Import packages

```
import pandas as pd
import numpy as np
from sklearn.feature_extraction.text import CountVectorizer
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import BernoulliNB
```

Import Dataset

```
data=pd.read_csv("/content/drive/MyDrive/machine-learning-data/Youtube01-Psy.csv")
print(data.sample(5))
```

```
                                 COMMENT_ID            AUTHOR  \
130    z12dzlpo0ueie1go404cfjjwsxf1g1lrtdo       Amir effect
116  z135f3rwwnjafnnrv04cirmifl3lipewgl40k          Akymonix
16     z13bgdvyluihfv11i22rgxwhuvabzz1os04       Zielimeek21
241      z124inzqgoyeh33uw23iibficv2kuf2nx  anthony Jennings
218    z12xczhjizrzuhvlp22gh3apkrasg3ggt04       Ben Stalker

                     DATE                                              CONTENT  \
130  2014-11-05T17:18:14  Can somebody wake me up when we get to 3 billi...
116  2014-11-05T07:01:34                                  Made in china....
16   2013-11-28T21:49:00                        I'm only checking the views
241  2014-11-07T23:26:04  People Who Say That "This Song Is Too Old Now,...
218  2014-11-07T19:27:45       GANGMAN STY- *D-D-D-D-D-D--DROP THE BASS!!*

     CLASS
130      0
116      0
16       0
241      0
218      0
```

Split the column

```
data = data[["CONTENT", "CLASS"]]
print(data.sample(10))
```

```
                                              CONTENT  CLASS
115                                     #2012bitches      0
67   OMG this oldspice spraytan party commercial om...      0
142  pls http://www10.vakinha.com.br/VaquinhaE.aspx...      1
165                                  Song name??        0
86                 Suscribe My Channel Please XD lol      1
343   Something to dance to, even if your sad JUST ...      0
188  Dear person reading this, You are beautiful an...      0
37   SUB 4 SUB PLEASE LIKE THIS COMMENT I WANT A SU...      1
277  Hey, join me on tsū, a publishing platform whe...      1
66                                     psy=korean      0
```

Indicate 0 for not spam and 1 for spam

```
data["CLASS"]= data["CLASS"].map({0:"Not SPAM", 1: "SPAM"})
print(data.sample(5))
```

```
                                              CONTENT     CLASS
259  Hey everyone, I am a new channel and will post...      SPAM
267  WHY DOES THIS HAVE 2 BILLION VIEWS THIS SONG I...  Not SPAM
81   Admit it you just came here to check the numbe...  Not SPAM
312  I still to this day wonder why this video is s...  Not SPAM
```

```
302  https://www.facebook.com/nicushorbboy add mee ...      SPAM
<ipython-input-22-3676386487f9>:1: SettingWithCopyWarning:
A value is trying to be set on a copy of a slice from a DataFrame.
Try using .loc[row_indexer,col_indexer] = value instead

See the caveats in the documentation: https://pandas.pydata.org/pandas-docs/stable/user_guide/indexing.htm
  data["CLASS"]= data["CLASS"].map({0:"Not SPAM", 1: "SPAM"})
```

Train Dataset with Bernoulli Algorithm

```
x=np.array(data["CONTENT"])
y=np.array(data["CLASS"])

cv=CountVectorizer()
x=cv.fit_transform(x)
xtrain, xtest, ytrain, ytest = train_test_split(x,y,test_size=0.2, random_state=42)

model=BernoulliNB()
model.fit(xtrain, ytrain)

print(model.score(xtest, ytest))
```

```
0.9857142857142858
```

Validate Trained Data

```
hWorld="please like http://www.grandi.com"
data=cv.transform([hWorld]).toarray()
print(model.predict(data))
```

```
['SPAM']
```