Import modules

```
!pip install SVM
```

```
Collecting SVM
  Downloading svm-0.1.0.tar.gz (3.4 kB)
  Preparing metadata (setup.py) ... ent already satisfied: requests in
/opt/conda/lib/python3.10/site-packages (from SVM) (2.31.0)
Requirement already satisfied: colorama in
/opt/conda/lib/python3.10/site-packages (from SVM) (0.4.6)
Collecting xmltodict (from SVM)
  Downloading xmltodict-0.13.0-py2.py3-none-any.whl (10.0 kB)
Requirement already satisfied: charset-normalizer<4,>=2 in
/opt/conda/lib/python3.10/site-packages (from requests->SVM) (3.2.0)
Requirement already satisfied: idna<4,>=2.5 in
/opt/conda/lib/python3.10/site-packages (from requests->SVM) (3.4)
Requirement already satisfied: urllib3<3,>=1.21.1 in
/opt/conda/lib/python3.10/site-packages (from requests->SVM) (1.26.15)

Requirement already satisfied: certifi>=2017.4.17 in
/opt/conda/lib/python3.10/site-packages (from requests->SVM)
(2023.11.17)
Building wheels for collected packages: SVM
  Building wheel for SVM (setup.py) ... e=svm-0.1.0-py3-none-any.whl
size=3465
sha256=9c4abd7ff26b21cabadb32683ea4faf2d1341bfba357b210ec0c6f566562660
6
  Stored in directory:
/root/.cache/pip/wheels/dc/0a/16/c3cfc069f00231db8d16bc70bc747c155395d
bd30843a61957
Successfully built SVM
Installing collected packages: xmltodict, SVM
Successfully installed SVM-0.1.0 xmltodict-0.13.0
```

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
import re  # used for pattern matching and text manipulation.
import string
import nltk #a powerful library for working with human language data.
from nltk.corpus import stopwords #for cleaning
from nltk.stem import LancasterStemmer ##for cleaning
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.svm import SVC
from sklearn.model_selection import train_test_split
from sklearn.metrics import accuracy_score
```

Read a data

```
train_path= "/kaggle/input/genre-classification-dataset-imdb/Genre
Classification Dataset/train_data.txt"
train_data = pd.read_csv(train_path, sep=":::", names=["TITLE",
"GENRE", "DESCRIPTION"], engine="python")

train_data
```

```
                                          TITLE          GENRE  \
1                    Oscar et la dame rose (2009)          drama
2                                Cupid (1997)       thriller
3                Young, Wild and Wonderful (1980)          adult
4                           The Secret Sin (1915)          drama
5                          The Unrecovered (2007)          drama
...                                           ...            ...
54210                            "Bonino" (1953)         comedy
54211                 Dead Girls Don't Cry (????)         horror
54212    Ronald Goedemondt: Ze bestaan echt (2008)    documentary
54213                     Make Your Own Bed (1944)         comedy
54214   Nature's Fury: Storm of the Century (2006)        history

                                                DESCRIPTION
1       Listening in to a conversation between his do...
2       A brother and sister with a past incestuous r...
3       As the bus empties the students for their fie...
4       To help their unemployed father make ends mee...
5       The film's title refers not only to the un-re...
...                                                   ...
54210   This short-lived NBC live sitcom centered on ...
54211   The NEXT Generation of EXPLOITATION. The sist...
54212   Ze bestaan echt, is a stand-up comedy about g...
54213   Walter and Vivian live in the country and hav...
54214   On Labor Day Weekend, 1935, the most intense ...

[54214 rows x 3 columns]
```

```
train_data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
Index: 54214 entries, 1 to 54214
Data columns (total 3 columns):
 #   Column       Non-Null Count  Dtype
---  ------       --------------  -----
 0   TITLE        54214 non-null  object
 1   GENRE        54214 non-null  object
 2   DESCRIPTION  54214 non-null  object
dtypes: object(3)
memory usage: 1.7+ MB
```

```
train_data.describe()
```

```
                                TITLE    GENRE  \
count                           54214    54214
unique                          54214       27
top      Oscar et la dame rose (2009)    drama
freq                                1    13613

                                                DESCRIPTION
count                                                 54214
unique                                                54086
top      Grammy - music award of the American academy ...
freq                                                     12
```

```
train_data.isnull().sum()
```

```
TITLE          0
GENRE          0
DESCRIPTION    0
dtype: int64
```

```
test_path= "/kaggle/input/genre-classification-dataset-imdb/Genre
Classification Dataset/test_data.txt"
test_data = pd.read_csv(train_path, sep=":::",
names=["ID","TITLE","DESCRIPTION"], engine="python")
```

```
test_data
```

```
                                             ID          TITLE  \
1                    Oscar et la dame rose (2009)          drama
2                                 Cupid (1997)       thriller
3                 Young, Wild and Wonderful (1980)          adult
4                           The Secret Sin (1915)          drama
5                         The Unrecovered (2007)          drama
...                                          ...            ...
54210                      "Bonino" (1953)         comedy
54211              Dead Girls Don't Cry (????)         horror
54212    Ronald Goedemondt: Ze bestaan echt (2008)    documentary
54213                Make Your Own Bed (1944)         comedy
54214   Nature's Fury: Storm of the Century (2006)        history

                                               DESCRIPTION
1       Listening in to a conversation between his do...
2       A brother and sister with a past incestuous r...
3       As the bus empties the students for their fie...
4       To help their unemployed father make ends mee...
5       The film's title refers not only to the un-re...
...                                                   ...
54210   This short-lived NBC live sitcom centered on ...
54211   The NEXT Generation of EXPLOITATION. The sist...
54212   Ze bestaan echt, is a stand-up comedy about g...
54213   Walter and Vivian live in the country and hav...
54214   On Labor Day Weekend, 1935, the most intense ...
```

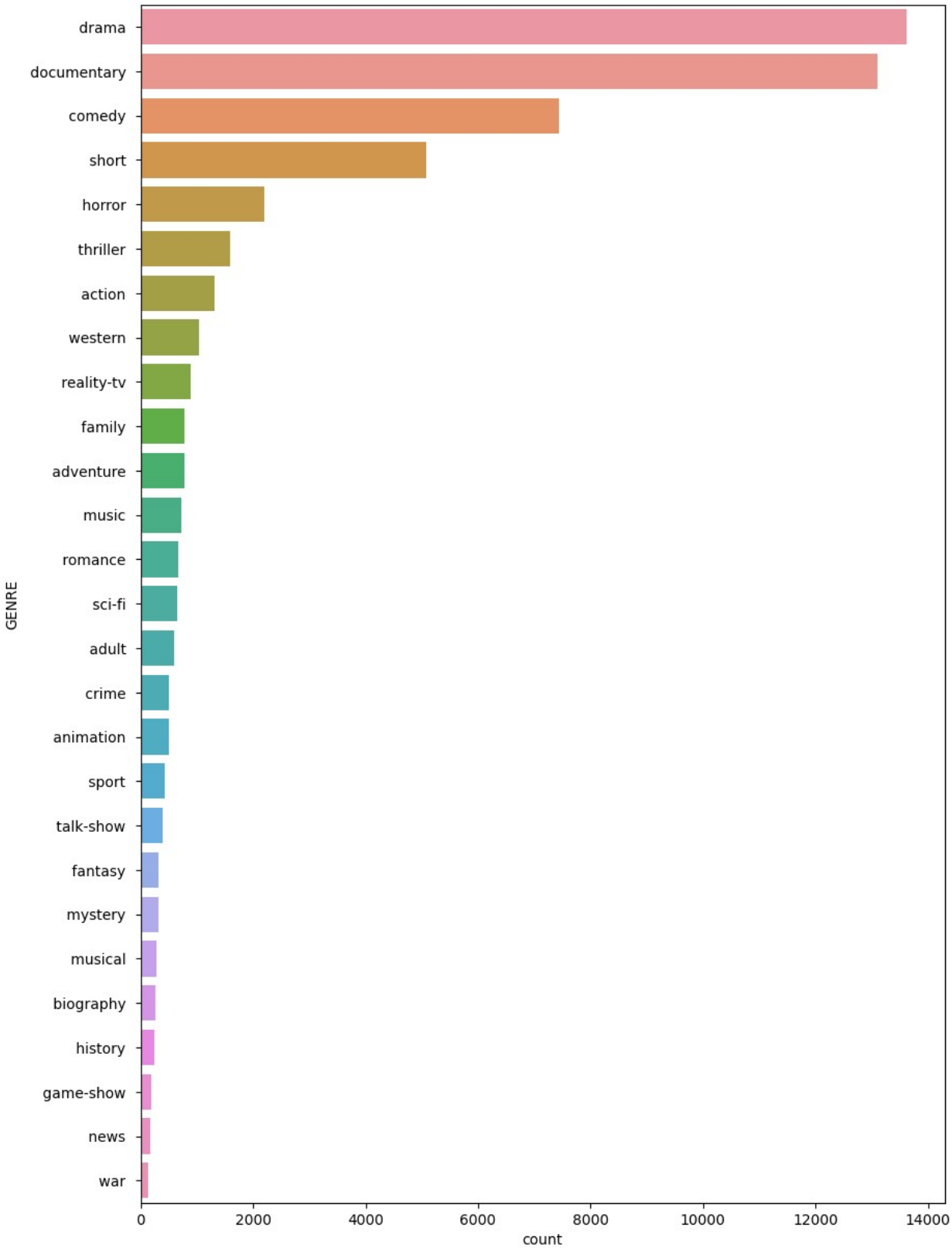
```
[54214 rows x 3 columns]
```

```
test_data.info()
```

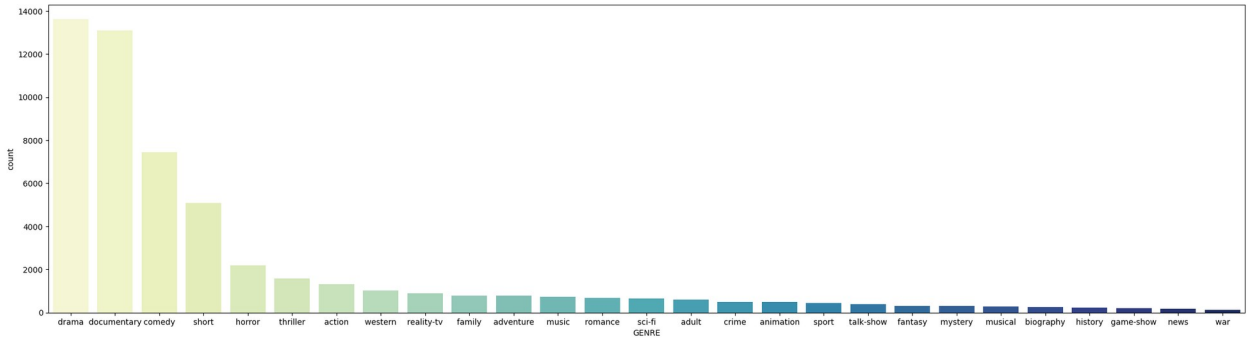
```
<class 'pandas.core.frame.DataFrame'>
Index: 54214 entries, 1 to 54214
Data columns (total 3 columns):
 #   Column       Non-Null Count  Dtype 
---  ------       --------------  ----- 
 0   ID           54214 non-null  object
 1   TITLE        54214 non-null  object
 2   DESCRIPTION  54214 non-null  object
dtypes: object(3)
memory usage: 1.7+ MB
```

EDA

```
plt.figure(figsize=(10,15))
sns.countplot(data=train_data, y="GENRE", order=
train_data["GENRE"].value_counts().index)
plt.show()
```


drama
documentary
comedy
short
horror
thriller
action
western
reality-tv
family
adventure
music
romance
sci-fi
adult
crime
animation
sport
talk-show
fantasy
mystery
musical
biography
history
game-show
news
war
GENRE
0
2000
4000
6000
8000
10000
12000
14000
count

```
plt.figure(figsize=(27,7))
sns.countplot(data=train_data, x="GENRE", order=
train_data["GENRE"].value_counts().index, palette = "YlGnBu")
plt.show()
```



Data Preprocessing

```
stemmer = LancasterStemmer()
stop_words = set(stopwords.words("english"))  # Stopwords set

def cleaning_data(text):
    text = text.lower()
    text = re.sub(r'@\S+', '', text)
    text = re.sub(r'http\S+', '', text)
    text = re.sub(r'.pic\S+', '', text)
    text = re.sub(r'[^a-zA-Z+]', ' ', text)  # Change to replace non-
characters with a space
    text = "".join([i for i in text if i not in string.punctuation])
    words = nltk.word_tokenize(text)
    # Use the predefined stop_words variable instead of redefining it
inside the function
    text = " ".join([i for i in words if i not in stop_words and
len(i) > 2])
    text = re.sub(r"\s+", " ", text).strip()  # Replace multiple
spaces with a single space
    return text

train_data["TextCleaning"] =
train_data["DESCRIPTION"].apply(cleaning_data)
test_data["TextCleaning"] =
test_data["DESCRIPTION"].apply(cleaning_data)

train_data
```

```
                                          TITLE          GENRE  \
1                    Oscar et la dame rose (2009)          drama
2                                 Cupid (1997)       thriller
```

```
3                  Young, Wild and Wonderful (1980)          adult
4                            The Secret Sin (1915)          drama
5                          The Unrecovered (2007)          drama
...                                           ...            ...
54210                             "Bonino" (1953)         comedy
54211                  Dead Girls Don't Cry (????)         horror
54212    Ronald Goedemondt: Ze bestaan echt (2008)    documentary
54213                      Make Your Own Bed (1944)         comedy
54214   Nature's Fury: Storm of the Century (2006)        history

                                                DESCRIPTION  \
1       Listening in to a conversation between his do...
2       A brother and sister with a past incestuous r...
3       As the bus empties the students for their fie...
4       To help their unemployed father make ends mee...
5       The film's title refers not only to the un-re...
...                                                   ...
54210   This short-lived NBC live sitcom centered on ...
54211   The NEXT Generation of EXPLOITATION. The sist...
54212   Ze bestaan echt, is a stand-up comedy about g...
54213   Walter and Vivian live in the country and hav...
54214   On Labor Day Weekend, 1935, the most intense ...

                                               TextCleaning
1      listening conversation doctor parents year old...
2      brother sister past incestuous relationship cu...
3      bus empties students field trip museum natural...
4      help unemployed father make ends meet edith tw...
5      film title refers recovered bodies ground zero...
...                                                  ...
54210  short lived nbc live sitcom centered bonino wo...
54211  next generation exploitation sisters kapa bay ...
54212  bestaan echt stand comedy growing facing fears...
54213  walter vivian live country difficult time keep...
54214  labor day weekend intense hurricane ever make ...

[54214 rows x 4 columns]
```

Using TF-IDF to vectorize the data

```
vectorize = TfidfVectorizer()

X_train = vectorize.fit_transform(train_data["TextCleaning"])

X_test = vectorize.transform(test_data["TextCleaning"])
```

Split the data into train data and test data

```
X = X_train
y = train_data["GENRE"]

X_train, X_test, Y_train, Y_test = train_test_split(X, y, test_size=
0.2, random_state=42)
```

Train the model

```
model = SVC()
model.fit(X_train, Y_train)
```

```
SVC()
```

```
model.score(X_train, Y_train)
```

```
0.9034608378870674
```

```
y_pred = model.predict(X_test)
y_pred
```

```
array([' comedy ', ' drama ', ' comedy ', ..., ' drama ', ' short ',
       ' horror '], dtype=object)
```

```
accuracy = accuracy_score(Y_test, y_pred)
print("Validation Accuracy:", accuracy)
```

```
Validation Accuracy: 0.5674628792769528
```