
4	4	0.713761	1	0.723594	0.722896	0.721493
5	5	0.713761	1	0.635424	0.669716	0.647702
6	6	0.708257	1	0.704649	0.774259	0.731880
7	7	0.766972	1	0.738045	0.762103	0.747143
8	8	0.737615	1	0.705379	0.785030	0.730417
9	9	0.728440	1	0.719580	0.755545	0.734357

	10 fold round	Accuracy	ngram	precision	recall	F1 score
0	0	0.686239	2	0.718548	0.718114	0.716507
1	1	0.656881	2	0.620428	0.734163	0.647044
2	2	0.647706	2	0.630883	0.705684	0.657317
3	3	0.644037	2	0.624621	0.719220	0.653211
4	4	0.662385	2	0.643960	0.686321	0.659797
5	5	0.695413	2	0.615250	0.629226	0.621417
6	6	0.664220	2	0.665866	0.732367	0.691992
7	7	0.682569	2	0.666472	0.674514	0.664502
8	8	0.666055	2	0.644315	0.662762	0.650537
9	9	0.636697	2	0.651454	0.658784	0.651785



	10 fold round	Accuracy	ngram	precision	recall	F1 score
0	0	0.511927	3	0.480510	0.620195	0.518275
1	1	0.469725	3	0.391749	0.488481	0.397739
2	2	0.504587	3	0.413263	0.493381	0.422046
3	3	0.500917	3	0.439023	0.561513	0.460871
4	4	0.469725	3	0.418000	0.584151	0.449145
5	5	0.486239	3	0.393377	0.494932	0.410566
6	6	0.466055	3	0.414139	0.581137	0.444511
7	7	0.488073	3	0.421692	0.659309	0.447025
8	8	0.467890	3	0.420049	0.646972	0.463659
9	9	0.462385	3	0.411049	0.533134	0.435329



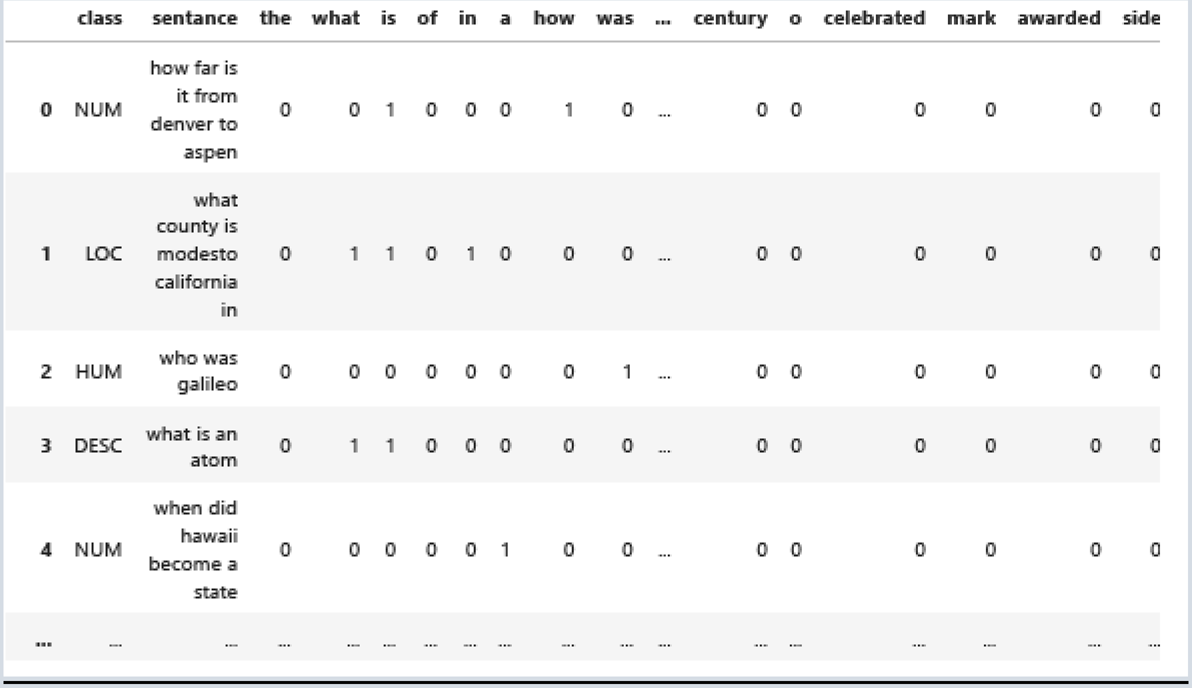


Test Data

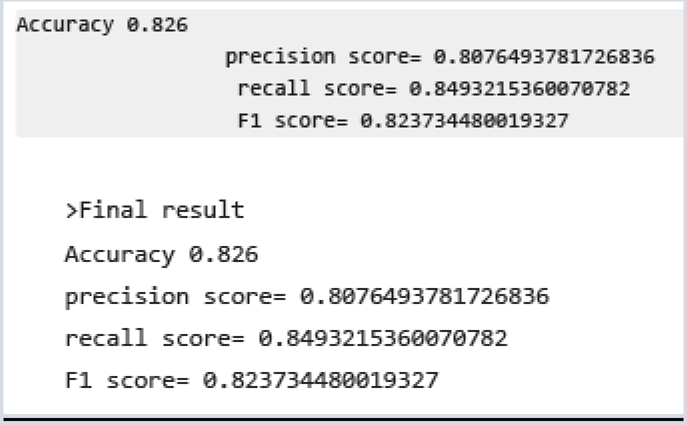
```
with open("test.txt") as f:
    testlines=f.readlines()
```

```
ngram=1
Wordslist,data=SplitAndCount(testlines,ngram)
testdf=CreateData(ngram,data,featurenames[0:-1])
```

```
testdf
```

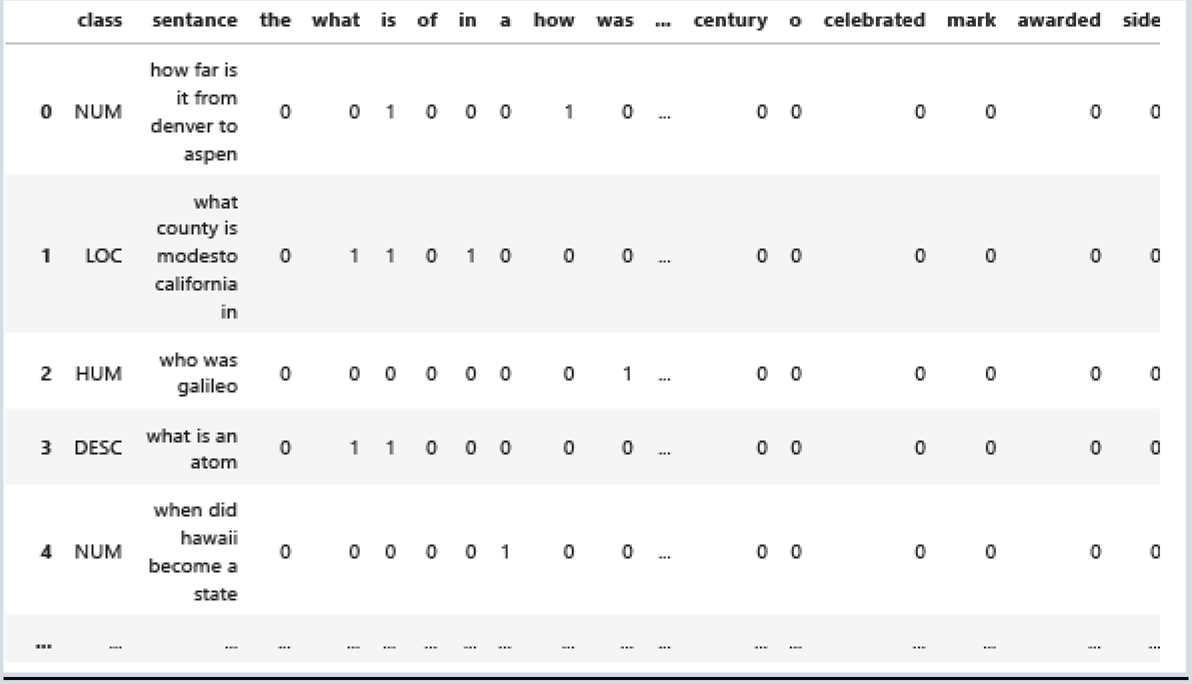


	class	sentance	the	what	is	of	in	a	how	was	...	century	o	celebrated	mark	awarded	side
0	NUM	how far is it from denver to aspen	0	0	1	0	0	0	1	0	...	0	0	0	0	0	0
1	LOC	what county is modesto california in	0	1	1	0	1	0	0	0	...	0	0	0	0	0	0
2	HUM	who was galileo	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0
3	DESC	what is an atom	0	1	1	0	0	0	0	0	...	0	0	0	0	0	0
4	NUM	when did hawaii become a state	0	0	0	0	0	1	0	0	...	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...



```
Accuracy 0.826
                 precision score= 0.8076493781726836
                  recall score= 0.8493215360070782
                  F1 score= 0.823734480019327
```

```
>Final result
Accuracy 0.826
precision score= 0.8076493781726836
recall score= 0.8493215360070782
F1 score= 0.823734480019327
```



	class	sentance	the	what	is	of	in	a	how	was	...	century	o	celebrated	mark	awarded	side
0	NUM	how far is it from denver to aspen	0	0	1	0	0	0	1	0	...	0	0	0	0	0	0
1	LOC	what county is modesto california in	0	1	1	0	1	0	0	0	...	0	0	0	0	0	0
2	HUM	who was galileo	0	0	0	0	0	0	0	1	...	0	0	0	0	0	0
3	DESC	what is an atom	0	1	1	0	0	0	0	0	...	0	0	0	0	0	0
4	NUM	when did hawaii become a state	0	0	0	0	0	1	0	0	...	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...