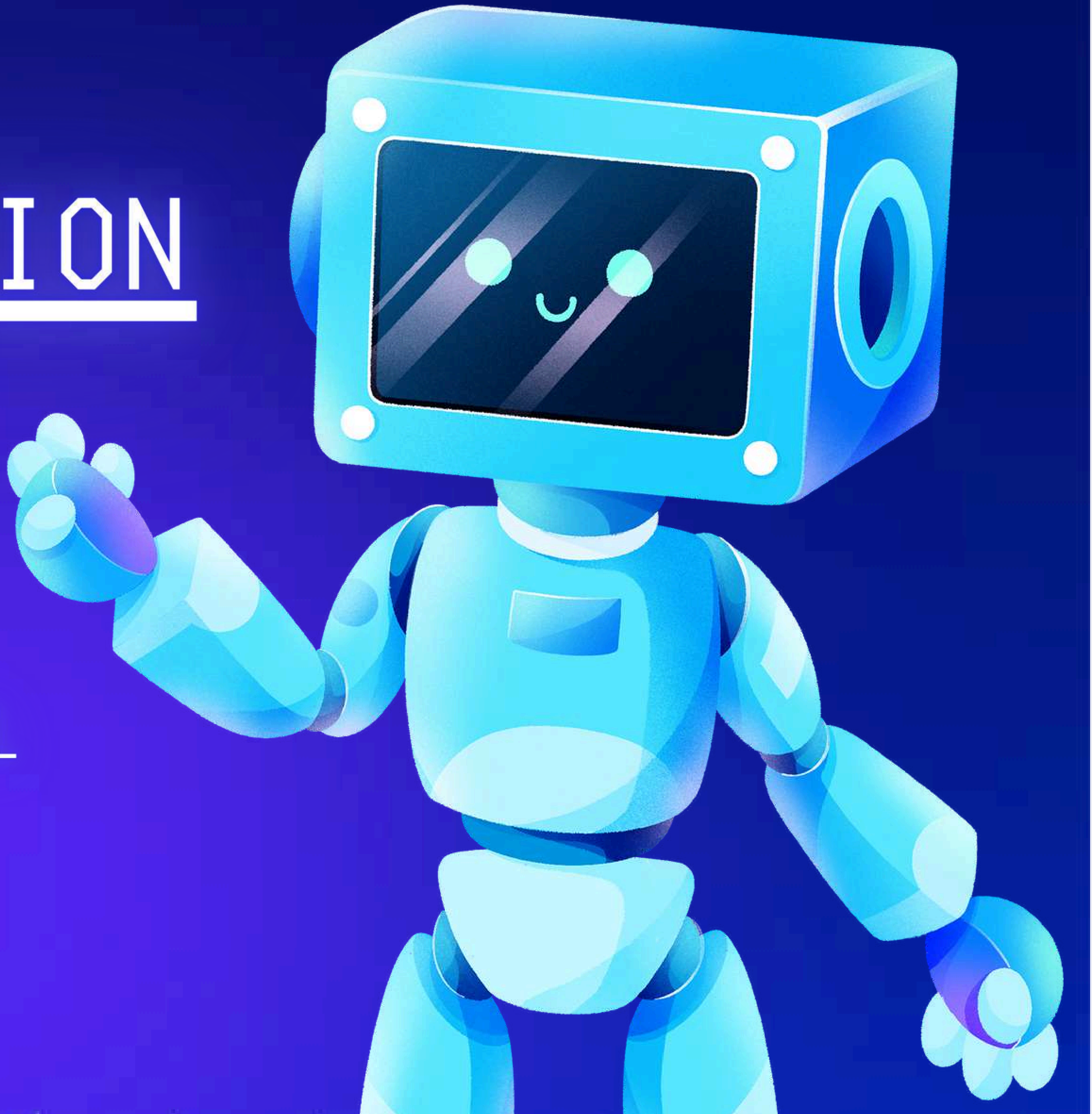
FILE SEGREGATION

ML PROJECT

By - BHAVYA SEHGAL



INTRODUCTION

PURPOSE

OVERVIEW

EXAMPLES

APPLICATIONS

NAIVE BAYE'S

Bayes' Theorem is stated as:

P(h|d) = (P(d|h) * P(h)) / P(d)

Where

- P(h|d) is the probability of hypothesis h given the data d. This is called the posterior probability.
- P(d|h) is the probability of data d given that the hypothesis h was true.
- P(h) is the probability of hypothesis h being true (regardless of the data). This is called the prior probability of h.
- P(d) is the probability of the data (regardless of the hypothesis).

PROCESS STEPS

01

COLLECTION OF DATA SET FROM SK LEARN

02

DEFINING CATEGORIES

03

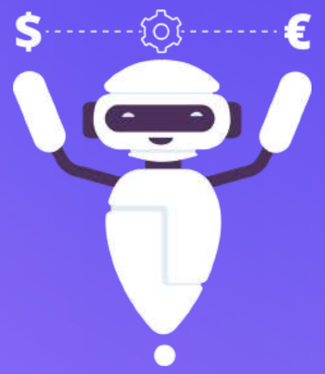
TESTING & TRAINING DATA

04

DISPLAY OF RESULTS AND HEAT MAP

CATEGORIES OF DATASETS

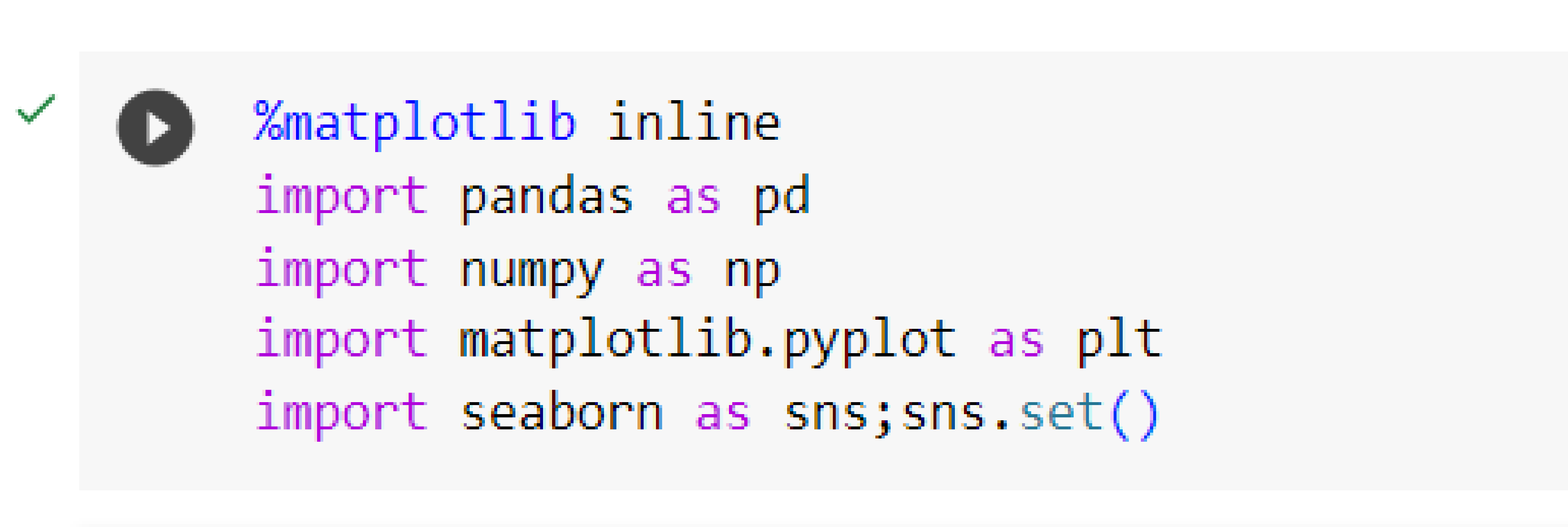
'alt.atheism', 'comp.graphics', 'comp.os.ms-windows.misc', 'comp.sys.ibm.pc.hardware', 'comp.sys.mac.hardware', 'misc.forsale',



'rec.autos', 'rec.motorcycles', 'rec.sport.baseball', 'rec.sport.hockey', 'sci.crypt', 'sci.electronics', 'sci.med',

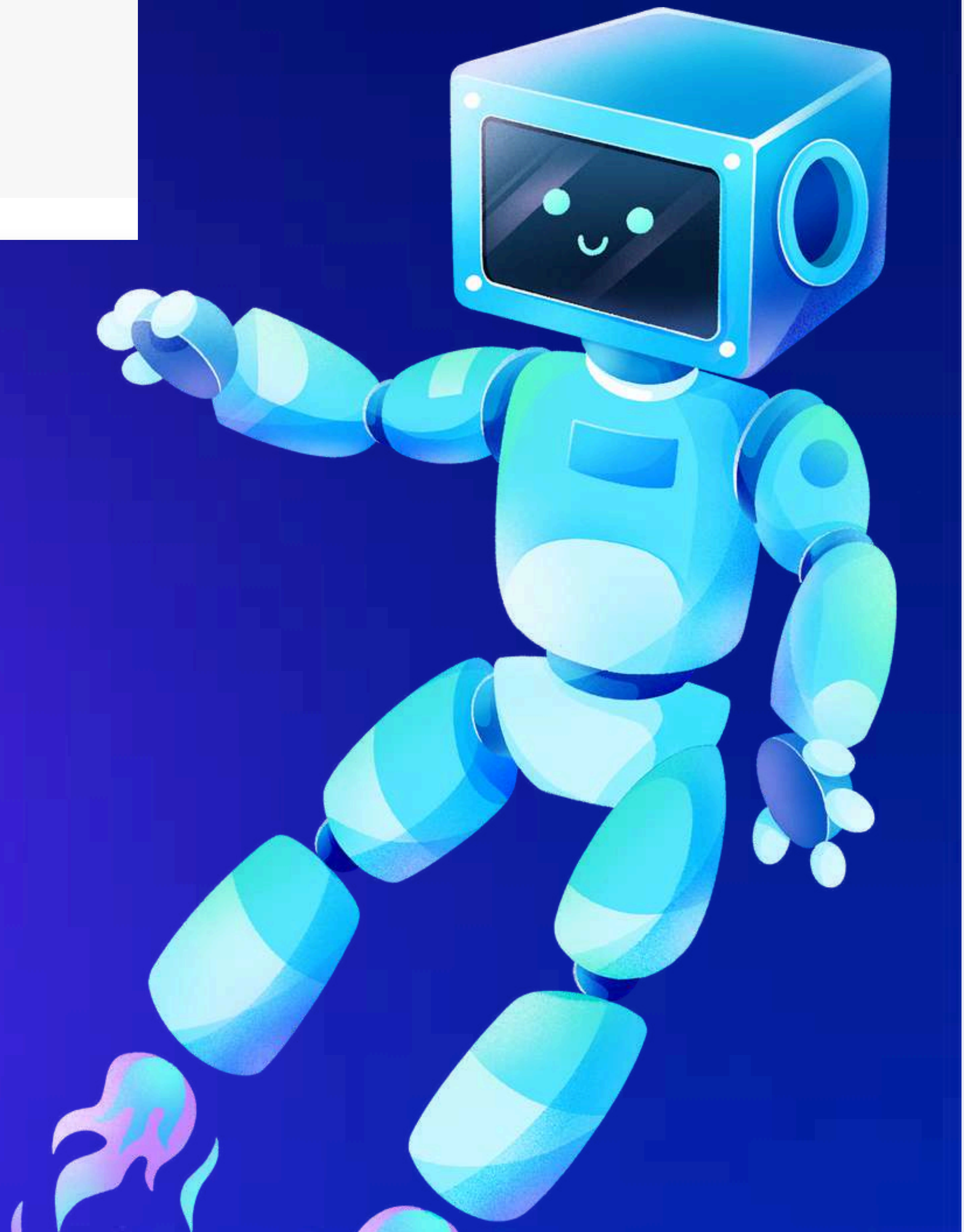
'sci.space', 'soc.religion.christian', 'talk.politics.guns', 'talk.politics.mideast', 'talk.politics.misc', 'talk.religion.misc' 'comp.windows.x',

WALKAROUND OF THE CODE

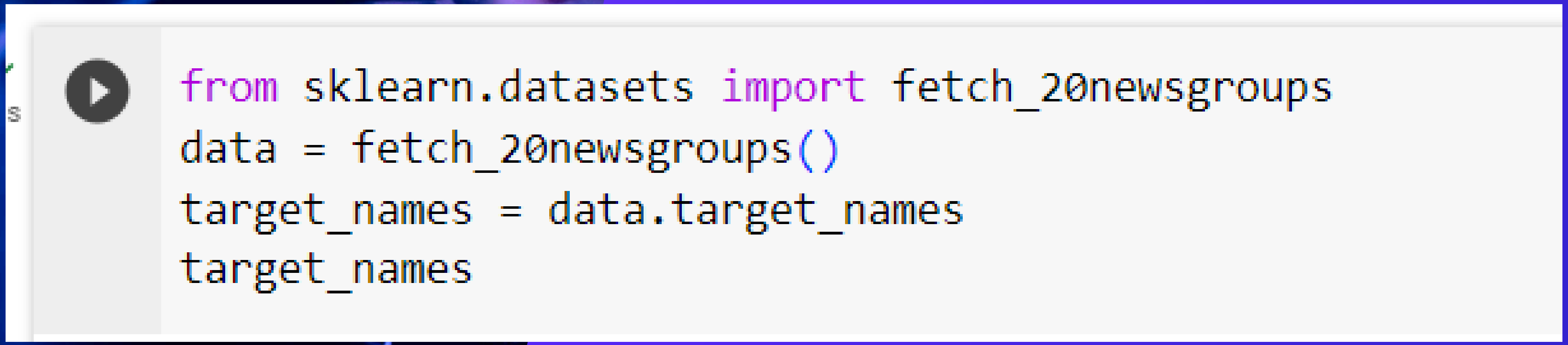


```python
%matplotlib inline
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns;sns.set()
```

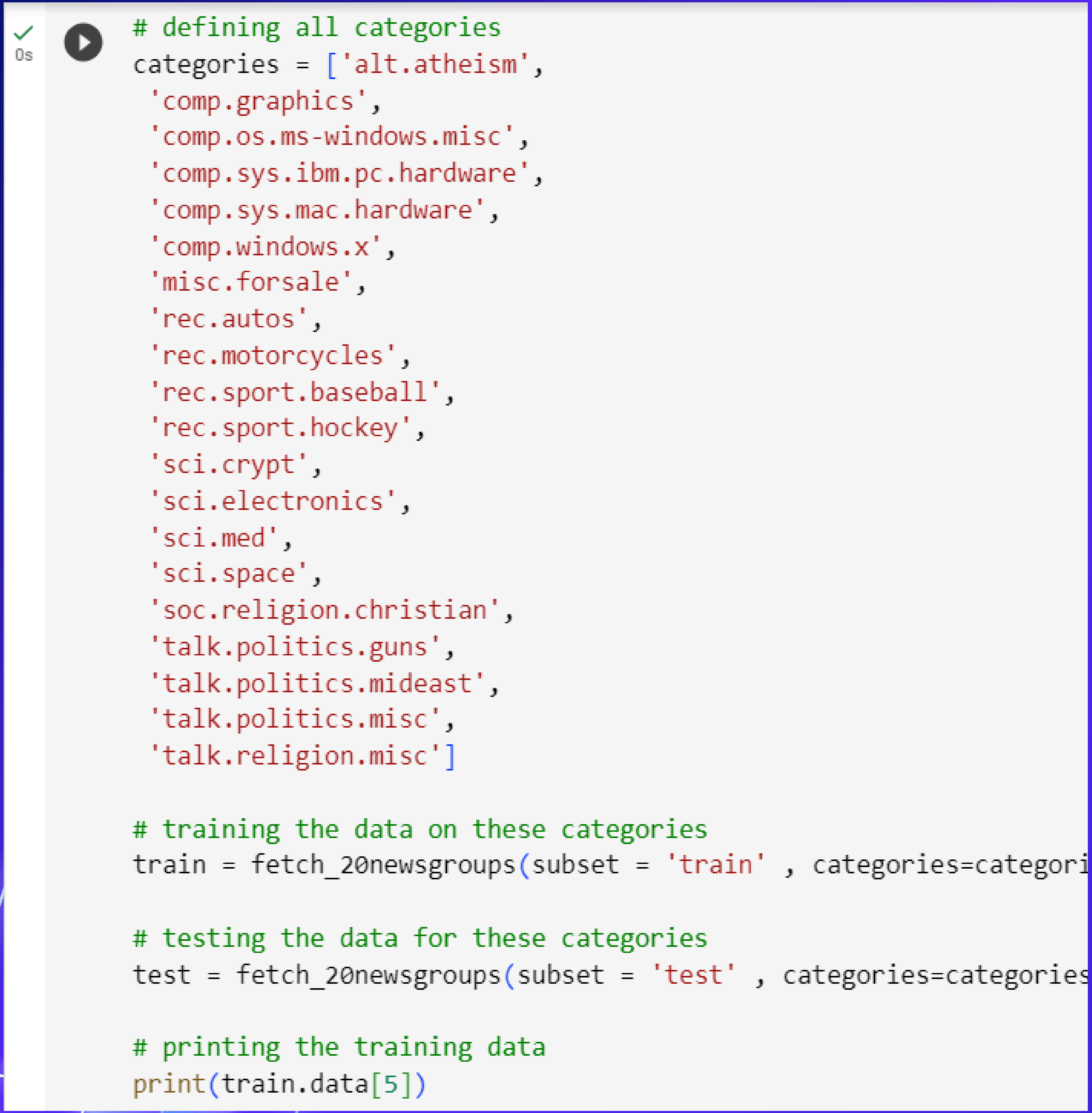
IMPORTING MODULES AND PACKAGES



FETCHING DATASETS FROM SKLEARN

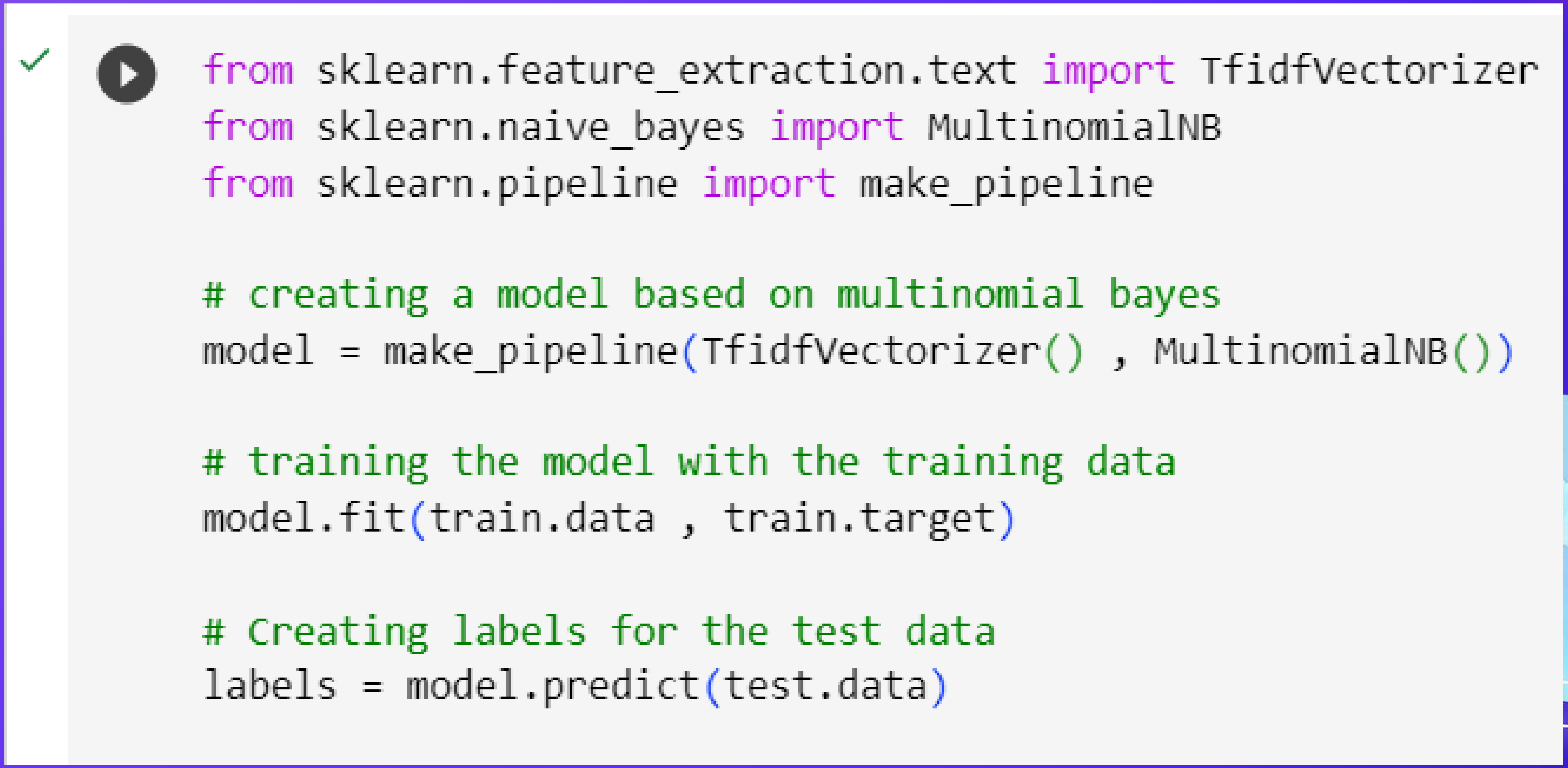


```
from sklearn.datasets import fetch_20newsgroups
data = fetch_20newsgroups()
target_names = data.target_names
target_names
```

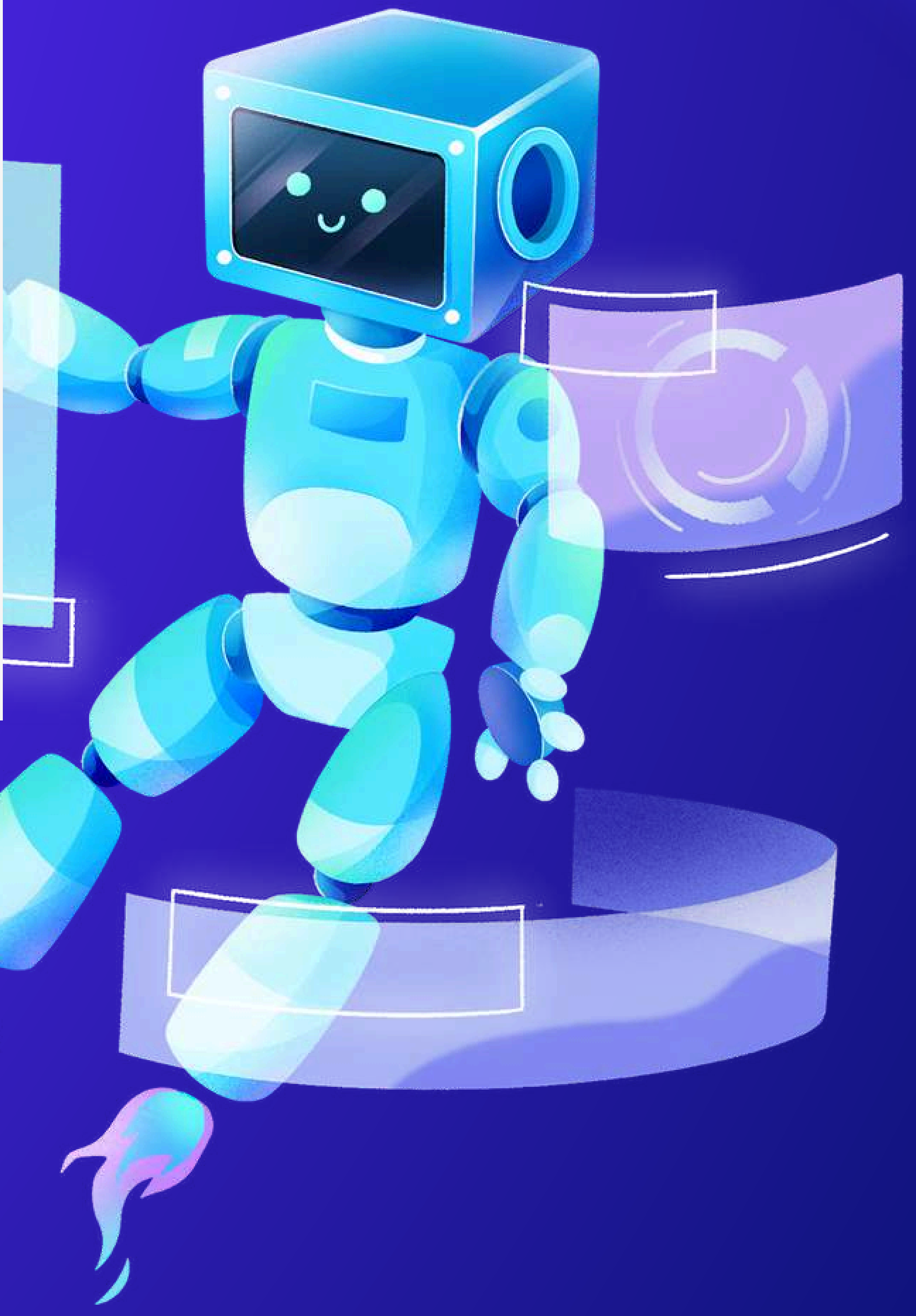


```
# defining all categories
categories = ['alt.atheism',
 'comp.graphics',
 'comp.os.ms-windows.misc',
 'comp.sys.ibm.pc.hardware',
 'comp.sys.mac.hardware',
 'comp.windows.x',
 'misc.forsale',
 'rec.autos',
 'rec.motorcycles',
 'rec.sport.baseball',
 'rec.sport.hockey',
 'sci.crypt',
 'sci.electronics',
 'sci.med',
 'sci.space',
 'soc.religion.christian',
 'talk.politics.guns',
 'talk.politics.mideast',
 'talk.politics.misc',
 'talk.religion.misc']

# training the data on these categories
train = fetch_20newsgroups(subset = 'train' , categories=categori

# testing the data for these categories
test = fetch_20newsgroups(subset = 'test' , categories=categories

# printing the training data
print(train.data[5])
```

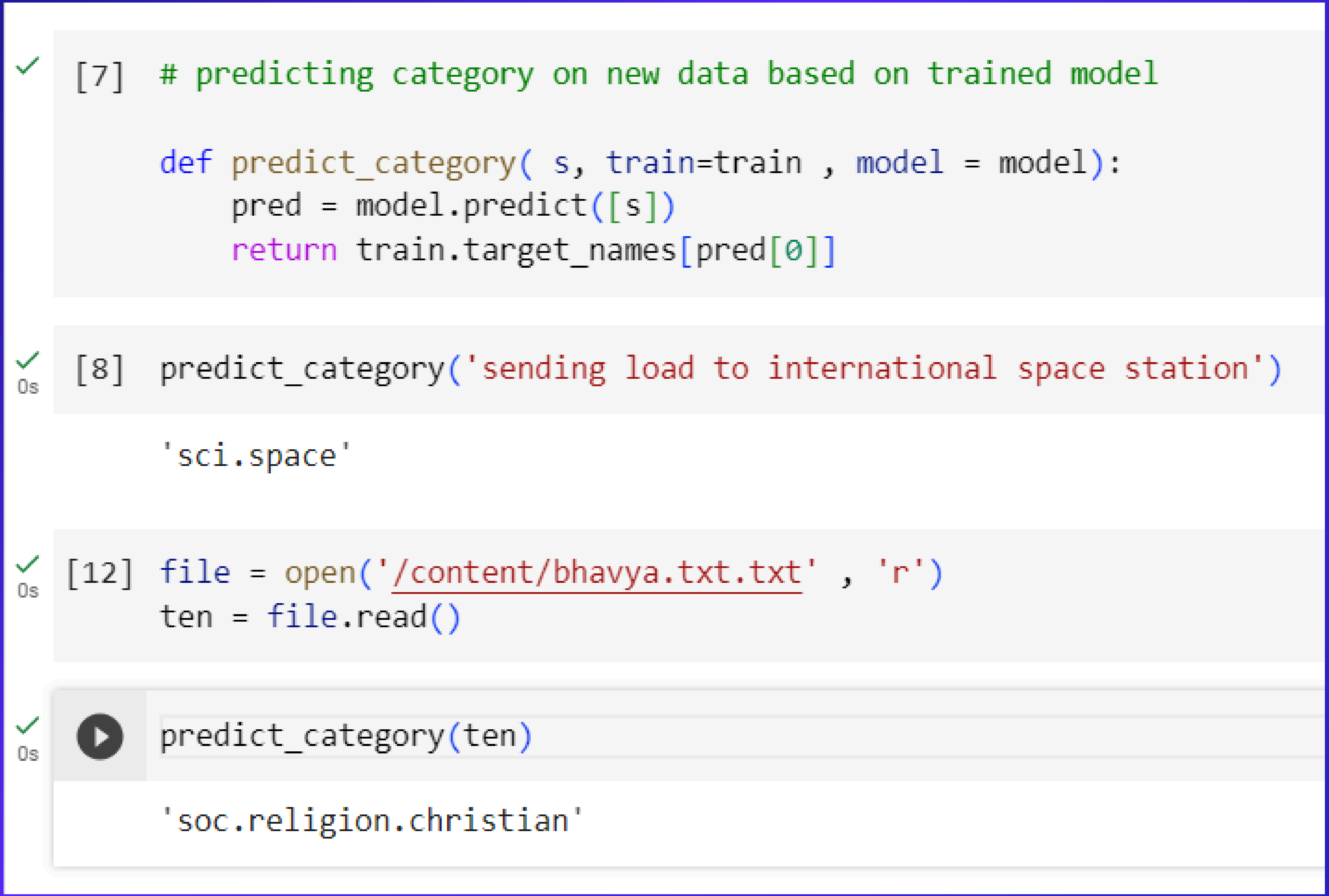
DEFINING ALL CATEGORIES AND TRAINING AND TESTING DATA



```
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.naive_bayes import MultinomialNB
from sklearn.pipeline import make_pipeline

# creating a model based on multinomial bayes
model = make_pipeline(TfidfVectorizer() , MultinomialNB())

# training the model with the training data
model.fit(train.data , train.target)

# Creating labels for the test data
labels = model.predict(test.data)
```



PREDICTING CATEGORIES



```
# predicting category on new data based on trained model

def predict_category( s, train=train , model = model):
    pred = model.predict([s])
    return train.target_names[pred[0]]
```

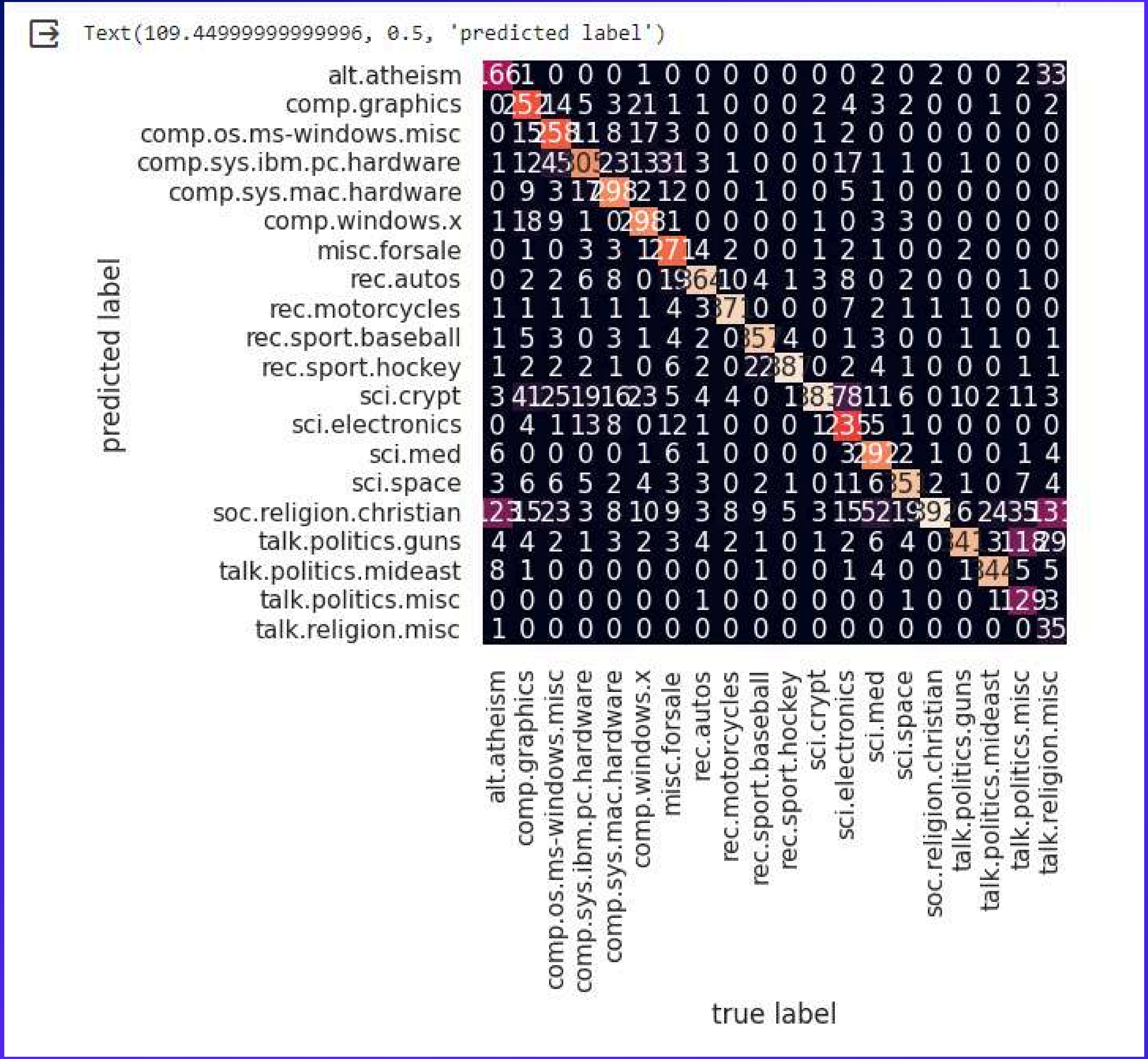
```
predict_category('sending load to international space station')
```

```
'sci.space'
```

```
file = open('/content/bhavya.txt.txt' , 'r')
ten = file.read()
```

```
predict_category(ten)
```

```
'soc.religion.christian'
```



```
Text(109.44999999999996, 0.5, 'predicted label')
```

HEAT MAP