Roll No :- 20BCE204 Practical No. :- 6

Implement Naive bayes theorm for given data and performe document classification

In [1]:
```
# import data
```

In [2]:
```
import pandas as pd
import random
import string

news = pd.read_table("pr-6.csv", delimiter=',',names=['label', 'message'],e

# news = news.sample()
print(news.shape)
print(news.head(5))
```

```
(50, 2)
    label                                            message
0     Eco   10 stocks from 5 sectors to stay on Ferris wheel
1  sports  Aaron Finch, Wade help Australia to tight T20 ...
2     Eco  Adani Green to launch $1 bn bond by December; ...
3     Eco  Adani, Tata Power plan to raise $1.3 bn in gre...
4     Eco  Any such move will delay investments in fuel p...
```

In [3]:
```
# Preprocessing and divide word dictionary in two classs wise
```

In [4]:

```
from nltk.corpus import stopwords
from sklearn.feature_extraction.text import CountVectorizer
import nltk

cleaning = [char for char in news['message'] if char not in string.punctuat
# print(cleaning)
# print(len(cleaning))
label = news['label']
# print(label[0])
X_train = list(cleaning[:35])
Y_train = list(label[:35])
X_test = list(cleaning[35:])
Y_test = list(label[35:])
print(Y_test)

# this will have train data only
dtext=[]
ps = nltk.stem.porter.PorterStemmer()

for i in X_train:
    dtext.append([ps.stem(word) for word in i.split() if word.lower() not 

print(len(dtext))

# create two least for each class

# print(len(set(label)))
eco=[]
sport=[]

for i in range(len(dtext)):
#     print(label[i])
    if label[i]=='Eco':
        eco.append(dtext[i])
    else:
        sport.append(dtext[i])

print(eco)
print("*********")
print(sport)

# two word set for each class again....

ecoc = []
sportc = []

for i in eco:
    ecoc+=i
for i in sport:
    sportc+=i


print(ecoc)
print(sportc)
```

```
['Eco', 'sports', 'Eco', 'Eco', 'Eco', 'Eco', 'Eco', 'sports', 'Eco', 'Eco'
, 'Eco', 'Eco', 'sports', 'Eco', 'Eco']
35
[['10', 'stock', '5', 'sector', 'stay', 'ferri', 'wheel'], ['adani', 'green
```

', 'launch', '$1', 'bn', 'bond', 'december;', 'tata', 'power', 'rais', '$32
0', 'mn', 'month'], ['adani,', 'tata', 'power', 'plan', 'rais', '$1.3', 'bn
', 'green', 'bond', 'new', 'project'], ['move', 'delay', 'invest', 'fuel',
'production,', 'compani', 'tell', 'govt', 'panel'], ['asian', 'develop', 'b
ank', 'provid', '$2.3-$2.5', 'bn', 'flood-hit', 'pakistan'], ['rs', '3,513'
, 'crore,', 'madhya', 'pradesh', 'bag', 'textil', 'pli', 'invest'], ['colle
g', 'optimist', 'bumper', 'recruit', 'rise', 'pre-plac', 'offer'], ['foreca
st', 'lower', '3.4%', 'recess', 'loom'], ['downsid', 'risk', 'materialise,'
, 'trade', 'growth', '2023', 'could', 'low', '-2.8', 'per', 'cent.'], ['sep
temb', 'anoth', '2.11', 'million', 'added,', 'take', 'total', 'count', '102
.61', 'million.'], ['first', 'nine', 'month', '2022,', 'home-grown', 'compa
ni', 'issu', 'green', 'bond', 'worth', '$1.79', 'billion,', 'contrast', '$4
.9', 'billion', 'rais', 'period', '2021'], ['india’', 'stockpile,', 'exampl
e,', 'tumbl', '$96', 'billion', 'year', '$538', 'billion.'], ['mcx', 'crude
', 'oil', 'pullback', 'rs', '7,700;', 'natur', 'ga', 'like', 'test', 'rs',
'600']]
****
[['aaron', 'finch,', 'wade', 'help', 'australia', 'tight', 't20', 'win', 'w
est', 'indi'], ['come', '89th', 'career', 'titl', 'tel', 'aviv', 'last', 'w
eekend'], ['devin', 'star', 'new', "zealand'", 'super', 'win'], ['devin', '
took', 'deliv', 'new', 'zealand', 'yet', 'anoth', 'seri', 'win', 'west', 'i
ndies.'], ['djokov', 'broke', 'overmatch', 'garin', 'five', 'time', 'contin
u', 'run'], ['erl', 'haaland', 'net', '2', 'man', 'citi', 'rout'], ['guardi
ola', 'deni', 'releas', 'claus'], ['icc', 'rankings:', 'suryakumar', 'slip'
, 'no.', '2', 't20i', 'bat', 'list', '..'], ['icc', 't20', 'world', 'cup',
'2022:', 'india’', 'nitin', 'menon', 'among', '16', 'umpir', 'name', 'mega'
, 'showpiec'], ['icc', 't20', 'world', 'cup:', 'rohit', 'sharma-l', 'team',
'india', 'depart', 'australia'], ['ind', 'vs', 'sa,', '3rd', 't20i:', 'domi
n', 'protea', 'hand', 'india', '49-run', 'defeat'], ['india', 'top', 'wrest
l', 'medal', 'talli', 'thrice', 'last', 'four', 'commonwealth', 'games.'],
['india', 'depart', 'world', 'cup', 'without', '15th', 'player'], ['india',
'disappoint', 'shoot'], ["india'", 'wrestl', 'fratern', 'rue', 'sport’', '2
026', 'cwg', 'axe'], ['indian', 'firm', 'write', 'csa', 'high', 'base', 'pr
ice', 'sa20'], ['ioc', 'consult', 'saudi', 'arabia', 'choic', '2029', 'asia
n', 'winter', 'game'], ['lionel', 'messi', 'save', 'pari', 'saint-germain',
'blush', 'benfica'], ['miss', 'flight', 'cost', 'shimron', 'hetmyer', 'plac
e', 'west', 'indies’', 'icc', 't20', 'world', 'cup', '2022', 'squad'], ['na
tion', 'game', '2022:', 'read', 'gita', 'made', 'calmer,', 'say', 'amlan',
'borgohain'], ['nation', 'game', '2022:', 'uttar', 'pradesh’', 'ram', 'babo
o', 'break', 'nation', 'record', "men'", '35km', 'race', 'walk'], ['novak',
'djokov', 'demolish', 'cristian', 'garin', 'astana', 'first', 'round']]
['10', 'stock', '5', 'sector', 'stay', 'ferri', 'wheel', 'adani', 'green',
'launch', '$1', 'bn', 'bond', 'december;', 'tata', 'power', 'rais', '$320',
'mn', 'month', 'adani,', 'tata', 'power', 'plan', 'rais', '$1.3', 'bn', 'gr
een', 'bond', 'new', 'project', 'move', 'delay', 'invest', 'fuel', 'product
ion,', 'compani', 'tell', 'govt', 'panel', 'asian', 'develop', 'bank', 'pro
vid', '$2.3-$2.5', 'bn', 'flood-hit', 'pakistan', 'rs', '3,513', 'crore,',
'madhya', 'pradesh', 'bag', 'textil', 'pli', 'invest', 'colleg', 'optimist'
, 'bumper', 'recruit', 'rise', 'pre-plac', 'offer', 'forecast', 'lower', '3
.4%', 'recess', 'loom', 'downsid', 'risk', 'materialise,', 'trade', 'growth
', '2023', 'could', 'low', '-2.8', 'per', 'cent.', 'septemb', 'anoth', '2.1
1', 'million', 'added,', 'take', 'total', 'count', '102.61', 'million.', 'f
irst', 'nine', 'month', '2022,', 'home-grown', 'compani', 'issu', 'green',
'bond', 'worth', '$1.79', 'billion,', 'contrast', '$4.9', 'billion', 'rais'
, 'period', '2021', 'india’', 'stockpile,', 'example,', 'tumbl', '$96', 'bi
llion', 'year', '$538', 'billion.', 'mcx', 'crude', 'oil', 'pullback', 'rs'
, '7,700;', 'natur', 'ga', 'like', 'test', 'rs', '600']
['aaron', 'finch,', 'wade', 'help', 'australia', 'tight', 't20', 'win', 'we
st', 'indi', 'come', '89th', 'career', 'titl', 'tel', 'aviv', 'last', 'week
end', 'devin', 'star', 'new', "zealand'", 'super', 'win', 'devin', 'took',
'deliv', 'new', 'zealand', 'yet', 'anoth', 'seri', 'win', 'west', 'indies.'

```
, 'djokov', 'broke', 'overmatch', 'garin', 'five', 'time', 'continu', 'run'
, 'erl', 'haaland', 'net', '2', 'man', 'citi', 'rout', 'guardiola', 'deni',
'releas', 'claus', 'icc', 'rankings:', 'suryakumar', 'slip', 'no.', '2', 't
20i', 'bat', 'list', '..', 'icc', 't20', 'world', 'cup', '2022:', 'india'',
'nitin', 'menon', 'among', '16', 'umpir', 'name', 'mega', 'showpiec', 'icc'
, 't20', 'world', 'cup:', 'rohit', 'sharma-l', 'team', 'india', 'depart', '
australia', 'ind', 'vs', 'sa,', '3rd', 't20i:', 'domin', 'protea', 'hand',
'india', '49-run', 'defeat', 'india', 'top', 'wrestl', 'medal', 'talli', 't
hrice', 'last', 'four', 'commonwealth', 'games.', 'india', 'depart', 'world
', 'cup', 'without', '15th', 'player', 'india', 'disappoint', 'shoot', "ind
ia'", 'wrestl', 'fratern', 'rue', 'sport'', '2026', 'cwg', 'axe', 'indian',
'firm', 'write', 'csa', 'high', 'base', 'price', 'sa20', 'ioc', 'consult',
'saudi', 'arabia', 'choic', '2029', 'asian', 'winter', 'game', 'lionel', 'm
essi', 'save', 'pari', 'saint-germain', 'blush', 'benfica', 'miss', 'flight
', 'cost', 'shimron', 'hetmyer', 'place', 'west', 'indies'', 'icc', 't20',
'world', 'cup', '2022', 'squad', 'nation', 'game', '2022:', 'read', 'gita',
'made', 'calmer,', 'say', 'amlan', 'borgohain', 'nation', 'game', '2022:',
'uttar', 'pradesh'', 'ram', 'baboo', 'break', 'nation', 'record', "men'", '
35km', 'race', 'walk', 'novak', 'djokov', 'demolish', 'cristian', 'garin',
'astana', 'first', 'round']
```

In [5]:
```
# example how test data is going to be performe
```

In [6]:
```
# print(X_test[0])
# print(Y_test[3])


# itw = [ps.stem(word) for word in it.split() if word.lower() not in stopw
# print(itw)
```

In [7]:
```
#  p(class/word) = p(word/class)*p(class)
```

In [8]:
```
pE = len(eco)/ len(eco)+len(sport)
pS = len(sport)/ len(eco)+len(sport)
```

In [9]:
```
# calculate probability for one word to be in one class and compare probab
```

In [10]:

```
# p(word/class) = word in class + 1 / (totalword in class + total word in 

pred = []
for i in range(len(X_test)):
    itw = [ps.stem(word) for word in X_test[i].split() if word.lower() not
    print(itw)

    # word in eco
    x=1
    for i in range(len(itw)):
        x *= ( (ecoc.count(itw[i]) + 1) / (len(ecoc) + len(ecoc)+len(sportc
    x*= pE
#     print(x)

    y=1
    # word in sportc
    for i in range(len(itw)):
        y *= ( (sportc.count(itw[i]) + 1) / (len(sportc) + len(ecoc)+len(sp
    y*= pS
#     print(y)

    if y>x:
        pred.append('sport')
    else:
        pred.append('eco')

print(pred)
print(Y_test)

for i in range(len(pred)):
    if pred[i]=='eco':
        pred[i]=0
    else:
        pred[i]=1
    if Y_test[i]=='Eco':
        Y_test[i]=0
    else:
        Y_test[i]=1

from sklearn.metrics import accuracy_score

accuracy_score(pred,Y_test)
```

```
['100', 'million', 'covid-19', 'vaccin', 'dose', 'wast', 'india', 'septembe
r-end']
['pakistan', 'doesn't', 'virat', 'kohli', 'no.4,', 'babar', 'azam', 'play']
['pakistan'', '$14', 'billion', 'reserv', 'aren't', 'enough', 'cover', 'thr
ee', 'month', 'import']
['reserv', 'declin', '$1', 'trillion,', '7.8%,', 'year', '$12', 'trillion,'
, 'biggest', 'drop', 'sinc', 'bloomberg', 'start', 'compil', 'data', '2003'
]
['ril', 'bat', 'cap', 'domest', 'ga', 'price', 'govt', 'panel', 'seek', 're
view']
['flood', 'caus', 'collect', 'loss', 'usd', '40', 'billion.']
['stock', 'zeel', 'appreci', '13', 'per', 'cent,', 'compar', '3.5', 'per',
'cent', 'rise', 's&p', 'bse', 'sensex']
['12', 'medals,', '6', 'gold,', 'birmingham', 'made', 'top', 'contributor',
"india'", 'medal', 'talli', '61']
['trade', 'volum', 'soar', 'demat', 'talli', 'surpass', '102.5', 'million',
'account']
['trade', 'volum', 'soar', 'demat', 'talli', 'surpass', '102.5', 'million',
'account']
['vnit', 'receiv', '170', 'ppo', 'current', '2022-23', 'batch', 'last', 'ye
ar']
['world', 'currenc', 'reserv', 'shrink', '$1', 'trn', 'yr', 'record', 'draw
down']
['world', 'tabl', 'tenni', 'championship', '2022:', 'indian', 'men', 'enter
', 'pre-quarterfin', 'despit', 'franc', 'loss']
['wto', 'slash', 'global', '2023', 'trade', 'growth', 'forecast', '1%', 're
cess', 'loom']
['zee', 'entertain', 'gain', '6%', "cci'", 'condit', 'nod', 'merger', 'soni
']
['sport', 'eco', 'eco', 'eco', 'eco', 'eco', 'eco', 'sport', 'eco', 'eco',
'eco', 'sport', 'sport', 'eco', 'eco']
['Eco', 'sports', 'Eco', 'Eco', 'Eco', 'Eco', 'Eco', 'sports', 'Eco', 'Eco'
, 'Eco', 'Eco', 'sports', 'Eco', 'Eco']
```

Out[10]:
```
0.8
```

In [ ]:

In [11]:
```
# p(class/word) = p(class/word)*p(class) + (1-p(class/word))*(1-p(class))
```

In [12]:

```
# p(word/class) = word in class + 1 / (totalword in class + total word in 

predb = []
for i in range(len(X_test)):
    itw = [ps.stem(word) for word in X_test[i].split() if word.lower() not
    print(itw)

    # word in eco
    x=1
    for i in range(len(itw)):
        x *= ( (ecoc.count(itw[i]) + 1) / (len(ecoc) + len(ecoc)+len(sportc
    tx = x;
    x*= pE
    x += (tx*(1-pE))

#     print(x)

    y=1
    # word in sportc
    for i in range(len(itw)):
        y *= ( (sportc.count(itw[i]) + 1) / (len(sportc) + len(ecoc)+len(sp
    ty = y;
    y*= pS
    y += (ty*(1-pS))

#     print(y)

    if y>x:
        predb.append('sport')
    else:
        predb.append('eco')

print(predb)
# print(Y_test)

for i in range(len(pred)):
    if predb[i]=='eco':
        predb[i]=0
    else:
        predb[i]=1
#     if Y_test[i]=='Eco':
#         Y_test[i]=0
#     else:
#         Y_test[i]=1

from sklearn.metrics import accuracy_score

accuracy_score(predb,Y_test)
```

```
['100', 'million', 'covid-19', 'vaccin', 'dose', 'wast', 'india', 'septembe
r-end']
['pakistan', 'doesn’t', 'virat', 'kohli', 'no.4,', 'babar', 'azam', 'play']
['pakistan’', '$14', 'billion', 'reserv', 'aren’t', 'enough', 'cover', 'thr
ee', 'month', 'import']
['reserv', 'declin', '$1', 'trillion,', '7.8%,', 'year', '$12', 'trillion,'
, 'biggest', 'drop', 'sinc', 'bloomberg', 'start', 'compil', 'data', '2003'
]
['ril', 'bat', 'cap', 'domest', 'ga', 'price', 'govt', 'panel', 'seek', 're
view']
['flood', 'caus', 'collect', 'loss', 'usd', '40', 'billion.']
['stock', 'zeel', 'appreci', '13', 'per', 'cent,', 'compar', '3.5', 'per',
'cent', 'rise', 's&p', 'bse', 'sensex']
['12', 'medals,', '6', 'gold,', 'birmingham', 'made', 'top', 'contributor',
"india'", 'medal', 'talli', '61']
['trade', 'volum', 'soar', 'demat', 'talli', 'surpass', '102.5', 'million',
'account']
['trade', 'volum', 'soar', 'demat', 'talli', 'surpass', '102.5', 'million',
'account']
['vnit', 'receiv', '170', 'ppo', 'current', '2022-23', 'batch', 'last', 'ye
ar']
['world', 'currenc', 'reserv', 'shrink', '$1', 'trn', 'yr', 'record', 'draw
down']
['world', 'tabl', 'tenni', 'championship', '2022:', 'indian', 'men', 'enter
', 'pre-quarterfin', 'despit', 'franc', 'loss']
['wto', 'slash', 'global', '2023', 'trade', 'growth', 'forecast', '1%', 're
cess', 'loom']
['zee', 'entertain', 'gain', '6%', "cci'", 'condit', 'nod', 'merger', 'soni
']
['eco', 'eco', 'eco', 'eco', 'eco', 'eco', 'eco', 'sport', 'eco', 'eco', 'e
co', 'sport', 'sport', 'eco', 'eco']
```

Out[12]:
```
0.8666666666666667
```