Imported libraries:

1. Pandas – for reading through the dataset and performing operations like head(), info(), describe(), isnull() etc
2. Seaborn – to plot graphs for visualization
3. Matplotlib – for enhancing the data visualization by adding title, label etc features
4. Nltk lib – Natural Language Toolkit
5. Re – regular expression – to identify patterns in a string and to work on them accordingly
6. sklearn

To load a dataset –

* move the dataset into the same folder as your code
* create a path variable and put the name of the file to be loaded
* create another data variable which will take the path variable as a attribute of the read\_csv function.
* If the dataset is in txt format specify the separator accordingly and assigns the columns present in it in the names attribute in a list

Preprocessing –

* Checked for null values from both training dataset and testing dataset

Cleaning and processing data –

* The description of the movie was to be cleaned and processed in order to guess the genre and hence a function was created that removes stop words, spaces, punctuations to clean the data.
* This was then applied on the description column and created a new column that stored the clean data

**TF-IDF -**

The TfidfVectorizer is a tool used to convert text data into numerical features that can be used by machine learning algorithms. Here’s a simplified explanation of how it works and why it's useful:

What is TF-IDF?

TF-IDF stands for Term Frequency-Inverse Document Frequency. It is a statistical measure used to evaluate the importance of a word in a document relative to a collection of documents (corpus). It combines two key ideas:

1. Term Frequency (TF): Measures how often a word appears in a document.
2. Inverse Document Frequency (IDF): Measures how important a word is across all documents. Words that appear in many documents are considered less important.