

Step 1 : Loading and Exploring the Dataset

```
In [1]: import pandas as pd
# Loading the dataset
df = pd.read_csv(r"C:\Users\Ritik\Downloads\IMDB.csv")
df.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 2199 entries, 0 to 2198
Data columns (total 8 columns):
#   Column      Non-Null Count  Dtype
---  -
0   Unnamed: 0   2199 non-null  int64
1   movie_id     2199 non-null  object
2   movie_name   2199 non-null  object
3   year         2134 non-null  object
4   genre        2199 non-null  object
5   overview     2199 non-null  object
6   director     2199 non-null  object
7   cast         2199 non-null  object
dtypes: int64(1), object(7)
memory usage: 137.6+ KB
```

```
In [31]: df.head(11)
```

```
Out[31]:
```

	Unnamed: 0	movie_id	movie_name	year	genre	overview	director	cast	Processed_Plot
0	0	tt15354916	Jawan	2023	Action, Thriller	A high-octane action thriller which outlines t...	Atlee	Shah Rukh Khan, Nayanthara, Vijay Sethupathi, ...	A high-octane action thriller which outline th...
1	1	tt15748830	Jaane Jaan	2023	Crime, Drama, Mystery	A single mother and her daughter who commit a ...	Sujoy Ghosh	Kareena Kapoor, Jaideep Ahlawat, Vijay Varma, ...	A single mother and her daughter who commit a ...
2	2	tt11663228	Jailer	2023	Action, Comedy, Crime	A retired jailer goes on a manhunt to find his...	Nelson Dilipkumar	Rajinikanth, Mohanlal, Shivarajkumar, Jackie S...	A retired jailer go on a manhunt to find his s...
3	3	tt14993250	Rocky Aur Rani Kii Prem Kahaani	2023	Comedy, Drama, Family	Flamboyant Punjabi Rocky and intellectual Beng...	Karan Johar	Ranveer Singh, Alia Bhatt, Dharmendra, Shabana...	Flamboyant Punjabi Rocky and intellectual Beng...
4	4	tt15732324	OMG 2	2023	Comedy, Drama	An unhappy civilian asks the court to mandate ...	Amit Rai	Pankaj Tripathi, Akshay Kumar, Yami Gautam, Pa...	An unhappy civilian asks the court to mandate ...
5	5	tt18266472	Sukhee	2023	Drama	Much to the dismay of her husband, a middle-cl...	Sonal Joshi	Shilpa Shetty Kundra, Amit Sadh, Chaitannya Ch...	Much to the dismay of her husband, a middle-cl...
6	6	tt18561736	The Great Indian Family	2023	Family	Ved Vyas Tripathi, aka Bhajan Kumar, is a devo...	Vijay Krishna Acharya	Alka Amin, Bhuvan Arora, Manushi Chhillar, Sri...	Ved Vyas Tripathi, aka Bhajan Kumar, is a devo...
7	7	tt3691740	The BFG	2016	Adventure, Family, Fantasy	An orphan little girl befriends a benevolent g...	Steven Spielberg	Mark Rylance, Ruby Barnhill, Penelope Wilton, ...	An orphan little girl befriends a benevolent g...
8	8	tt12844910	Pathaan	2023	Action, Adventure, Thriller	An Indian agent races against a doomsday clock...	Siddharth Anand	Shah Rukh Khan, Deepika Padukone, John Abraham...	An Indian agent race against a doomsday clock ...
9	9	tt15464390	Mastaney	2023	Action, Drama, History	Set in 1739, Nadar Shah's undefeated army was ...	Sharan Art	Tarsem Jassar, Simi Chahal, Gurpreet Ghuggi, K...	Set in 1739, Nadar Shah's undefeated army wa a...
10	10	tt1187043	3 Idiots	2009	Comedy, Drama	Two friends are searching for their long lost ...	Rajkumar Hirani	Aamir Khan, Madhavan, Mona Singh, Sharman Joshi	Two friend are searching for their long lost c...

Step 2 : Text Preprocessing : Tokenization, Lemmatizing

```
In [3]: import nltk
from nltk.corpus import stopwords
from nltk.tokenize import word_tokenize
#from nltk.stem import PorterStemmer
from nltk.stem import WordNetLemmatizer

nltk.download('stopwords')
nltk.download('punkt')
nltk.download('wordnet')

# Creating a Lemmatizer
lemmatizer = WordNetLemmatizer()

# Function for Lemmatization
def lemmatize_text(text):
    return ' '.join([lemmatizer.lemmatize(word) for word in text.split()])

# Applying Lemmatization to the 'overview' column
df['Processed_Plot'] = df['overview'].apply(lemmatize_text)
```

```
[nltk_data] Downloading package stopwords to
[nltk_data] C:\Users\Ritik\AppData\Roaming\nltk_data...
[nltk_data] Package stopwords is already up-to-date!
[nltk_data] Downloading package punkt to
[nltk_data] C:\Users\Ritik\AppData\Roaming\nltk_data...
[nltk_data] Package punkt is already up-to-date!
[nltk_data] Downloading package wordnet to
[nltk_data] C:\Users\Ritik\AppData\Roaming\nltk_data...
[nltk_data] Package wordnet is already up-to-date!
```

In [34]: *#displaying the Processed_Plot*
df.head(6)

Out[34]:

	Unnamed: 0	movie_id	movie_name	year	genre	overview	director	cast	Processed_Plot
0	0	tt15354916	Jawan	2023	Action, Thriller	A high-octane action thriller which outlines t...	Atlee	Shah Rukh Khan, Nayanthara, Vijay Sethupathi, ...	A high-octane action thriller which outline th...
1	1	tt15748830	Jaane Jaan	2023	Crime, Drama, Mystery	A single mother and her daughter who commit a ...	Sujoy Ghosh	Kareena Kapoor, Jaideep Ahlawat, Vijay Varma, ...	A single mother and her daughter who commit a ...
2	2	tt11663228	Jailer	2023	Action, Comedy, Crime	A retired jailer goes on a manhunt to find his...	Nelson Dilipkumar	Rajinikanth, Mohanlal, Shivarajkumar, Jackie S...	A retired jailer go on a manhunt to find his s...
3	3	tt14993250	Rocky Aur Rani Kii Prem Kahaani	2023	Comedy, Drama, Family	Flamboyant Punjabi Rocky and intellectual Beng...	Karan Johar	Ranveer Singh, Alia Bhatt, Dharmendra, Shabana...	Flamboyant Punjabi Rocky and intellectual Beng...
4	4	tt15732324	OMG 2	2023	Comedy, Drama	An unhappy civilian asks the court to mandate ...	Amit Rai	Pankaj Tripathi, Akshay Kumar, Yami Gautam, Pa...	An unhappy civilian asks the court to mandate ...
5	5	tt18266472	Sukhee	2023	Drama	Much to the dismay of her husband, a middle-cl...	Sonal Joshi	Shilpa Shetty Kundra, Amit Sadh, Chaitannya Ch...	Much to the dismay of her husband, a middle-cl...

Step 3 : Vectorize Text Data

```
In [16]: from sklearn.feature_extraction.text import TfidfVectorizer

tfidf_vectorizer = TfidfVectorizer(max_features=30000)
tfidf_matrix = tfidf_vectorizer.fit_transform(df['Processed_Plot'])
```

Step 4 : Importing KMeans & Creating Clusters to plot it

```
In [5]: from sklearn.cluster import KMeans
from sklearn.feature_extraction.text import TfidfVectorizer
from sklearn.decomposition import PCA
import matplotlib.pyplot as plt

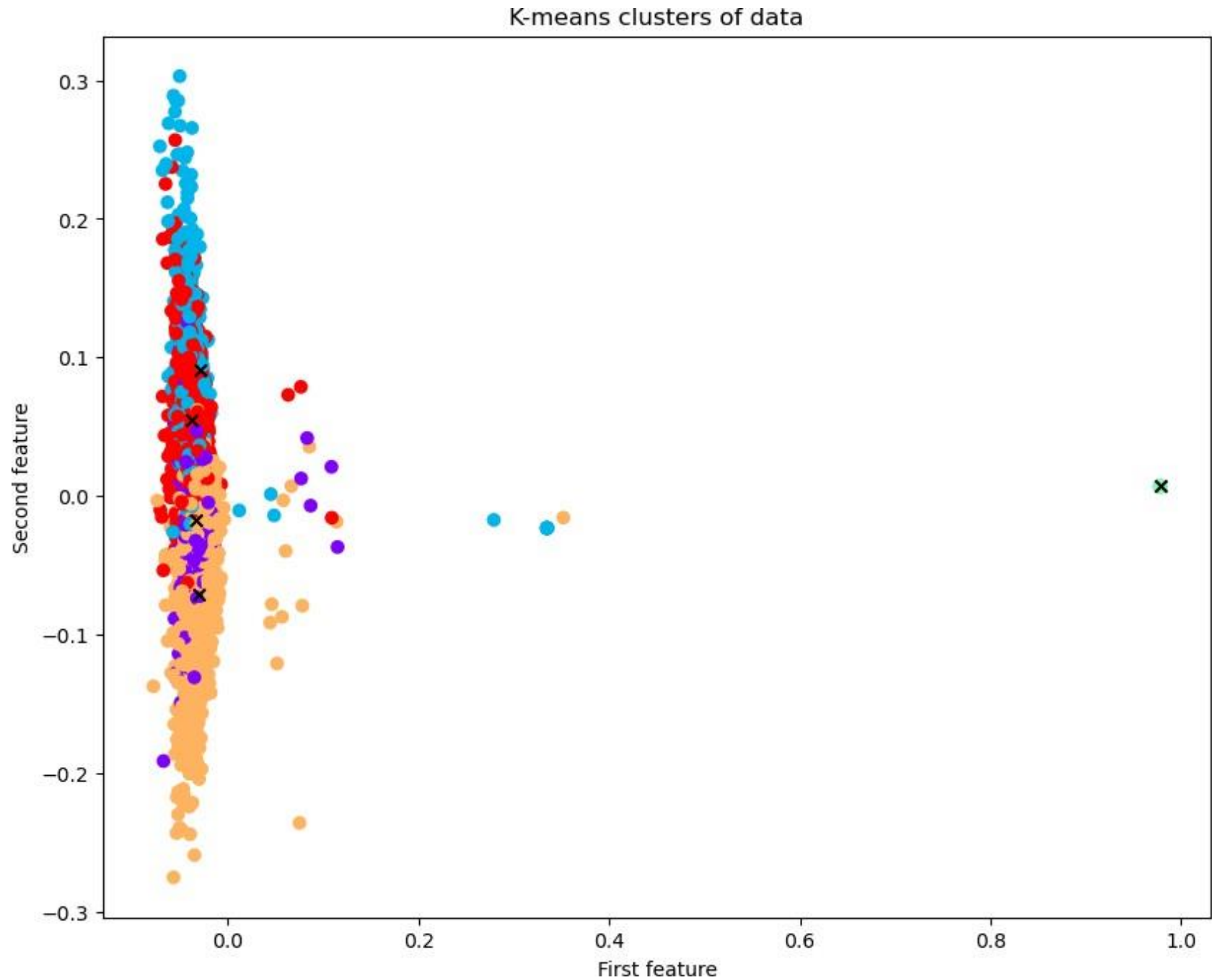
kmeans = KMeans(n_clusters=5, random_state=92)
kmeans.fit(tfidf_matrix)

y_pred = kmeans.predict(tfidf_matrix)
centers_pred = kmeans.cluster_centers_

pca = PCA(n_components=2, random_state=92)
tfidf_matrix_reduced = pca.fit_transform(tfidf_matrix.toarray())
centers_reduced = pca.transform(centers_pred)

plt.figure(figsize=(10,8))
plt.scatter(tfidf_matrix_reduced[:,0], tfidf_matrix_reduced[:,1], c=y_pred, cmap='rainbow')
plt.scatter(centers_reduced[:,0], centers_reduced[:,1], c='black', marker='x')
plt.xlabel('First feature')
plt.ylabel('Second feature')
plt.title('K-means clusters of data')
plt.show()
```

```
C:\Users\Ritik\anaconda3\anaconda\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)
```



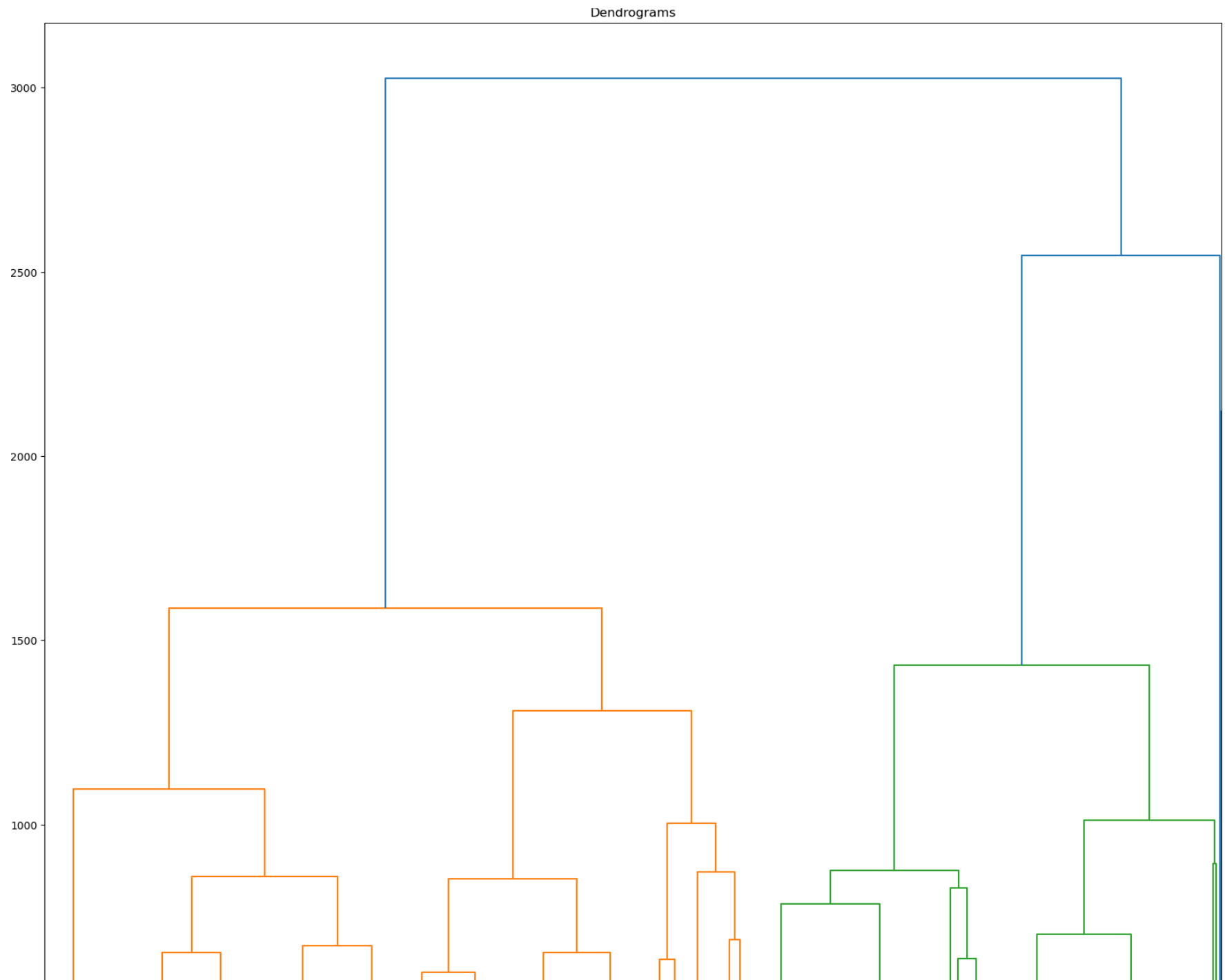
Step 5 : Compute Similarity

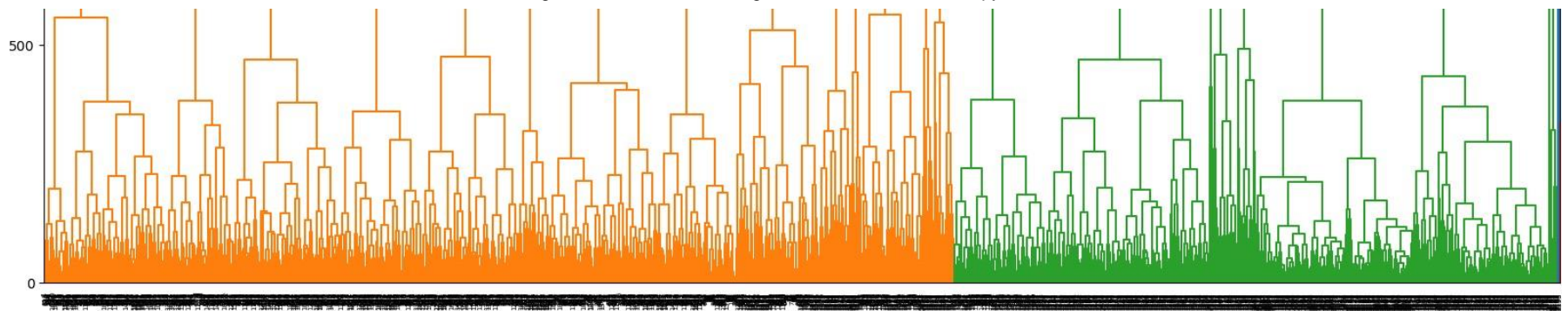
```
In [6]: from sklearn.metrics.pairwise import cosine_similarity
similarity_matrix = cosine_similarity(tfidf_matrix, tfidf_matrix)
# Convert the similarity matrix to a DataFrame
similarity_df = pd.DataFrame(similarity_matrix, index=df['movie_name'], columns=df[''])
```

Step 6 : Importing Matplot, Linkage

```
In [7]: from scipy.cluster import hierarchy
from scipy.cluster.hierarchy import linkage, dendrogram
import numpy as np
import matplotlib.pyplot as plt
# Creating mergings matrix
mergings = linkage(similarity_matrix, method='complete')

# Creating Dendrogram for our data
Z = hierarchy.linkage(mergings, method='average')
plt.figure()
fig = plt.gcf()
fig.set_size_inches(20, 20)
plt.title("Dendrograms")
# Dendrogram plotting using Linkage matrix
dendrogram = hierarchy.dendrogram(Z)
```





Step 7 : Finding Similar Movies

For checking similarity user we will enter the movie name of different types to check we get results or not. Below given names can be tested as all are of different type :-

Examples : K.G.F: Chapter 1 , Ramayana: The Legend of Prince Rama , M.S. Dhoni: The Untold Story , Jab Harry Met Sejal

K.G.F: Chapter 1

Ramayana: The Legend of Prince Rama

M.S. Dhoni: The Untold Story

Jab Harry Met Sejal

```
In [11]: #function for checking the movies similar
def find_similar_movies(movie_title, similarity_matrix, num_recommendations=10):
    similar_movies = similarity_matrix[movie_title].sort_values(ascending=False)[1:num_recommendations+1]
    return similar_movies

# Find similar movies for a given title
#Note : Entering the correct movie title is important so can use the above given movie list
movie_title = input("Enter the movie name : ")
similar_movies = find_similar_movies(movie_title, similarity_df)

# Displaying the result
print(f"Movies similar to '{movie_title}':\n{similar_movies}")
```

Enter the movie name : Ramayana: The Legend of Prince Rama

Movies similar to 'Ramayana: The Legend of Prince Rama':

movie_name

Adipurush 0.310326

Sampoorna Ramayana 0.225799

Ramayana: The Epic 0.224849

Asoka 0.224611

The Intern 0.182134

Goliyon Ki Rasleela Ram-Leela 0.166881

Dharam Sankat Mein 0.164505

Panipat 0.143879

Vinaya Vidheya Rama 0.142840

King of Kotha 0.142306

Name: Ramayana: The Legend of Prince Rama, dtype: float64