Movie recommendation system with clustering algorithms using Movie lens dataset

GROUP 4

Ayesha Siddiqua

Shusmoy Chowdhury

16th November 2023

Missouri State UNIVERSITY

Outline

- Introduction
- Problem Specification
- Background
- Data Preprocessing
- Clustering Methodology
- Results
- Conclusion

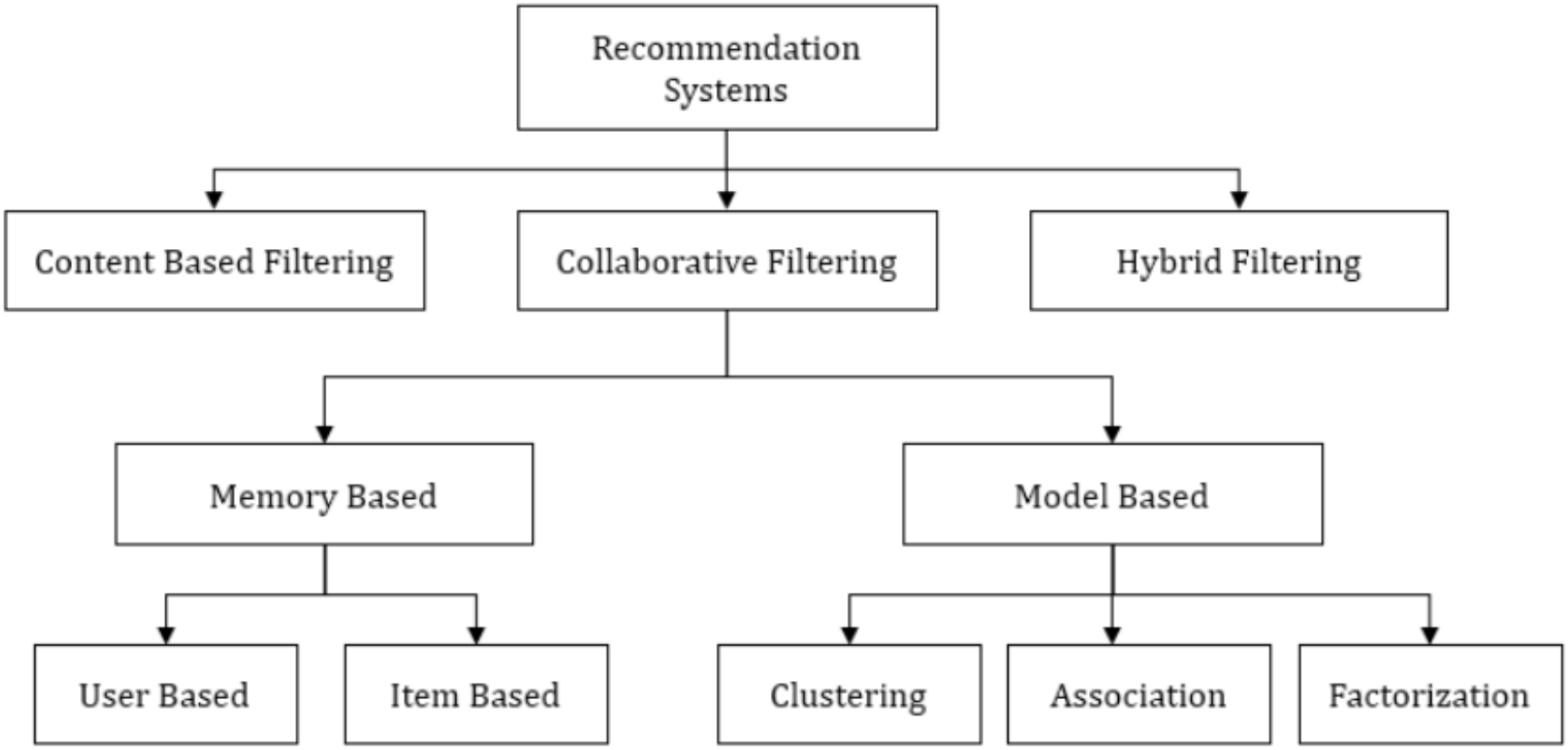
Introduction

- The Movie recommendation system is Based on users' past ratings and observed behaviors.
- filters and predicts only those movies that a corresponding user is most likely to want to watch
- The system will be created using clustering algorithms
- The algorithms will run on MovieLens dataset

Problem Specification

- Cluster the similar kinds of movies based on the movie features
- Provide better suggestions of movies based on users' choice
- Maintain user satisfaction by suggesting similar movies
- Compare the recommendation results of various clustering algorithms

Background


Recommendation Systems
Content Based Filtering
Collaborative Filtering
Hybrid Filtering
Memory Based
Model Based
User Based
Item Based
Clustering
Association
Factorization

Dataset

HTTPS://GROUPLENS.ORG/DATASETS/MOVIELENS/LATEST/

- Created by 330975 users between January 09, 1995 and July 20, 2023
- Contains 33832162 ratings and 2328315 tag applications across 86537 movies
- Movies.csv file contains movieId, title and genres
- Tags.csv file contains userId, movieId, tag and timestamp
- Ratings.csv file contains userId, movieId, rating and timestamp

Data Preprocessing

JOIN DATASETS

- Calculated the average rating and user count for individual movies.
- Modify tag dataframe to group tags of each movie
- Merged the dataframe with movies and tag dataset

```
+-------+--------------------+--------------------+---------+------------------+--------------------+
|movieId|               title|              genres|UserCount|     AverageRating|                Tags|
+-------+--------------------+--------------------+---------+------------------+--------------------+
|      1|    Toy Story (1995)|Adventure|Animati...|    76813|3.8935076093890357|[match, girl, fri...|
|      2|      Jumanji (1995)|Adventure|Childre...|    30209|3.2781786884703235|[bridge, friendsh...|
|      3|Grumpier Old Men ...|      Comedy|Romance|    15820|3.1712705436156763|[old, CLV, good s...|
|      4|Waiting to Exhale...|Comedy|Drama|Romance|     3028|2.8683949801849407|[CLV, single moth...|
|      5|Father of the Bri...|              Comedy|    15801|3.0769571546104677|[father, confiden...|
```

Data Preprocessing

DATA TRANSFORMATION

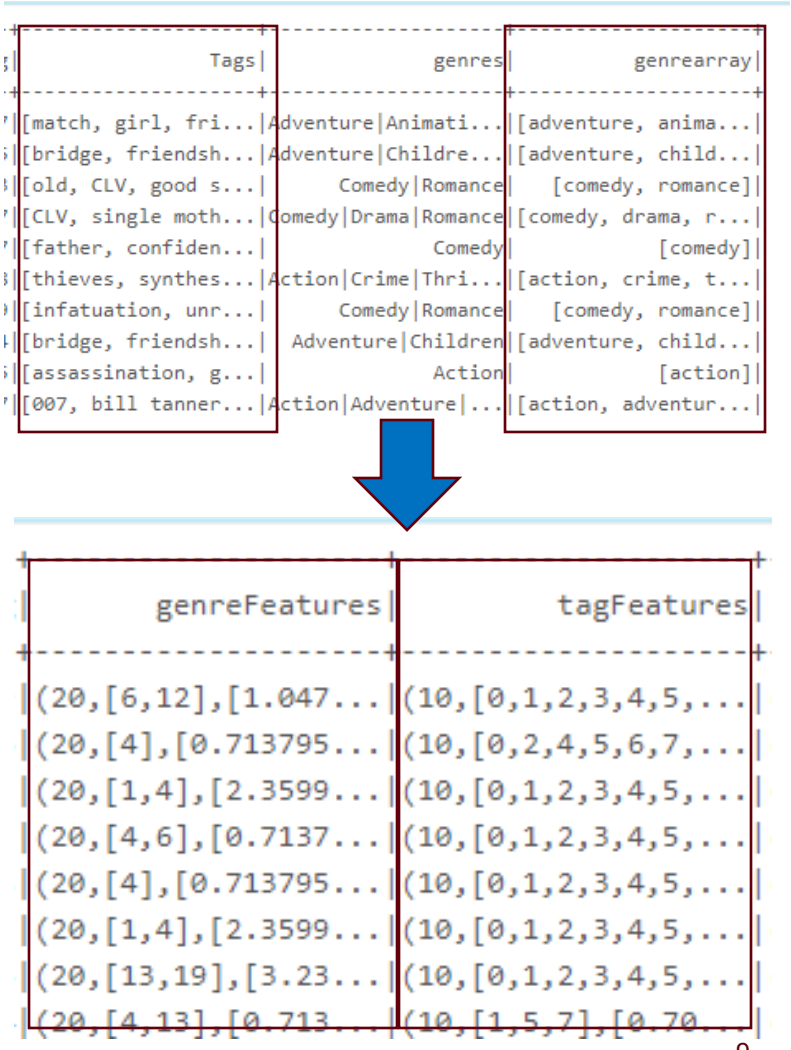
- Transformed the genre column to set of words using RegexTokenizer

genres	movieId	words
Adventure|Animation|Children|Comedy|Fantasy	1	[adventure, animation, children, comedy, fantasy]
Adventure|Children|Fantasy	2	[adventure, children, fantasy]
Comedy|Romance	3	[comedy, romance]
Comedy|Drama|Romance	4	[comedy, drama, romance]
Comedy	5	[comedy]
Action|Crime|Thriller	6	[action, crime, thriller]
Comedy|Romance	7	[comedy, romance]
Adventure|Children	8	[adventure, children]
Action	9	[action]
Action|Adventure|Thriller	10	[action, adventure, thriller]

Data Preprocessing

DATA TRANSFORMATION

- Split the tags and genres into a set of words.
- Used hastingtf to transform the words into double values.



Tags	genres	genrearray
[match, girl, fri...]	Adventure|Animati...	[adventure, anima...]
[bridge, friendsh...]	Adventure|Childre...	[adventure, child...]
[old, CLV, good s...]	Comedy|Romance	[comedy, romance]
[CLV, single moth...]	Comedy|Drama|Romance	[comedy, drama, r...]
[father, confiden...]	Comedy	[comedy]
[thieves, synthes...]	Action|Crime|Thri...	[action, crime, t...]
[infatuation, unr...]	Comedy|Romance	[comedy, romance]
[bridge, friendsh...]	Adventure|Children	[adventure, child...]
[assassination, g...]	Action	[action]
[007, bill tanner...]	Action|Adventure|...	[action, adventur...]

genreFeatures	tagFeatures
(20,[6,12],[1.047...	(10,[0,1,2,3,4,5,...
(20,[4],[0.713795...	(10,[0,2,4,5,6,7,...
(20,[1,4],[2.3599...	(10,[0,1,2,3,4,5,...
(20,[4,6],[0.7137...	(10,[0,1,2,3,4,5,...
(20,[4],[0.713795...	(10,[0,1,2,3,4,5,...
(20,[1,4],[2.3599...	(10,[0,1,2,3,4,5,...
(20,[13,19],[3.23...	(10,[0,1,2,3,4,5,...
(20,[4,13],[0.713...	(10,[1,5,7],[0.70...

9

Data Preprocessing

FEATURE EXTRACTION

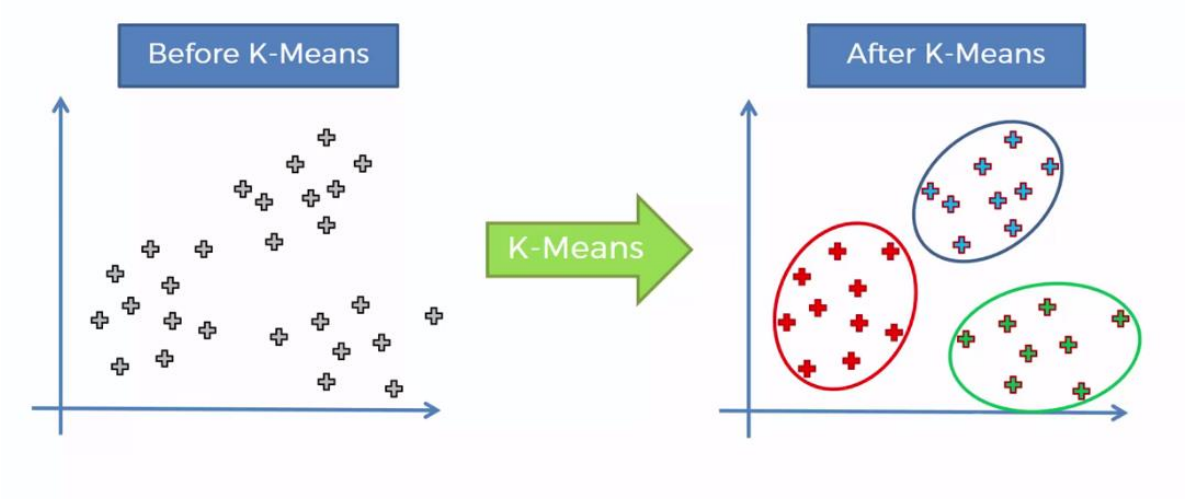
Used vectorassembler to merge multiple feature columns(genres, tags, average rating, user count) into one single feature col

```
------------------------------------------------------------------------------------------------------------------------+
|features
|
+------------------------------------------------------------------------------------------------------------------------------------
------------------------------------------------------------------------------------------------------------------------+
|(32,[0,7,13,21,22,23,24,25,26,27,28,29,30,31],[4715.0,1.0477876515859816,0.9977800042094694,14.453127287160152,15.508238386540583,21.297160583557744,14.43840573923044,10
4,13.747437807867417,10.05554549998418,11.978152227450025,24.63387442889018,19.084089019664674,2.683881230116649])                                                      |
|(32,[0,5,21,23,25,26,27,28,29,30,31],[3063.0,0.7137955447148369,1.256793677144361,0.687005180114766,1.7378323864539,1.9639196868382025,0.6284715937490113,1.409194379700
896913939,1.4136362236788647,3.604309500489716])                                                                                                                        |
|(32,[0,2,5,21,22,23,24,25,26,27,28,29,30,31],[1941.0,2.359925312753744,0.7137955447148369,1.8851905157165416,1.4098398533218712,3.435025900573829,1.9688735098950 6,2.317
3092797912254683,1.8854147812470339,0.7045971898500014,2.7370971587655757,2.120454335518297,3.42091705306543])                                                          |
|(32,[0,5,7,21,22,23,24,25,26,27,28,29,30,31],[3497.0,0.7137955447148369,1.0477876515859816,1.8851905157165416,2.1147597799828066,0.687005180114766,1.96887350989506,2.896
4,3.927839373676405,2.513886374996045,3.522985949250007,2.7370971587655757,2.120454335518297,4.038318558764655])                                                          |
|(32,[0,5,21,22,23,24,25,26,27,28,29,30,31],[12067.0,0.7137955447148369,3.1419841928609027,6.344279339948421,2.748020720459064,2.62516467986008,1.1585549243026,4.58247926
86374996045,1.4091943797000028,2.7370971587655757,0.7068181118394323,3.235021132012928])                                                                                |
```

Missouri State UNIVERSITY

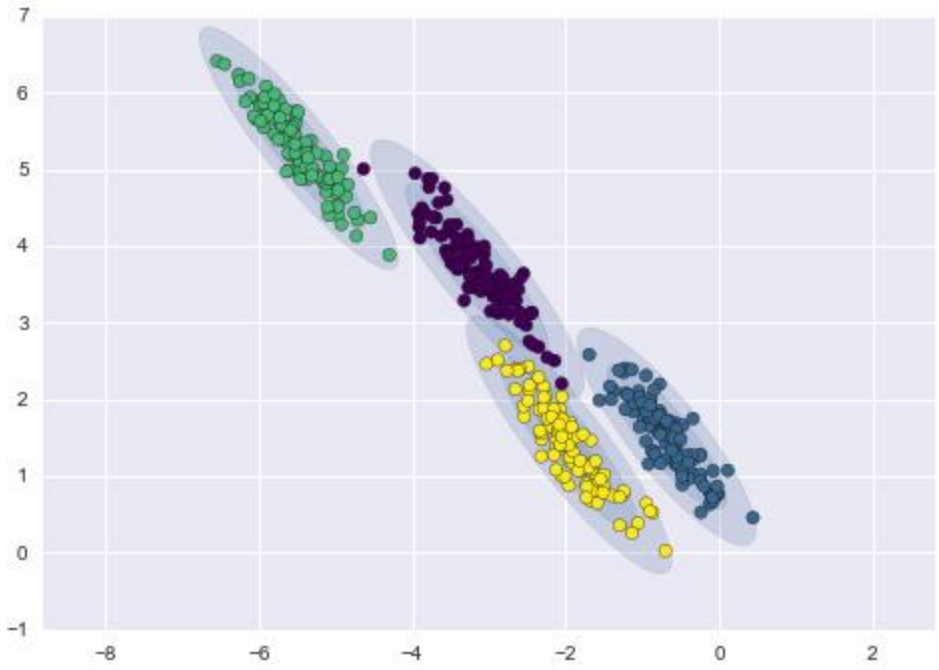
Clustering Models

K MEANS



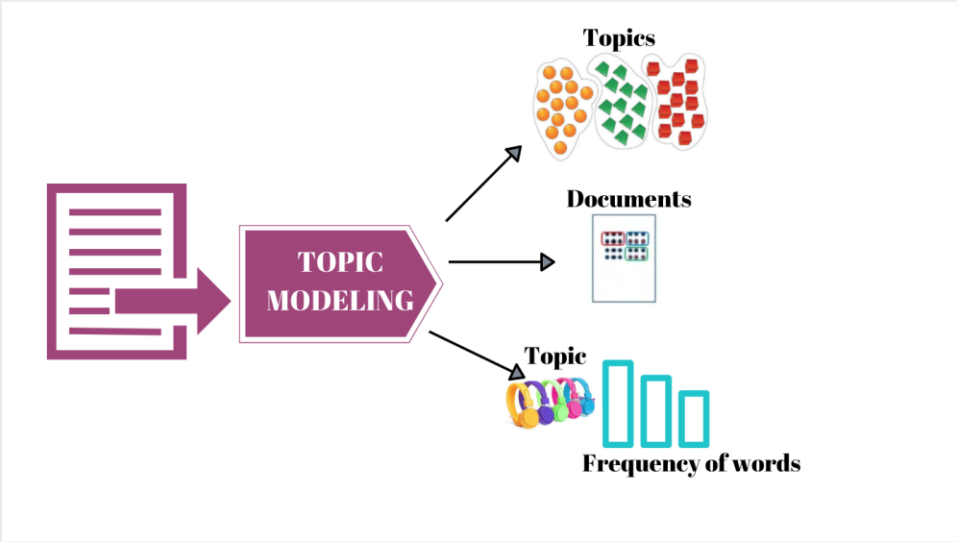
Clustering Models

GAUSSIAN MIXTURE MODEL



Clustering Models

LATENT DIRICHLET ALLOCATION



Clustering Models

RECOMMENDATION CRITERIA

Movies are filtered from the same cluster based on

1. Maximum Average rating
2. Maximum number of user rated the movies

Results

CLUSTER SIZE

K Means

prediction	count
0	40814
1	41
2	229
3	4
4	14
5	133
6	30
7	669
8	1802
9	439
10	202
11	89
12	5
13	52
14	268
15	3913
16	2
17	344
18	10
19	1094

GMM

prediction	count
0	20204
6	29950

LDA

prediction	count
0	96
2	1472
3	2202
5	5132
6	65
8	533
9	2438
11	2372
12	25497
13	2219
14	9
15	1353
17	2021
18	4745

Results

MOVIE RECOMMENDATION FOR A SINGLE USER

K Means

title	AverageRating	UserCount
Parasite (2019)	4.3299459633841435	12399
Lives of Others, ...	4.201409789323618	12626
Spider-Man: Into ...	4.192053284336242	10885
Cinema Paradiso (...	4.123029556650247	12180
Manchurian Candid...	4.07504873294347	10773
Knives Out (2019)	4.058669623059867	11275
African Queen, Th...	4.044937975190076	12495
Raging Bull (1980)	4.036668142245832	13554
Manhattan (1979)	4.01723984715187	11253
Hoop Dreams (1994)	4.0028130594152245	11731

GMM

title	AverageRating	UserCount
Awaken (2013)	5.0	3
Love, Kennedy (2017)	5.0	3
Placebo: Soulmate...	5.0	2
The Brooklyn Bank...	5.0	2
Hart to Hart: Til...	5.0	2
Christmas on Salv...	5.0	2
War Arrow (1954)	5.0	2
The Fallen of Wor...	5.0	2
Nico the Unicorn ...	5.0	2
This is a Hijack ...	5.0	1
The Sandwich Man ...	5.0	1
Plato's Reality M...	5.0	1
Asphalt Angels	5.0	1
Tony 10 (2012)	5.0	1

LDA

title	AverageRating	UserCount
Planet Earth II (...	4.451739343459089	2041
Planet Earth (2006)	4.448092868988391	3015
Band of Brothers ...	4.423985890652557	2835
James Acaster: Co...	4.421052631578948	19
Shawshank Redempt...	4.416792045528881	122296
The Work of Direc...	4.395833333333333	24
Cosmos	4.3432	625
Come From Away (2...	4.342105263157895	19
Parasite (2019)	4.3299459633841435	12399
Godfather, The (1...	4.32660258119567	75004

Results

SILHOUETTE COEFFICIENT

Algorithm Name	SILHOUETTE COEFFICIENT
K Means	**0.9029066662575397**
GMM	**0.22617731117331444**
LDA	**-0.5828395886832082**

Conclusion

- Explored different clustering algorithm for movie recommendation
- Different clustering models provide different size of clusters
- From the evaluation matrix, K means provides better results compared to others.
- Different feature extraction methods can be explored in the future extension of this work.

Thank You