

```
In [1]: !pip install numpy  
!pip install pandas  
!pip install matplotlib  
!pip install seaborn  
!pip install -U scikit-learn
```

Requirement already satisfied: numpy in c:\users\divya\anaconda3\lib\site-packages (1.24.3)
Requirement already satisfied: pandas in c:\users\divya\anaconda3\lib\site-packages (2.0.3)
Requirement already satisfied: python-dateutil>=2.8.2 in c:\users\divya\anaconda3\lib\site-packages (from pandas) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in c:\users\divya\anaconda3\lib\site-packages (from pandas) (2023.3.post1)
Requirement already satisfied: tzdata>=2022.1 in c:\users\divya\anaconda3\lib\site-packages (from pandas) (2023.3)
Requirement already satisfied: numpy>=1.21.0 in c:\users\divya\anaconda3\lib\site-packages (from pandas) (1.24.3)
Requirement already satisfied: six>=1.5 in c:\users\divya\anaconda3\lib\site-packages (from python-dateutil>=2.8.2->pandas) (1.16.0)
Requirement already satisfied: matplotlib in c:\users\divya\anaconda3\lib\site-packages (3.7.2)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\divya\anaconda3\lib\site-packages (from matplotlib) (1.0.5)
Requirement already satisfied: cyclor>=0.10 in c:\users\divya\anaconda3\lib\site-packages (from matplotlib) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\divya\anaconda3\lib\site-packages (from matplotlib) (4.25.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\divya\anaconda3\lib\site-packages (from matplotlib) (1.4.4)
Requirement already satisfied: numpy>=1.20 in c:\users\divya\anaconda3\lib\site-packages (from matplotlib) (1.24.3)
Requirement already satisfied: packaging>=20.0 in c:\users\divya\anaconda3\lib\site-packages (from matplotlib) (23.1)
Requirement already satisfied: pillow>=6.2.0 in c:\users\divya\anaconda3\lib\site-packages (from matplotlib) (9.4.0)
Requirement already satisfied: pyparsing<3.1,>=2.3.1 in c:\users\divya\anaconda3\lib\site-packages (from matplotlib) (3.0.9)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\divya\anaconda3\lib\site-packages (from matplotlib) (2.8.2)
Requirement already satisfied: six>=1.5 in c:\users\divya\anaconda3\lib\site-packages (from python-dateutil>=2.7->matplotlib) (1.16.0)
Requirement already satisfied: seaborn in c:\users\divya\anaconda3\lib\site-packages (0.12.2)
Requirement already satisfied: numpy!=1.24.0,>=1.17 in c:\users\divya\anaconda3\lib\site-packages (from seaborn) (1.24.3)
Requirement already satisfied: pandas>=0.25 in c:\users\divya\anaconda3\lib\site-packages (from seaborn) (2.0.3)
Requirement already satisfied: matplotlib!=3.6.1,>=3.1 in c:\users\divya\anaconda3\lib\site-packages (from seaborn) (3.7.2)
Requirement already satisfied: contourpy>=1.0.1 in c:\users\divya\anaconda3\lib\site-packages (from matplotlib!=3.6.1,>=3.1->seaborn) (1.0.5)
Requirement already satisfied: cyclor>=0.10 in c:\users\divya\anaconda3\lib\site-packages (from matplotlib!=3.6.1,>=3.1->seaborn) (0.11.0)
Requirement already satisfied: fonttools>=4.22.0 in c:\users\divya\anaconda3\lib\site-packages (from matplotlib!=3.6.1,>=3.1->seaborn) (4.25.0)
Requirement already satisfied: kiwisolver>=1.0.1 in c:\users\divya\anaconda3\lib\site-packages (from matplotlib!=3.6.1,>=3.1->seaborn) (1.4.4)
Requirement already satisfied: packaging>=20.0 in c:\users\divya\anaconda3\lib\site-packages (from matplotlib!=3.6.1,>=3.1->seaborn) (23.1)
Requirement already satisfied: pillow>=6.2.0 in c:\users\divya\anaconda3\lib\site-packages (from matplotlib!=3.6.1,>=3.1->seaborn) (9.4.0)
Requirement already satisfied: pyparsing<3.1,>=2.3.1 in c:\users\divya\anaconda3\lib\site-packages (from matplotlib!=3.6.1,>=3.1->seaborn) (3.0.9)
Requirement already satisfied: python-dateutil>=2.7 in c:\users\divya\anaconda3\lib\site-packages (from matplotlib!=3.6.1,>=3.1->seaborn) (2.8.2)
Requirement already satisfied: pytz>=2020.1 in c:\users\divya\anaconda3\lib\site-packages (from pandas>=0.25->seaborn) (2023.3.post1)
Requirement already satisfied: tzdata>=2022.1 in c:\users\divya\anaconda3\lib\site-packages (from pandas>=0.25->seaborn) (2023.3)

Requirement already satisfied: six>=1.5 in c:\users\divya\anaconda3\lib\site-packages (from python-dateutil>=2.7->matplotlib!=3.6.1,>=3.1->seaborn) (1.16.0)
Requirement already satisfied: scikit-learn in c:\users\divya\anaconda3\lib\site-packages (1.5.0)
Requirement already satisfied: numpy>=1.19.5 in c:\users\divya\anaconda3\lib\site-packages (from scikit-learn) (1.24.3)
Requirement already satisfied: scipy>=1.6.0 in c:\users\divya\anaconda3\lib\site-packages (from scikit-learn) (1.11.1)
Requirement already satisfied: joblib>=1.2.0 in c:\users\divya\anaconda3\lib\site-packages (from scikit-learn) (1.2.0)
Requirement already satisfied: threadpoolctl>=3.1.0 in c:\users\divya\anaconda3\lib\site-packages (from scikit-learn) (3.5.0)

```
In [2]: import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import re
import nltk
import string
from nltk.corpus import stopwords
from nltk.stem import LancasterStemmer
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.model_selection import train_test_split
from sklearn.naive_bayes import MultinomialNB
from sklearn.metrics import accuracy_score, classification_report
```

```
In [3]: train_path = "train_data.txt"
train_data = pd.read_csv(train_path, sep=':::', names=['Title', 'Genre', 'Description'], engine='python')
train_data
```

Out[3]:

	Title	Genre	Description
1	Oscar et la dame rose (2009)	drama	Listening in to a conversation between his do...
2	Cupid (1997)	thriller	A brother and sister with a past incestuous r...
3	Young, Wild and Wonderful (1980)	adult	As the bus empties the students for their fie...
4	The Secret Sin (1915)	drama	To help their unemployed father make ends mee...
5	The Unrecovered (2007)	drama	The film's title refers not only to the un-re...
...
54210	"Bonino" (1953)	comedy	This short-lived NBC live sitcom centered on ...
54211	Dead Girls Don't Cry (????)	horror	The NEXT Generation of EXPLOITATION. The sist...
54212	Ronald Goedemondt: Ze bestaan echt (2008)	documentary	Ze bestaan echt, is a stand-up comedy about g...
54213	Make Your Own Bed (1944)	comedy	Walter and Vivian live in the country and hav...
54214	Nature's Fury: Storm of the Century (2006)	history	On Labor Day Weekend, 1935, the most intense ...

54214 rows × 3 columns

```
In [4]: test_path = "test_data.txt"
test_data = pd.read_csv(test_path, sep=':::', names=['Id', 'Title', 'Description'], engine='python')
test_data
```

Out[4]:

	Id	Title	Description
0	1	Edgar's Lunch (1998)	L.R. Brane loves his life - his car, his apar...
1	2	La guerra de papá (1977)	Spain, March 1964: Quico is a very naughty ch...
2	3	Off the Beaten Track (2010)	One year in the life of Albin and his family ...
3	4	Meu Amigo Hindu (2015)	His father has died, he hasn't spoken with hi...
4	5	Er nu zhai (1955)	Before he was known internationally as a mart...
...
54195	54196	"Tales of Light & Dark" (2013)	Covering multiple genres, Tales of Light & Da...
54196	54197	Der letzte Mohikaner (1965)	As Alice and Cora Munro attempt to find their...
54197	54198	Oliver Twink (2007)	A movie 169 years in the making. Oliver Twist...
54198	54199	Slipstream (1973)	Popular, but mysterious rock D.J Mike Mallard...
54199	54200	Curitiba Zero Grau (2010)	Curitiba is a city in movement, with rhythms ...

54200 rows × 3 columns

In [5]: `train_data.describe()`

Out[5]:

	Title	Genre	Description
count	54214	54214	54214
unique	54214	27	54086
top	Oscar et la dame rose (2009)	drama	Grammy - music award of the American academy ...
freq	1	13613	12

In [6]: `test_data.describe()`

Out[6]:

	Id
count	54200.000000
mean	27100.500000
std	15646.336632
min	1.000000
25%	13550.750000
50%	27100.500000
75%	40650.250000
max	54200.000000

In [7]: `train_data.isnull().sum()`

Out[7]: Title 0
Genre 0
Description 0
dtype: int64

In [8]: `test_data.isnull().sum()`

Out[8]: Id 0
Title 0
Description 0
dtype: int64

In [9]: `class_distribution = train_data['Genre'].value_counts()
print("Class Distribution:")
print(class_distribution)`

Class Distribution:

Genre

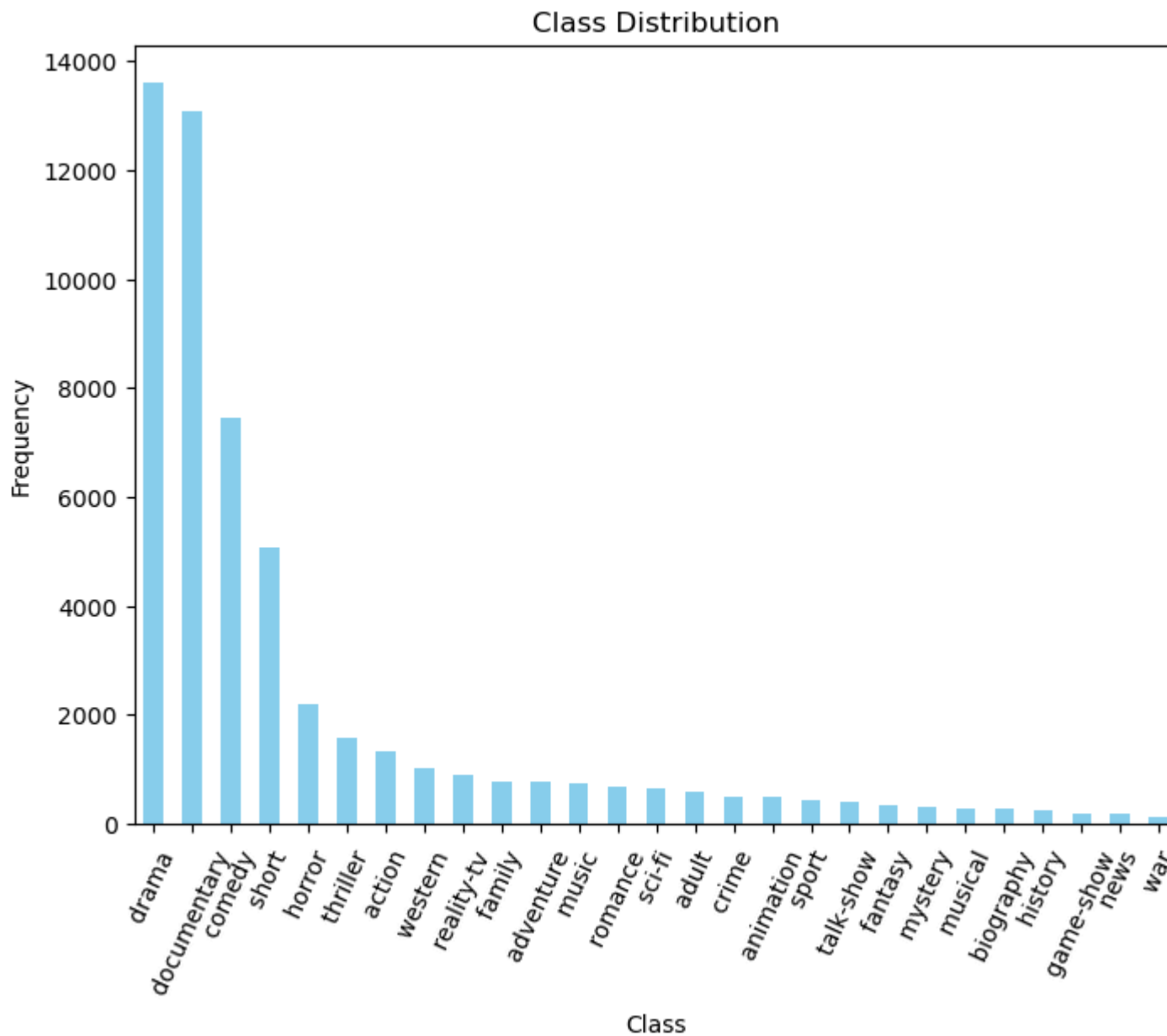
drama	13613
documentary	13096
comedy	7447
short	5073
horror	2204
thriller	1591
action	1315
western	1032
reality-tv	884
family	784
adventure	775
music	731
romance	672
sci-fi	647
adult	590
crime	505
animation	498
sport	432
talk-show	391
fantasy	323
mystery	319
musical	277
biography	265
history	243
game-show	194
news	181
war	132

Name: count, dtype: int64

```
In [10]: imbalance_ratio = class_distribution.min() / class_distribution.max()
print("Imbalance Ratio:", imbalance_ratio)
```

Imbalance Ratio: 0.009696613531183427

```
In [11]: plt.figure(figsize=(8, 6))
class_distribution.plot(kind='bar', color='skyblue')
plt.title('Class Distribution')
plt.xlabel('Class')
plt.ylabel('Frequency')
plt.xticks(rotation=65)
plt.show()
```



```
In [1]: tfidf_vectorizer = TfidfVectorizer(max_features=5000)
X_train_tfidf = tfidf_vectorizer.fit_transform(train_data['Description'])
y_train = train_data['Genre']
```



```
nb_classifier = MultinomialNB()
nb_classifier.fit(X_train_tfidf, y_train)

y_train_pred = nb_classifier.predict(X_train_tfidf)
print("Accuracy on training set:", accuracy_score(y_train, y_train_pred))
print("Classification Report on training set:\n", classification_report(y_train, y_train_pred))
```

```
-----
NameError                                Traceback (most recent call last)
Cell In[1], line 1
----> 1 tfidf_vectorizer = TfidfVectorizer(max_features=5000)
      2 X_train_tfidf = tfidf_vectorizer.fit_transform(train_data['Description'])
      3 y_train = train_data['Genre']

NameError: name 'TfidfVectorizer' is not defined
```

```
In [14]: tfidf_vectorizer = TfidfVectorizer(max_features=5000)
X_test = tfidf_vectorizer.fit_transform(test_data['Description'])
```

```
In [15]: X_test_predictions = nb_classifier.predict(X_test)
test_data['Predicted_Genre'] = X_test_predictions
```

```
In [16]: test_data.to_csv('predicted_genres.csv', index=False)

print(test_data)
```

	Id	Title \
0	1	Edgar's Lunch (1998)
1	2	La guerra de papá (1977)
2	3	Off the Beaten Track (2010)
3	4	Meu Amigo Hindu (2015)
4	5	Er nu zhai (1955)
...
54195	54196	"Tales of Light & Dark" (2013)
54196	54197	Der letzte Mohikaner (1965)
54197	54198	Oliver Twink (2007)
54198	54199	Slipstream (1973)
54199	54200	Curitiba Zero Grau (2010)

	Description	Predicted_Genre
0	L.R. Brane loves his life - his car, his apar...	drama
1	Spain, March 1964: Quico is a very naughty ch...	drama
2	One year in the life of Albin and his family ...	documentary
3	His father has died, he hasn't spoken with hi...	documentary
4	Before he was known internationally as a mart...	documentary
...
54195	Covering multiple genres, Tales of Light & Da...	drama
54196	As Alice and Cora Munro attempt to find their...	drama
54197	A movie 169 years in the making. Oliver Twist...	documentary
54198	Popular, but mysterious rock D.J Mike Mallard...	drama
54199	Curitiba is a city in movement, with rhythms ...	short

[54200 rows x 4 columns]

```
In [17]: import pickle
with open('tfidf_vectorizer.pkl', 'wb') as file:
    pickle.dump(tfidf_vectorizer, file)
with open('nb_classifier.pkl', 'wb') as file:
    pickle.dump(nb_classifier, file)

print("Models pickled successfully.")
```

Models pickled successfully.

In []: