In [2]:

```python
import pandas as pd
import matplotlib.pyplot as plt
import seaborn as sns
%matplotlib inline
```

In [3]:

```python
df = pd.read_csv('C://Users//SAIHARSHITHA//Downloads//movies.csv')
```

In [4]:

```python
df.head()
```

Out[4]:

	rank	title	genre	wiki_plot	imdb_plot	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	Unnamed: 10	Unnamed: 11	Unnamed: 12
0	0	The Godfather	[u' Crime', u' Drama']	On the day of his only daughter's wedding, Vit...	In late summer 1945, guests are gathered for t...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
1	1	The Shawshank Redemption	[u' Crime', u' Drama']	In 1947, banker Andy Dufresne is convicted of ...	In 1947, Andy Dufresne (Tim Robbins), a banker...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
2	2	Schindler's List	[u' Biography', u' Drama', u' History']	In 1939, the Germans move Polish Jews into the...	The relocation of Polish Jews from surrounding...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
3	3	Raging Bull	[u' Biography', u' Drama', u' Sport']	In a brief scene in 1964, an aging, overweight...	The film opens in 1964, where an older and fat...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN
4	4	Casablanca	[u' Drama', u' Romance', u' War']	It is early December 1941. American expatriate...	In the early years of World War II, December 1...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN

In [5]:

```python
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 102 entries, 0 to 101
Data columns (total 13 columns):
 #   Column       Non-Null Count  Dtype 
---  ------       --------------  ----- 
 0   rank         101 non-null    object
 1   title        101 non-null    object
 2   genre        101 non-null    object
 3   wiki_plot    101 non-null    object
 4   imdb_plot    91 non-null     object
 5   Unnamed: 5   1 non-null      object
 6   Unnamed: 6   1 non-null      object
 7   Unnamed: 7   1 non-null      object
 8   Unnamed: 8   1 non-null      object
 9   Unnamed: 9   1 non-null      object
 10  Unnamed: 10  1 non-null      object
 11  Unnamed: 11  1 non-null      object
 12  Unnamed: 12  1 non-null      object
dtypes: object(13)
memory usage: 10.5+ KB
```

In [6]:

```python
df.describe()
```

Out[6]:

	rank	title	genre	wiki_plot	imdb_plot	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	Unnamed: 10	Unnamed: 11	Unnamed: 12
count	101	101	101	101	91	1	1	1	1	1	1	1	1
unique	101	101	62	101	91	1	1	1	1	1	1	1	1
top	0	The Godfather	[u' Drama']	On the day of his only daughter's wedding, Vit...	In late summer 1945, guests are gathered for t...	but Luke and two mechanics assure him that R2...	an awards ceremony is held in a huge hall. Hu...	officers	and pilots are present. A door at the rear of...	Han	and Chewbacca. They walk down the aisle to wh...	along with several Rebel leaders and dignitar...	Luke and Han are awarded medals for bravery b...
freq	1	1	11	1	1	1	1	1	1	1	1	1	1

In [7]:

```python
df["plots"] = df["wiki_plot"].astype(str) + "\n" + df["imdb_plot"].astype(str)
```

In [8]:

```python
df.head()
```

Out[8]:

	rank	title	genre	wiki_plot	imdb_plot	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	Unnamed: 10	Unnamed: 11	Unnamed: 12	plots
0	0	The Godfather	[u' Crime', u' Drama']	On the day of his only daughter's wedding, Vit...	In late summer 1945, guests are gathered for t...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	On the day of his only daughter's wedding, Vit...
1	1	The Shawshank Redemption	[u' Crime', u' Drama']	In 1947, banker Andy Dufresne is convicted of ...	In 1947, Andy Dufresne (Tim Robbins), a banker...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	In 1947, banker Andy Dufresne is convicted of ...
2	2	Schindler's List	[u' Biography', u' Drama', u' History']	In 1939, the Germans move Polish Jews into the...	The relocation of Polish Jews from surrounding...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	In 1939, the Germans move Polish Jews into the...
3	3	Raging Bull	[u' Biography', u' Drama', u' Sport']	In a brief scene in 1964, an aging, overweight...	The film opens in 1964, where an older and fat...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	In a brief scene in 1964, an aging, overweight...
4	4	Casablanca	[u' Drama', u' Romance', u' War']	It is early December 1941. American expatriate...	In the early years of World War II, December 1...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	It is early December 1941. American expatriate...

In [9]:

```python
df["length"] = df["plots"].apply(len)
```

In [10]:

```python
df.head()
```

Out[10]:

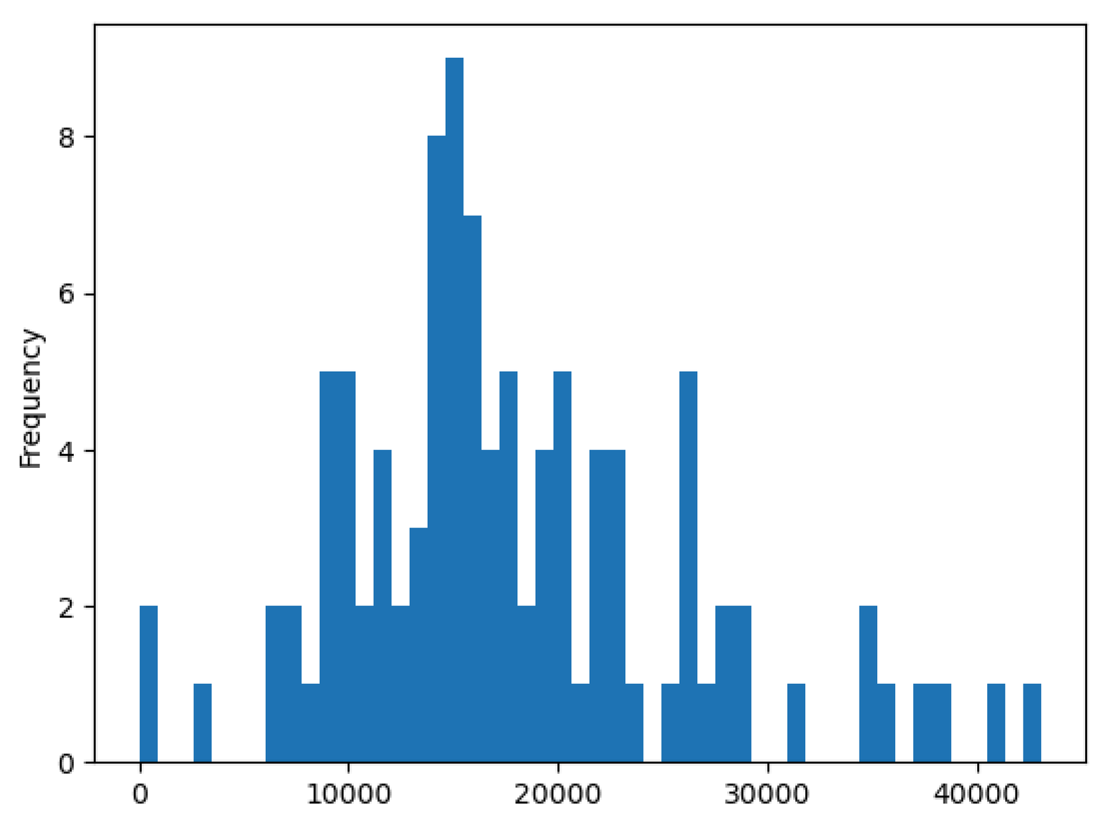
	rank	title	genre	wiki_plot	imdb_plot	Unnamed: 5	Unnamed: 6	Unnamed: 7	Unnamed: 8	Unnamed: 9	Unnamed: 10	Unnamed: 11	Unnamed: 12	plots	length
0	0	The Godfather	[u' Crime', u' Drama']	On the day of his only daughter's wedding, Vit...	In late summer 1945, guests are gathered for t...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	On the day of his only daughter's wedding, Vit...	26878
1	1	The Shawshank Redemption	[u' Crime', u' Drama']	In 1947, banker Andy Dufresne is convicted of ...	In 1947, Andy Dufresne (Tim Robbins), a banker...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	In 1947, banker Andy Dufresne is convicted of ...	22533
2	2	Schindler's List	[u' Biography', u' Drama', u' History']	In 1939, the Germans move Polish Jews into the...	The relocation of Polish Jews from surrounding...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	In 1939, the Germans move Polish Jews into the...	16008
3	3	Raging Bull	[u' Biography', u' Drama', u' Sport']	In a brief scene in 1964, an aging, overweight...	The film opens in 1964, where an older and fat...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	In a brief scene in 1964, an aging, overweight...	16108
4	4	Casablanca	[u' Drama', u' Romance', u' War']	It is early December 1941. American expatriate...	In the early years of World War II, December 1...	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	It is early December 1941. American expatriate...	14130

In [11]:

```python
df["length"].plot.hist(bins = 50)
```

Out[11]:

```
<Axes: ylabel='Frequency'>
```



In [12]:

```python
df["length"].describe()
```

Out[12]:

```
count      102.000000
mean     17811.862745
std       8272.882319
min          7.000000
25%      12849.750000
50%      16174.500000
75%      22020.750000
max      43049.000000
Name: length, dtype: float64
```

In [13]:

```python
import nltk
```

In [14]:

```python
import string
```

In [15]:

```python
from nltk.corpus import stopwords
from nltk.stem.snowball import SnowballStemmer
```

In [16]:

```python
stemmer = SnowballStemmer("english")
```

In [17]:

```python
def text_process(plot):
    no_punc = [char for char in plot if char not in string.punctuation]
    no_punc = ''.join(no_punc)
    no_punc_mod = [word for word in no_punc.split() if word.lower not in stopwords.words('english')]
    stems = [stemmer.stem(word) for word in no_punc_mod]
    return stems
```

In [18]:

```python
df['plots'].head(5).apply(text_process)
```

Out[18]:

```
0    [on, the, day, of, his, onli, daughter, wed, v...
1    [in, 1947, banker, andi, dufresn, is, convict,...
2    [in, 1939, the, german, move, polish, jew, int...
3    [in, a, brief, scene, in, 1964, an, age, overw...
4    [it, is, earli, decemb, 1941, american, expatr...
Name: plots, dtype: object
```

In [19]:

```python
from sklearn.feature_extraction.text import CountVectorizer
```

In [20]:

```python
bow_transformer = CountVectorizer(analyzer=text_process).fit(df['plots'])
```

In [21]:

```python
print(len(bow_transformer.vocabulary_))
```

```
12800
```

In [22]:

```python
plots_bow = bow_transformer.transform(df['plots'])
```

In [23]:

```python
print(plots_bow.shape)
```

```
(102, 12800)
```

In [24]:

```python
plots_bow.nnz
```

Out[24]:

```
63356
```

In [25]:

```python
from sklearn.feature_extraction.text import TfidfTransformer
```

In [26]:

```python
tfidf_transformer = TfidfTransformer().fit(plots_bow)
```

In [27]:

```python
plots_tfidf = tfidf_transformer.transform(plots_bow)
```

In [28]:

```python
print(plots_tfidf.shape)
```

```
(102, 12800)
```

In [29]:

```python
from sklearn.cluster import KMeans
```

In [30]:

```python
kmeans = KMeans(n_clusters=5)
```

In [31]:

```python
kmeans.fit(plots_tfidf)
```

```
C:\Users\SAIHARSHITHA\anaconda3\Lib\site-packages\sklearn\cluster\_kmeans.py:1412: FutureWarning: The default value of `n_init` will change from 10 to 'auto' in 1.4. Set the value of `n_init` explicitly to suppress the warning
  super()._check_params_vs_input(X, default_n_init=10)
```

Out[31]:

```
▾ KMeans
KMeans(n_clusters=5)
```

In [32]:

```python
clusters = kmeans.labels_.tolist()
```

In [33]:

```python
df["cluster"] = clusters
```

In [34]:

```python
df['cluster'].value_counts()
```

Out[34]:

```
cluster
2    67
3    26
1     5
0     2
4     2
Name: count, dtype: int64
```

In [35]:

```python
from sklearn.metrics.pairwise import cosine_similarity
```

In [36]:

```python
similarity_distance = 1 - cosine_similarity(plots_tfidf)
```

In [37]:

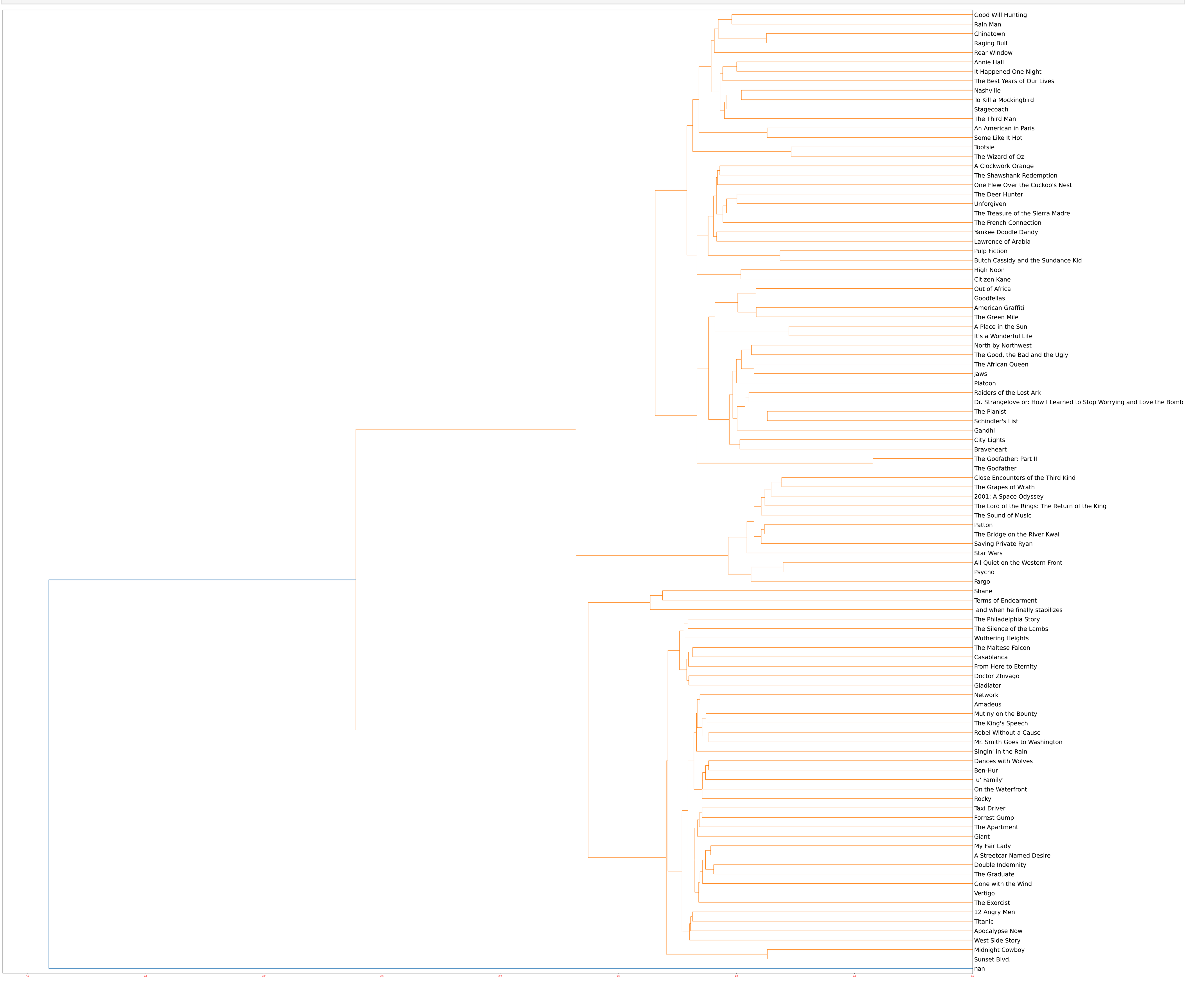
```python
from scipy.cluster.hierarchy import linkage, dendrogram
```

In [39]:

```python
mergings = linkage(similarity_distance, method="complete")

dendrogram_plot = dendrogram(mergings, orientation='left', labels=[title for title in df["title"]], leaf_font_size=27)

fig = plt.gcf()
p = [label.set_color('r') for label in plt.gca().get_xmajorticklabels()]
fig.set_size_inches(80,80)
plt.show()
```



In [50]:

```python
# Function to get 5 similar movies based on cluster and similarity distance
def get_similar_movies(movie_title):
    # Find the cluster ID for the given movie
    cluster_id = df.loc[df['title'] == movie_title, 'cluster'].values[0]

    # Get the indices of movies in the same cluster
    similar_movie_indices = df.loc[df['cluster'] == cluster_id].index

    # Calculate similarity distances between the given movie and other movies in the cluster
    distances = cosine_similarity(plots_tfidf[df.loc[df['title'] == movie_title].index], plots_tfidf[similar_movie_indices]).flatten()

    # Get the indices of the 5 most similar movies (excluding the input movie)
    most_similar_indices = distances.argsort()[:5]

    # Select the titles of these similar movies
    similar_movies = df.loc[similar_movie_indices[most_similar_indices], 'title'].tolist()

    return similar_movies

# Example usage:
movie_title = input("Enter a movie title: ")
similar_movies = get_similar_movies(movie_title)
print("Similar movies:")
for movie in similar_movies:
    print(movie)
```

```
Enter a movie title: The Philadelphia Story
Similar movies:
Terms of Endearment
Midnight Cowboy
From Here to Eternity
Casablanca
The Silence of the Lambs
```

In [ ]: