logistic regression

```
from google.colab import drive
drive.mount('/content/gdrive/')
```

```
Mounted at /content/gdrive/
```

```
%cd /content/drive/MyDrive/D341_19CSE453_NLP/Project
```

```
/content/drive/MyDrive/D341_19CSE453_NLP/Project
```

```
import pandas as pd
```

```
import pandas as pd
data = pd.read_csv('Data.csv',encoding = "ISO-8859-1")
```

```
data.head(5)
```

	Date	Label	Top1	Top2	Top3	Top4	Top5	Top6	Top7	Top8
0	2000-01-03	0	A 'hindrance to operations': extracts from the...	Scorecard	Hughes' instant hit buoys Blues	Jack gets his skates on at ice-cold Alex	Chaos as Maracana builds up for United	Depleted Leicester prevail as Elliott spoils E...	Hungry Spurs sense rich pickings	Gunners so wide of an easy target
1	2000-01-04	0	Scorecard	The best lake scene	Leader: German sleaze inquiry	Cheerio, boyo	The main recommendations	Has Cubie killed fees?	Has Cubie killed fees?	Has Cubie killed fees?
2	2000-01-05	0	Coventry caught on counter by Flo	United's rivals on the road to Rio	Thatcher issues defence before trial by video	Police help Smith lay down the law at Everton	Tale of Trautmann bears two more retellings	England on the rack	Pakistan retaliate with call for video of Walsh	Cullinan continues his Cape monopoly
3	2000-01-06	1	Pilgrim knows how to progress	Thatcher facing ban	McIlroy calls for Irish fighting spirit	Leicester bin stadium blueprint	United braced for Mexican wave	Auntie back in fashion, even if the dress look...	Shoaib appeal goes to the top	Hussain hurt by 'shambles' but lays blame on e...
4	2000-01-07	1	Hitches and Horlocks	Beckham off but United survive	Breast cancer screening	Alan Parker	Guardian readers: are you all whingers?	Hollywood Beyond	Ashes and diamonds	Whingers - a formidable minority

```
data.shape
```

```
(4101, 27)
```

```
new_data = pd.DataFrame(index=range(0,data.shape[0]),columns=['text','labels'])
```

```
for row in range(0,data.shape[0]):
  str1 = ' '.join(str(x) for x in data.iloc[row,2:25])
  new_data['text'][row] = str1
```

```
new_data['text'][0]
```

```
'A 'hindrance to operations': extracts from the leaked reports Scorecard Hughes' instant hit buoys Blu
es Jack gets his skates on at ice-cold Alex Chaos as Maracana builds up for United Depleted Leicester
prevail as Elliott spoils Everton's party Hungry Spurs sense rich pickings Gunners so wide of an easy
target Derby raise a glass to Strupar's debut double Southgate strikes, Leeds pay the penalty Hammers
hand Robson a youthful lesson Saints party like it's 1999 Wear wolves have turned into lambs Stump mik
e catches testy Gough's taunt Langer escapes to hit 167 Flintoff injury piles on woe for England Hunte
rs threaten Jospin with new battle of the Somme Kohl's successor drawn into scandal The difference bet
ween men and women Sara Denver, nurse turned solicitor Diana's landmine crusade put Tories in a panic
Yeltsin's resignation caught opposition flat-footed Russian roulette'
```

```
new_data['labels'] = data['Label'].values
```

```
new_data
```

	text	labels
0	A 'hindrance to operations': extracts from the...	0
1	Scorecard The best lake scene Leader: German s...	0
2	Coventry caught on counter by Flo United's riv...	0
3	Pilgrim knows how to progress Thatcher facing ...	1
4	Hitches and Horlocks Beckham off but United su...	1
...	...	...
4096	Barclays and RBS shares suspended from trading...	0
4097	2,500 Scientists To Australia: If You Want To ...	1
4098	Explosion At Airport In Istanbul Yemeni former...	1
4099	Jamaica proposes marijuana dispensers for tour...	1
4100	A 117-year-old woman in Mexico City finally re...	1

4101 rows × 2 columns

```
%run "Copy of preprocess.ipynb"
```

```
[nltk_data] Downloading package punkt to /root/nltk_data...
[nltk_data]   Package punkt is already up-to-date!
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
[nltk_data] Downloading package wordnet to /root/nltk_data...
[nltk_data]   Package wordnet is already up-to-date!
```

```
new_data['clean_text'] = new_data['text'].apply(preprocess)
```

```
new_data.head(10)
```

	text	labels	clean_text
0	A 'hindrance to operations': extracts from the...	0	oper extract leak report scorecard hugh instan...
1	Scorecard The best lake scene Leader: German s...	0	scorecard best lake scene leader german sleaz ...
2	Coventry caught on counter by Flo United's riv...	0	coventri caught counter flo unit rival road ri...
3	Pilgrim knows how to progress Thatcher facing ...	1	pilgrim know progress thatcher face ban mcilro...
4	Hitches and Horlocks Beckham off but United su...	1	hitch horlock beckham unit surviv breast cance...
5	Fifth round draw BBC unveils secret weapon in ...	1	fifth round draw bbc unveil secret weapon rate...
6	Man Utd 2 - 0 South Melbourne How North Atlant...	1	man utd south melbourn north atlant drift coul...
7	Newcastle seek new football supremo Liverpool ...	0	newcastl seek new footbal supremo liverpool ai...
8	Bungling officials on the carpet And in the re...	1	bungl offici carpet red raw corner mackenzi un...
9	Pompey plump for Pulis work ethic Roma under f...	1	pompey plump puli work ethic roma fire rolex r...

```
new_data = new_data.drop(labels=['text'],axis=1)
```

```
new_data
```

	labels	clean_text
0	0	oper extract leak report scorecard hugh instan...
1	0	scorecard best lake scene leader german sleaz ...
2	0	coventri caught counter flo unit rival road ri...
3	1	pilgrim know progress thatcher face ban mcilro...
4	1	hitch horlock beckham unit surviv breast cance...
...	...	...
4096	0	barclay rbs share suspend trade tank pope say ...
4097	1	scientist australia want save great barrier re...
4098	1	explos airport istanbul yemeni former presid t...
4099	1	jamaica propos marijuana dispens tourist airpo...
4100	1	woman mexico citi final receiv birth certif di...

4101 rows × 2 columns

```
from sklearn.feature_extraction.text import CountVectorizer
texts = new_data['clean_text'].values
CountVectorizer = CountVectorizer(ngram_range=(2,3))
X = CountVectorizer.fit_transform(texts)
y = new_data['labels'].values
print(X.shape)
print(y.shape)
```

```
(4101, 1199732)
(4101,)
```

```
from sklearn.model_selection import train_test_split
X_train, X_test, y_train, y_test = train_test_split(X, y, test_size = .20, shuffle = True,random_state = 0)
```

```
from sklearn.linear_model import LogisticRegression
```

```
model = LogisticRegression()
model = model.fit(X_train, y_train)
predictions = model.predict(X_test)
```

```
from sklearn.metrics import classification_report
from sklearn.metrics import f1_score
from sklearn.metrics import accuracy_score
from sklearn.metrics import confusion_matrix

print (classification_report(test["Label"], predictions))
print (accuracy_score(test["Label"], predictions))
```

```
                  precision    recall  f1-score   support

               0       0.89      0.83      0.86       186
               1       0.85      0.90      0.87       192

        accuracy                           0.87       378
       macro avg       0.87      0.86      0.86       378
    weighted avg       0.87      0.87      0.86       378

    0.8650793650793651
```