Mini-Project - Movie Recommendations: The Hood

M Srinivasan - IMT2021058
Sankalp Kothari - IMT2021028
Siddharth Kothari - IMT2021019
Munagala Kalyan Ram - IMT2021023

Overview - What have we implemented??

Implemented Methods -

1) KMeans with SVD

- Reduced SVD and randomized SVD
- Random initialisation and KMeans++ initialisation

2) KModes with SVD

3) Collaborative Filtering

- User-User and Item-Item filtering with different similarity metrics

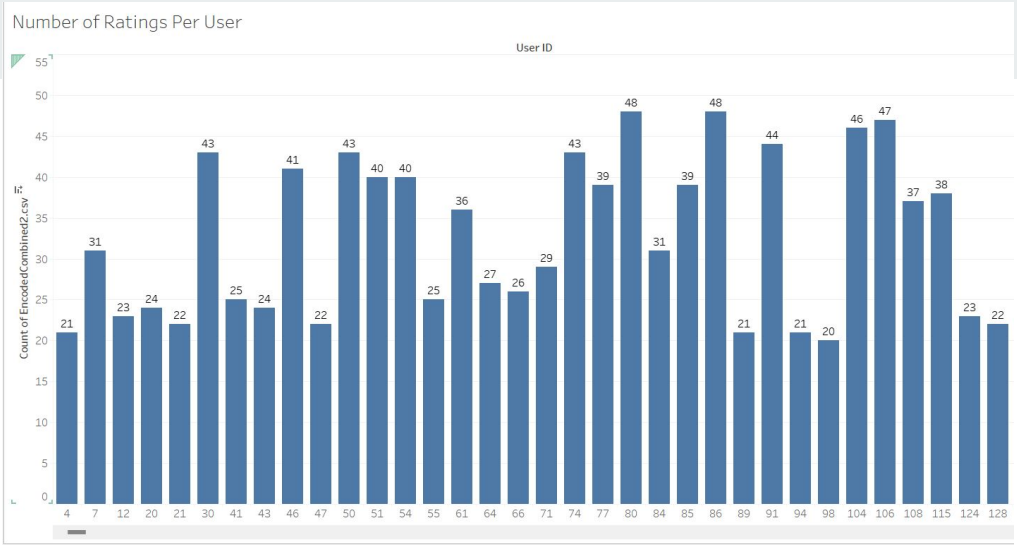
Biggest Issue

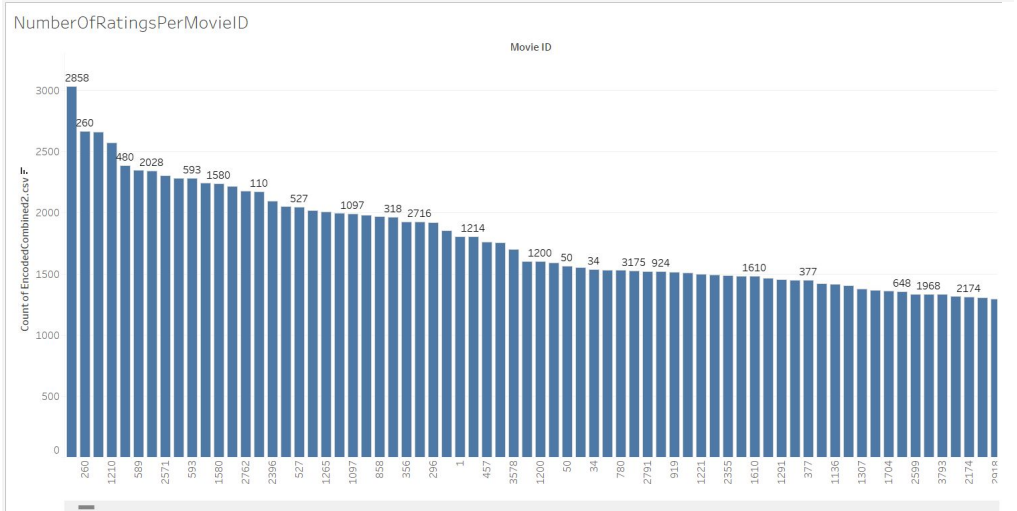
Run time given the complexity of the algorithms and the size of the dataset.

EDA / preprocessing

Users and their Ratings

- We plotted a graph of users and the number of movies they have rated and watched.
- We could see that more than 25% users had seen and rated more than 50 movies

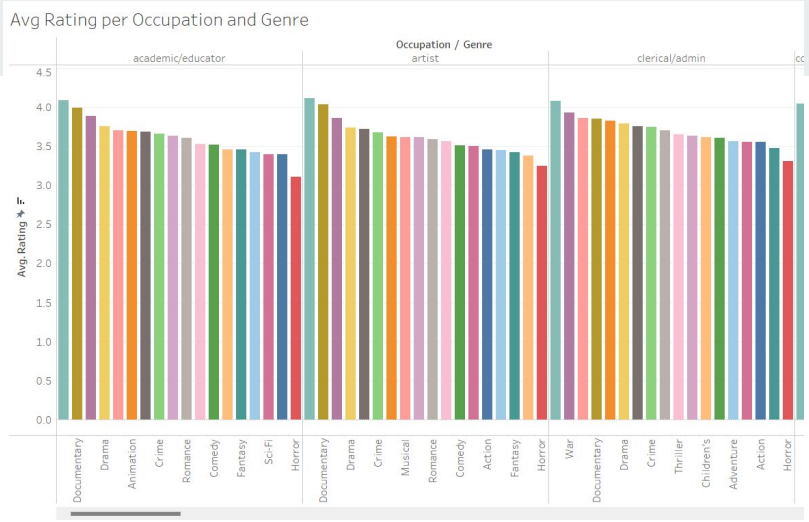


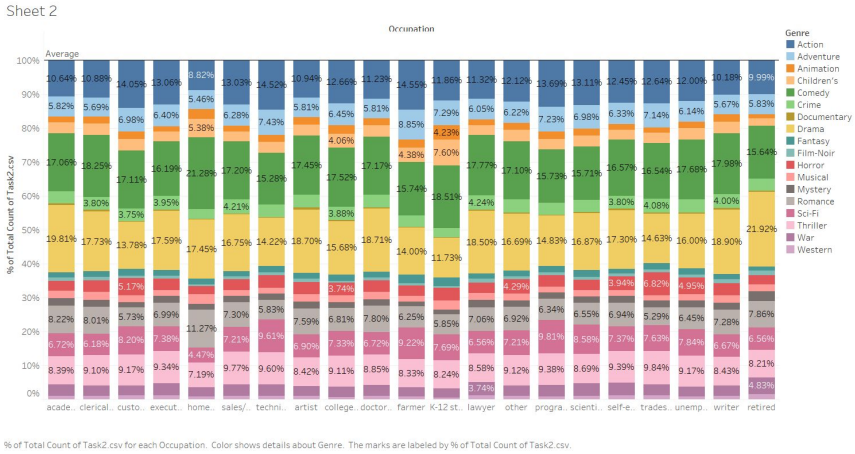


EDA / preprocessing

Genres and Occupation Ratings

- For each occupation we plotted the average ratings for all genres to see if there is any relation.
- There was a uniform distribution across all the occupations and genres, so nothing could be inferred from this.

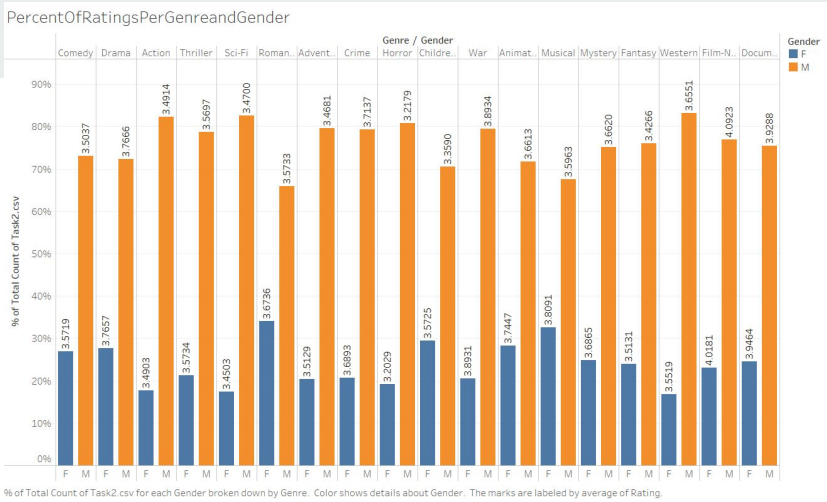


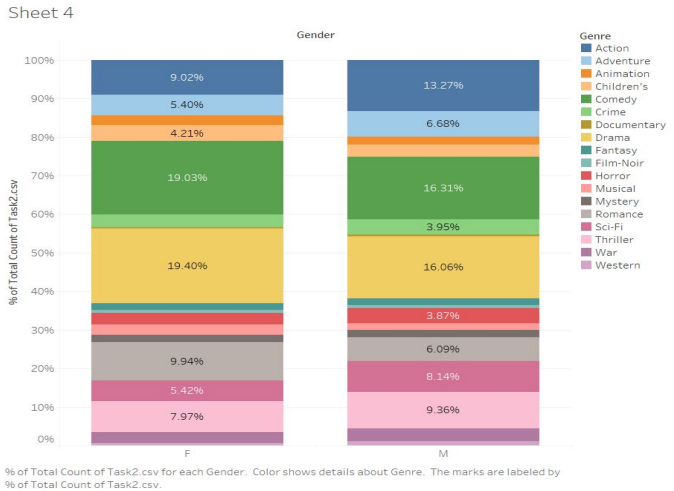


EDA / preprocessing

Genres Ratings along with Gender

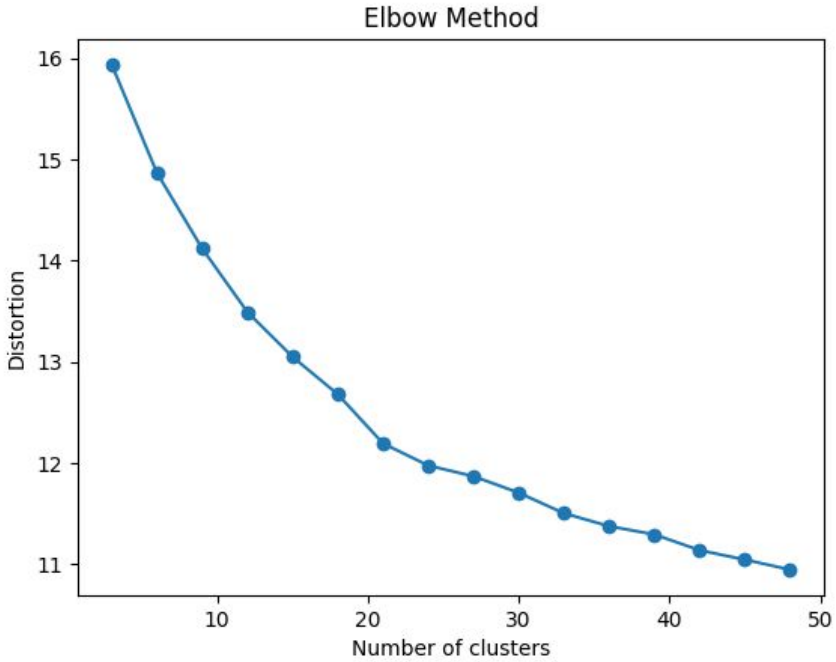
- We plotted the movie ratings according to each genre and further according to each gender.
- We couldn't infer much from it, apart from what is usually expected from the society(males prefer watching Action and Horror/War whereas Females prefer Musical and RomComs).





KMeans with SVD

- We first performed SVD for the user-genre matrix with smaller datasets and then ran it on the complete dataset. Filled the null values of a genre with the average rating of the genre
- For the user queried, the cluster he belongs to is found and the movies that the user hasn't watched and are common / highly rated in that cluster are selected and then recommended to the user.
- The image on the right is the output of the elbow method, from which we got the optimal number of centers.



SVD Followed By Kmeans

- The new user with userID- 7000 was added and the results for this user were as follows.

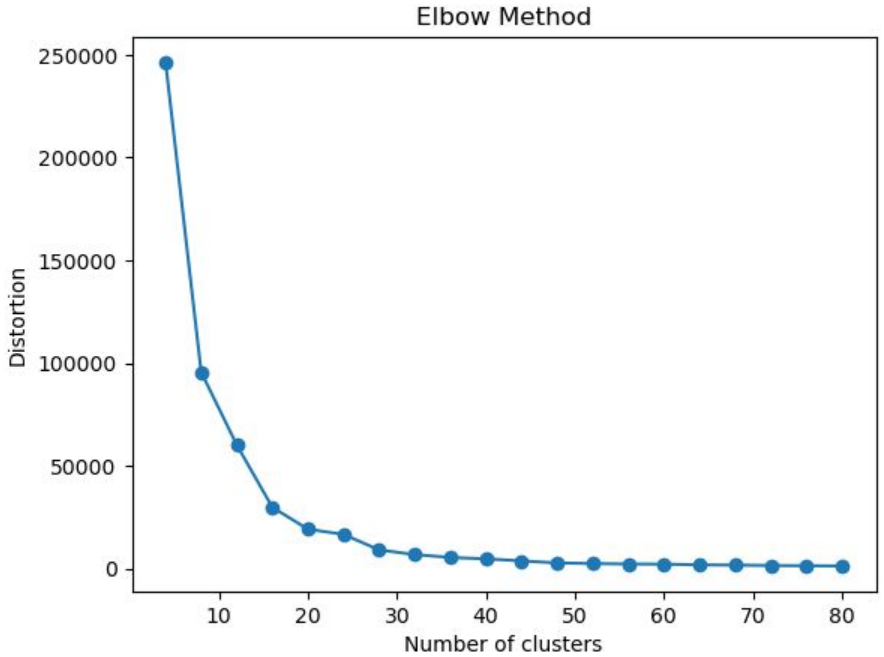
```
Movies Watched by User : 7000
Title : Big Green, The (1995) , MovieID : 54 , Rating : 5.0 , Genres : Children's|Comedy
Title : Last Detail, The (1973) , MovieID : 3200 , Rating : 5.0 , Genres : Comedy|Drama
Title : Leaving Las Vegas (1995) , MovieID : 25 , Rating : 4.0 , Genres : Drama|Romance
Title : Aliens (1986) , MovieID : 1200 , Rating : 4.0 , Genres : Action|Sci-Fi|Thriller|War
Title : Bio-Dome (1996) , MovieID : 65 , Rating : 3.0 , Genres : Comedy
Title : Heidi Fleiss: Hollywood Madam (1995) , MovieID : 99 , Rating : 2.0 , Genres : Documentary
Title : Free Willy (1993) , MovieID : 455 , Rating : 2.0 , Genres : Adventure|Children's|Drama
Title : Men Cry Bullets (1997) , MovieID : 2980 , Rating : 1.0 , Genres : Drama
```

We can see that he has seen a lot of Drama and Comedy movies, so the predictions from KMeans were also the same

```
Recommended movies for User 7000
Title : Shall We Dance? (Shall We Dansu?) (1996) , MovieID : 1537 , Rating : 5.0 , Genres : Comedy
Title : Run Lola Run (Lola rennt) (1998) , MovieID : 2692 , Rating : 5.0 , Genres : Action|Crime|Romance
Title : King of Masks, The (Bian Lian) (1996) , MovieID : 2609 , Rating : 5.0 , Genres : Drama
Title : Three Colors: White (1994) , MovieID : 308 , Rating : 5.0 , Genres : Drama
Title : Patton (1970) , MovieID : 1272 , Rating : 5.0 , Genres : Drama|War
```

SVD Followed By Kmeans++

- Similar to the previous approach, but in this case, the initialization of the centroids was according to the KMeans++ Paper.
- For the user queried, the cluster he belongs to is found and the movies that are common / highly rated in that cluster is selected and then recommended to the user (the ones he hasn't watched).
- The image on the right is the output from the elbow method for Kmeans++ from which we take the optimal number of clusters.



SVD Followed By Kmeans++

- The new user with userID- 7000 was added and the results for this user were as follows.

```
Movies Watched by User : 7000
Title : Big Green, The (1995) , MovieID : 54 , Rating : 5.0 , Genres : Children's|Comedy
Title : Last Detail, The (1973) , MovieID : 3200 , Rating : 5.0 , Genres : Comedy|Drama
Title : Leaving Las Vegas (1995) , MovieID : 25 , Rating : 4.0 , Genres : Drama|Romance
Title : Aliens (1986) , MovieID : 1200 , Rating : 4.0 , Genres : Action|Sci-Fi|Thriller|War
Title : Bio-Dome (1996) , MovieID : 65 , Rating : 3.0 , Genres : Comedy
Title : Heidi Fleiss: Hollywood Madam (1995) , MovieID : 99 , Rating : 2.0 , Genres : Documentary
Title : Free Willy (1993) , MovieID : 455 , Rating : 2.0 , Genres : Adventure|Children's|Drama
Title : Men Cry Bullets (1997) , MovieID : 2980 , Rating : 1.0 , Genres : Drama
```

We can see that he has seen a lot of Drama and Comedy movies, so the predictions from KMeans++ were also the same.

```
Recommended movies for User 7000
Title : Vie est belle, La (Life is Rosey) (1987) , MovieID : 771 , Rating : 5.0 , Genres : Comedy|Drama
Title : Nightwatch (1997) , MovieID : 1355 , Rating : 5.0 , Genres : Horror|Thriller
Title : Man Who Knew Too Much, The (1956) , MovieID : 2183 , Rating : 5.0 , Genres : Thriller
Title : Conversation, The (1974) , MovieID : 3730 , Rating : 5.0 , Genres : Drama|Mystery
Title : Roman Holiday (1953) , MovieID : 916 , Rating : 5.0 , Genres : Comedy|Romance
```

KModes followed by SVD on the clusters

- First performed the KModes on the user-ratings data and then grouped the users. We find the cluster the user is part of and then do the SVD on the ratings for that people in that cluster.
- For the user queried, the cluster he belongs to is found and the movies that are common / highly rated in that cluster is selected and then recommended to the user (the ones he hasn't watched).
- The elbow method for this took a lot of time to run because of the size of the dataset, so we choose the centre very close to the ones done for SVD_Kmeans, but the results we got from that were very good so we continued with the same number of clusters.

KModes followed by SVD on the clusters

The user has seen very less ‘War’ movies but has rated all of them very high. He has watched a lot of Drama movies and has rated them pretty good, the KModes model has recommended a lot of Drama movies to him.


```
Recommended movies for User  1
Title : Seven Samurai (The Magnificent Seven) (Shichinin no samurai) (1954) , MovieID : 2019 , Rating :
4.606315789499392 , Genres : Action|Drama
Title : Godfather, The (1972) , MovieID : 858 , Rating : 4.595789473801125 , Genres : Action|Crime|Drama
Title : Silence of the Lambs, The (1991) , MovieID : 593 , Rating : 4.589473684995558 , Genres : Drama|Thriller
Title : Shawshank Redemption, The (1994) , MovieID : 318 , Rating : 4.586842105157931 , Genres : Drama
Title : Usual Suspects, The (1995) , MovieID : 50 , Rating : 4.570526315819115 , Genres : Crime|Thriller
```

Novel Ideas in the SVD - KMeans approach

1. Using randomized svd (multiplies a gaussian matrix of a certain size on the initial matrix to obtain a sampled matrix which is smaller in size and hence runs faster)
2. Using the matrix which has userid's as row and genres as columns (with the value for each cell being the average of the ratings provided by that user to movies in that genre).
3. Adding the user's latent variable representation that is related to the genres (obtained by running svd on the user genre matrix) with the user details (age,gender,occupation).
4. When creating combined csv file, raw join resulted in file of size 95 MB, but after changing the data types from int64 to int16 or int8 based on the column and removing the title and zip-code column, along with removing null values, reduced to 47 MB. This was also used in Collaborative Filtering.
5. For Kmeans we tried various initialization methods such as random and the approach mentioned in the Kmeans++ paper.

Collaborative Filtering

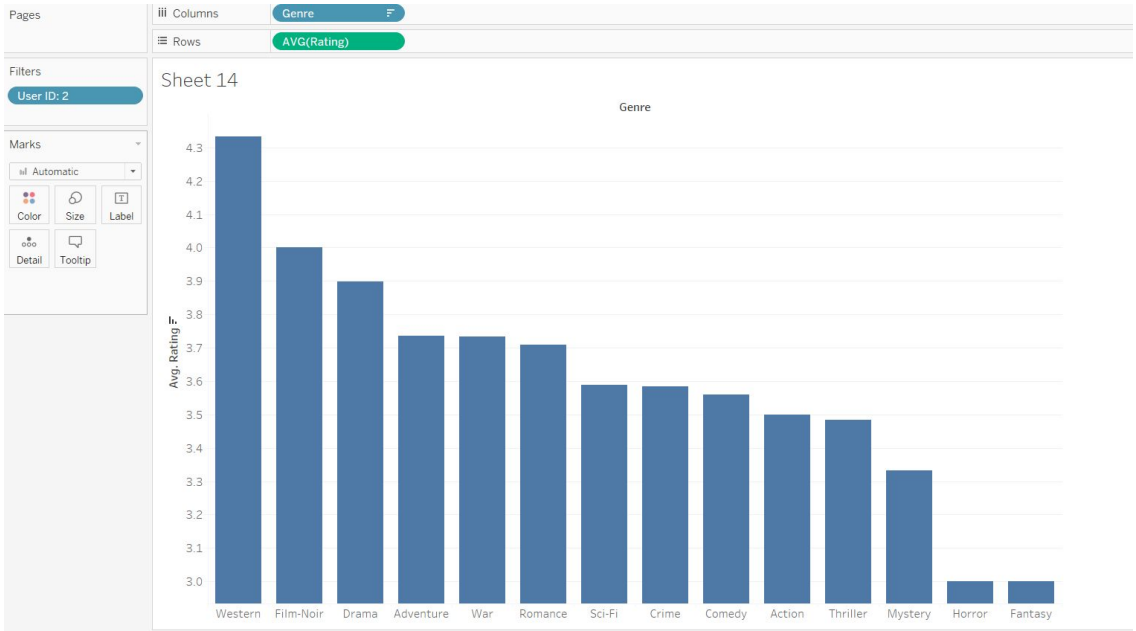
We implemented the following 4 similarity metrics -

1. Cosine Similarity
2. Pearson's Correlation Coefficient (Unweighted)
3. Weighted Pearson's Correlation Coefficient (IDF as the weights)
4. Weighted Pearson's Correlation Coefficient (Variance as the weights)

For Cosine Similarity and Unweighted PCC, we implemented with both User User and Item Item Based Similarity.

For Weighted versions, we implemented User User only. So a total of 6 combinations.

User 2



1. This user has watched a lot of Drama movies, and it is expected that he should be recommended more of them (as he has rated them highly as well).
2. Western and Film-Noir should be recommended to him, which as we shall see later, are not recommended at all.
3. He is the dummy user for all the collaborative filtering results.

Collaborative Filtering - User User

Cosine Similarity

```
Title : Apple, The (Sib) (1998) , MovieID : 2503 , Genres : Drama
Title : Gate of Heavenly Peace, The (1995) , MovieID : 787 , Genres : Documentary
Title : Hour of the Pig, The (1993) , MovieID : 578 , Genres : Drama|Mystery
Title : Jar, The (Khomreh) (1992) , MovieID : 758 , Genres : Drama
Title : I Am Cuba (Soy Cuba/Ya Kuba) (1964) , MovieID : 3245 , Genres : Drama
Title : Follow the Bitch (1998) , MovieID : 1830 , Genres : Comedy
Title : Schlafes Bruder (Brother of Sleep) (1995) , MovieID : 989 , Genres : Drama
Title : Foreign Student (1994) , MovieID : 572 , Genres : Drama
Title : Mamma Roma (1962) , MovieID : 557 , Genres : Drama
Title : Song of Freedom (1936) , MovieID : 3382 , Genres : Drama
```

- It takes more amount of time to train this. (around 200 mins for PCC and 150 mins for Cosine Similarity)
- Predictions are done for User with User ID 2 as a test.

Pearson's Correlation Coefficient

```
Title : Gate of Heavenly Peace, The (1995) , MovieID : 787 , Genres : Documentary
Title : I Am Cuba (Soy Cuba/Ya Kuba) (1964) , MovieID : 3245 , Genres : Drama
Title : Leather Jacket Love Story (1997) , MovieID : 1851 , Genres : Drama|Romance
Title : Identification of a Woman (Identificazione di una donna) (1982) , MovieID : 1360 , Genres : Drama
Title : Wirey Spindell (1999) , MovieID : 3228 , Genres : Comedy
Title : Trois (2000) , MovieID : 3291 , Genres : Thriller
Title : Foreign Student (1994) , MovieID : 572 , Genres : Drama
Title : Zachariah (1971) , MovieID : 3236 , Genres : Western
Title : Low Life, The (1994) , MovieID : 730 , Genres : Drama
Title : Loves of Carmen, The (1948) , MovieID : 3209 , Genres : Drama
```

Collaborative Filtering - Item Item

Cosine Similarity

```
Title : Anna (1996) , MovieID : 1316, Genres : Drama
Title : Soft Toilet Seats (1999) , MovieID : 3290, Genres : Comedy
Title : Legal Deceit (1997) , MovieID : 1709, Genres : Thriller
Title : Loves of Carmen, The (1948) , MovieID : 3209, Genres : Drama
Title : Eden (1997) , MovieID : 1815, Genres : Drama
Title : Nueba Yol (1995) , MovieID : 133, Genres : Comedy|Drama
Title : Roula (1995) , MovieID : 642, Genres : Drama
Title : Silence of the Palace, The (Saimt el Qusur) (1994) , MovieID : 127, Genres : Drama
Title : Chain of Fools (2000) , MovieID : 3323, Genres : Comedy|Crime
Title : Song of Freedom (1936) , MovieID : 3382, Genres : Drama
```

- It takes lesser amount of time to train this. (around 120 mins for PCC and 90 mins for Cosine Similarity)
- Predictions are done for User with User ID 2 as a test.

Pearson's Correlation Coefficient

```
Title : Pawnbroker, The (1965) , MovieID : 3789, Genres : Drama
Title : Cinema Paradiso (1988) , MovieID : 1172, Genres : Comedy|Drama|Romance
Title : Rain Man (1988) , MovieID : 1961, Genres : Drama
Title : E.T. the Extra-Terrestrial (1982) , MovieID : 1097, Genres : Children's|Drama|Fantasy|Sci-Fi
Title : My Fair Lady (1964) , MovieID : 914, Genres : Musical|Romance
Title : King and I, The (1956) , MovieID : 2565, Genres : Musical
Title : Good Will Hunting (1997) , MovieID : 1704, Genres : Drama
Title : Shawshank Redemption, The (1994) , MovieID : 318, Genres : Drama
Title : Schindler's List (1993) , MovieID : 527, Genres : Drama|War
Title : Jeanne and the Perfect Guy (Jeanne et le garon formidable) (1998) , MovieID : 2591, Genres : Comedy|Romance
```

Weighted Collaborative Filtering

Variance as weights

```
Title : Believers, The (1987) , MovieID : 1332, Genres : Horror|Thriller
Title : Birds, The (1963) , MovieID : 1333, Genres : Horror
Title : Blob, The (1958) , MovieID : 1334, Genres : Horror|Sci-Fi
Title : Blood Beach (1981) , MovieID : 1335, Genres : Action|Horror
Title : Body Parts (1991) , MovieID : 1336, Genres : Horror
Title : Body Snatcher, The (1945) , MovieID : 1337, Genres : Horror
Title : Bram Stoker's Dracula (1992) , MovieID : 1339, Genres : Horror|Romance
Title : Bride of Frankenstein (1935) , MovieID : 1340, Genres : Horror
Title : Candyman (1992) , MovieID : 1342, Genres : Horror
Title : Contender, The (2000) , MovieID : 3952, Genres : Drama|Thriller
```

- It takes lesser amount of time to train this. (around 120 mins for variance and IDF)
- Predictions are done for User with User ID 2 as a test.

IDF as weights

```
Title : Even Dwarfs Started Small (Auch Zwerge haben klein angefangen) (1971) , MovieID : 3202, Genres : Drama
Title : Identification of a Woman (Identificazione di una donna) (1982) , MovieID : 1360, Genres : Drama
Title : Trois (2000) , MovieID : 3291, Genres : Thriller
Title : Low Life, The (1994) , MovieID : 730, Genres : Drama
Title : Foreign Student (1994) , MovieID : 572, Genres : Drama
Title : Zachariah (1971) , MovieID : 3236, Genres : Western
Title : Wirey Spindell (1999) , MovieID : 3228, Genres : Comedy
Title : Crude Oasis, The (1995) , MovieID : 821, Genres : Romance
Title : Little Indian, Big City (Un indien dans la ville) (1994) , MovieID : 641, Genres : Comedy
Title : Loves of Carmen, The (1948) , MovieID : 3209, Genres : Drama
```

Novel Ideas in Collaborative Filtering

1. There wasn't much scope of implementing new ideas in the Collaborative Filtering approach. Hence, we just implemented all possible combinations of algorithms.
2. One new thing we tried out was running Collaborative Filtering with weights as the Variance of the ratings for a movie, instead of the IDF approach.
3. The rationale was to provide less weightage to universally good or universally bad movies (they should not affect the similarity metric between 2 users). Variance was a good way to capture this, hence we used it. It gave more or less the same results as IDF.

Attention areas - Some Additional Information during Demo

Risk 1

- The Collaborative filtering and PCC take a lot of time to run as compared to the SVD-KMeans approach. Hence we have saved some results to be shown.

Risk 2

- The KModes-SVD one will also take a considerable amount of time to run, hence we have saved some of the results for this as well.

Bibliography

- Numpy Documentation - https://numpy.org/doc/
- Pandas Documentation - https://pandas.pydata.org/docs/index.html
- Getting eigenvalues and eigenvectors from scratch - https://jamesmccaffrey.wordpress.com/2023/12/22/eigenvalues-and-eigenvectors-from-scratch-using-python/
- MovieLens 1M Dataset - https://grouplens.org/datasets/movielens/1m/
- Randomized SVD - https://gregorygundersen.com/blog/2019/01/17/randomized-svd/
- Kmeans - https://github.com/tugot17/K-Means-Algorithm-From-Scratch/blob/master/k-means.py
- Medium article https://asdkazmi.medium.com/ai-movies-recommendation-system-with-clustering-based-k-means-algorithm-f04467e02fcd

Thank you !