```
!pip install "scikit_learn==0.22.2.post1"
```

```
Looking in indexes: https://pypi.org/simple, https://us-python.pkg.dev/colab-wheels/pub
Requirement already satisfied: scikit_learn==0.22.2.post1 in /usr/local/lib/python3.7/di
Requirement already satisfied: scipy>=0.17.0 in /usr/local/lib/python3.7/dist-packages (
Requirement already satisfied: joblib>=0.11 in /usr/local/lib/python3.7/dist-packages (
Requirement already satisfied: numpy>=1.11.0 in /usr/local/lib/python3.7/dist-packages (
```

```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
import pandas as pd
import io
from google.colab import files


uploaded = files.upload()
dataframe = pd.read_csv(io.BytesIO(uploaded['federalist.csv']))
```

Choose Files federalist.csv

- **federalist.csv**(text/csv) - 1100616 bytes, last modified: 11/8/2022 - 100% done

```
Saving federalist.csv to federalist (5).csv
```

Author column to categorical data

```
dataframe['author'] = dataframe['author'].astype("category")
uniques = dataframe['author'].unique()
print(dataframe[0:4])
```

```
     author                                                text
0  HAMILTON  FEDERALIST. No. 1 General Introduction For the...
1       JAY  FEDERALIST No. 2 Concerning Dangers from Forei...
2       JAY  FEDERALIST No. 3 The Same Subject Continued (C...
3       JAY  FEDERALIST No. 4 The Same Subject Continued (C...
```

Split Data

```
from nltk.corpus import stopwords
import nltk
nltk.download('stopwords')

input = dataframe['text']
output = dataframe['author']

inp_train, inp_test, tar_train, tar_test = train_test_split(input, output, test_size=0.2, tra

print(inp_train.shape, tar_train.shape)
```

```
print(inp_test.shape, tar_test.shape)

stopwords = set(stopwords.words('english'))
vectorizer = TfidfVectorizer(stop_words=stopwords)
v_inp_train = vectorizer.fit_transform(inp_train)
v_inp_test = vectorizer.transform(inp_test)

print(v_inp_train.shape)
print(v_inp_test.shape)
```

```
(66,) (66,)
(17,) (17,)
(66, 7876)
(17, 7876)
[nltk_data] Downloading package stopwords to /root/nltk_data...
[nltk_data]   Package stopwords is already up-to-date!
```

Bernoulli Bayes Model

```
from sklearn.naive_bayes import BernoulliNB
from sklearn.metrics import accuracy_score, precision_score, recall_score, f1_score, confusio

bayes = BernoulliNB()
bayes.fit(v_inp_train, tar_train)

pred = bayes.predict(v_inp_test)

print('accuracy score: ', accuracy_score(tar_test, pred))
```

```
accuracy score:  0.5882352941176471
```

Redo with Vectorization limited to 1000 and added unigrams + bigrams

```
vectorizer = TfidfVectorizer(stop_words=stopwords, max_features=1000, ngram_range=(1,2))
v_inp_train = vectorizer.fit_transform(inp_train)
v_inp_test = vectorizer.transform(inp_test)

bayes = BernoulliNB()
bayes.fit(v_inp_train, tar_train)

pred = bayes.predict(v_inp_test)

print('accuracy score: ', accuracy_score(tar_test, pred))
```

```
accuracy score:  0.9411764705882353
```

Logistic Regression

```
from sklearn.linear_model import LogisticRegression

vectorizer = TfidfVectorizer()
v_inp_train = vectorizer.fit_transform(inp_train)
v_inp_test = vectorizer.transform(inp_test)

classifier = LogisticRegression(multi_class='multinomial', solver='lbfgs',class_weight='balan
classifier.fit(v_inp_train, tar_train)
pred = classifier.predict(v_inp_test)

print('accuracy score: ', accuracy_score(tar_test, pred))
```

```
accuracy score:  0.9411764705882353
```

Neural Network

```
from sklearn.neural_network import MLPClassifier

vectorizer = TfidfVectorizer()
v_inp_train = vectorizer.fit_transform(inp_train)
v_inp_test = vectorizer.transform(inp_test)

nn = MLPClassifier(solver='lbfgs',
                   alpha=5e-5,
                   hidden_layer_sizes=(100, 17, 7),
                   random_state=1)

nn.fit(v_inp_train, tar_train)
pred = nn.predict(v_inp_test)
print('accuracy score: ', accuracy_score(tar_test, pred))
```

```
accuracy score:  0.7647058823529411
```

Colab paid products - Cancel contracts here

✓ 22s completed at 6:37 PM