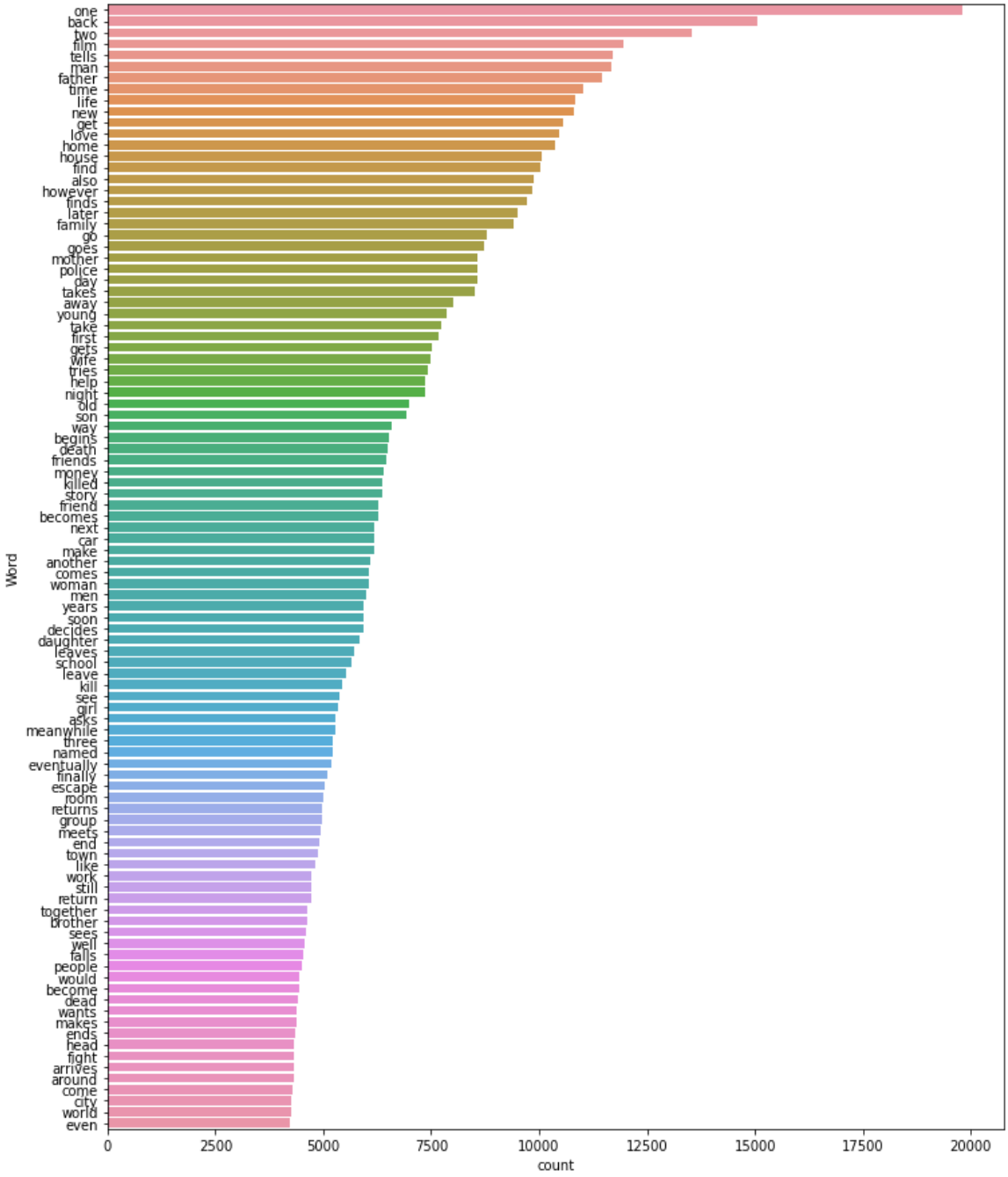
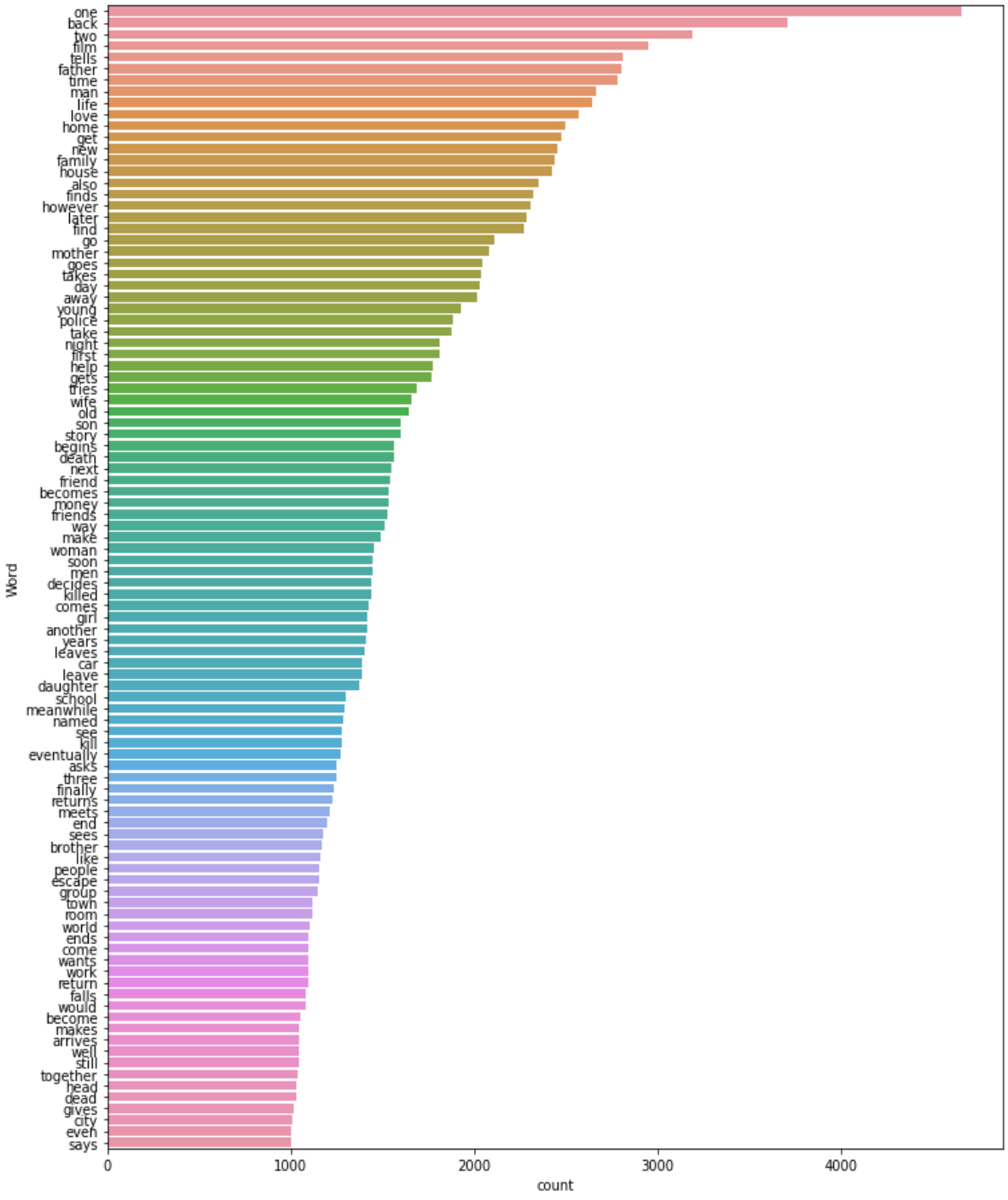
Report

The given project is about predicting a model on the given movie data. It is a multi-label classification problem i.e.; a record can have multiple labels and the number of labels per instance is not fixed. The following steps have been taken and implemented to build the model:

1. The required libraries such as **pandas**, **numpy**, **scikit-learn**, **natural language toolkit** libraries have been imported.
2. Load and read the train and test data files that have been provided.
3. I defined a function "**clean**_text" that will convert all the strings that are in the given data to the lower-case letters and remove all the punctuation marks in order to get better results.
4. Now, I have called the function that is mentioned above and passed plot and genre columns as arguments.
5. In order to find out the **frequency** of the words that has been used, a function has been developed that will help us to visualize the most frequent words that has been used. In the function, the natural language toolkit library has been used.
6. After visualizing, I came to know that most frequent words are the **stop words**. As these stop words carry less meaning and causes hindrance to develop the model, I removed the stop words from all the records in the plot column. The following graphs are the frequent words that are used in the train data and the test data after the stop words were removed:


Word
one
back
two
film
tells
man
father
time
life
new
get
love
home
house
find
also
however
finds
later
family
go
goes
mother
police
day
takes
away
young
take
first
gets
wife
tries
help
night
old
son
way
begins
death
friends
money
killed
story
friend
becomes
next
car
make
another
comes
woman
men
years
soon
decides
daughter
leaves
school
leave
kill
see
girl
asks
meanwhile
three
named
eventually
finally
escape
room
returns
group
meets
end
town
like
work
still
return
together
brother
sees
well
falls
people
would
become
dead
wants
makes
ends
head
fight
arrives
around
come
city
world
even
0
2500
5000
7500
10000
12500
15000
17500
20000
count



7. The stop words can be recognized by downloading the natural language toolkit stop words file. Based on that, the words that are present in the plot column can be matched with the library and thus removed.
8. Next, I removed all the **duplicates** that are present in the train and test data.

9. Now, I encoded the train and test data to fetch the target variables with the help of "**CounterVectorizer**" which is from natural language toolkit.
10. I used **TF-IDF** to extract the features from the train and test dataset. I have used 1000 most frequent words in the data as my features.
11. I am using the **SGD classifier** to build the model. And "**OneVsRestClassifier**" is also used in order to solve this problem as a binary relevance.
12. Later, I fit the model with the train set and predict the model on the validation set.
13. At last, I have calculated the accuracy, recall, precision and the f1 scores of the model. The results of accuracy, recall, precision and the f1 score are:

    Accuracy: `4.747871644%`
    Recall: `0.136`
    Precision: `44.001950585%`
    F1 Score: `20.823076923%`

14. The F1 score that is obtained is based on the **threshold** value "0.5". That is, the probabilities that are greater than or equal to "0.5" are converted to value "1" and the rest are converted to "0". To **improve** the model score, I have **reduced** the threshold value to "0.1" which gave the f1 score: `35.007881371%`.