Movie Recommendation System using Spark

Venkata Bharath Malapati

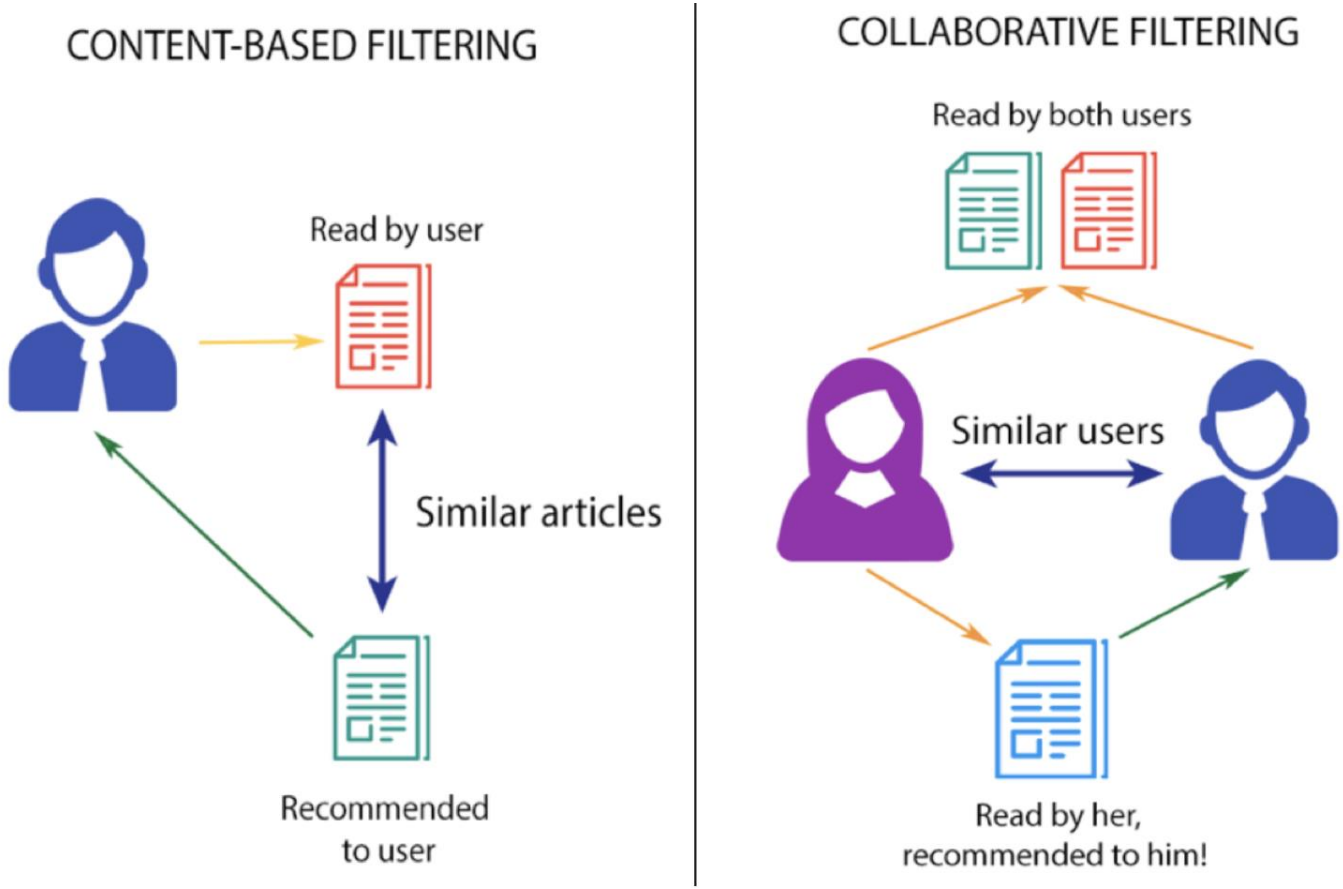
Purna Sai Pushkal kalipindi

Recommended Systems

Recommending products/content based on His/Her choice/preferences

- Amazon
- Netflix’s
- Youtube
- Google
- Facebook
- Alibaba

Types Of Recommended Systems



Spark for Recommended systems

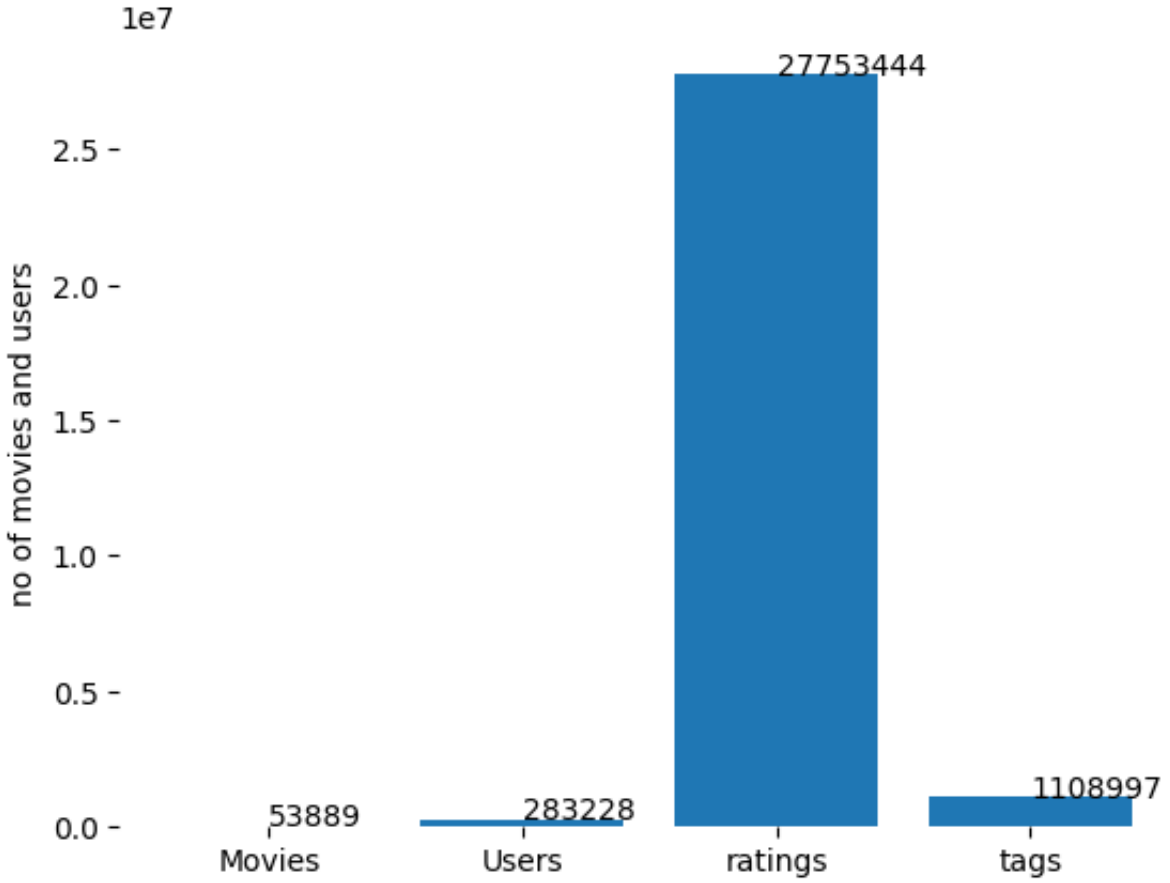
- Computing Similarity for Each new user and new preferences arrives from the existing user's
- Training of the model for each new user and for every new preferences of existing user (CBF)
- Its very expensive and it's a scalability problem
- distributed computation engine such as Spark to perform model computation is a must in any real-world recommendation engine

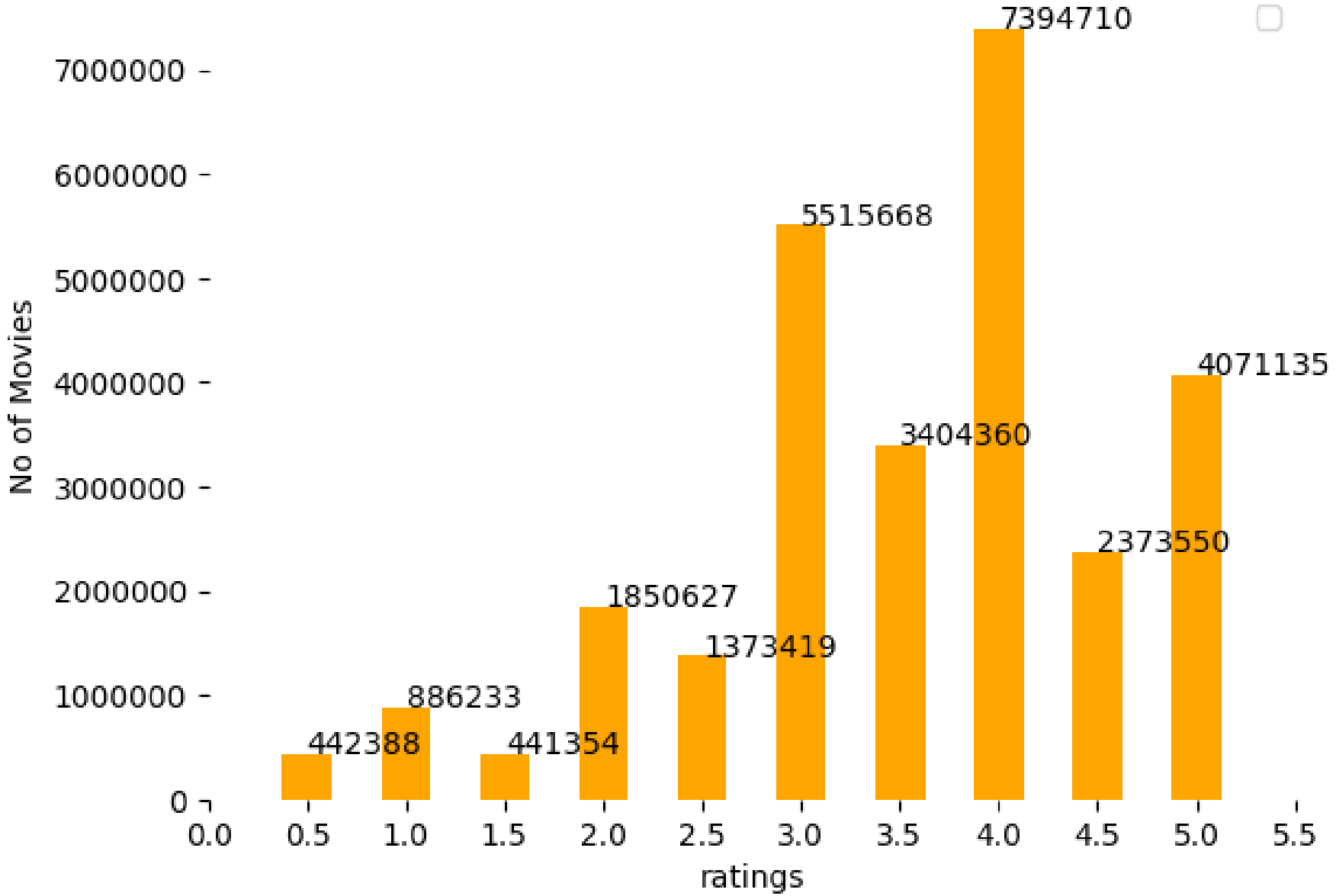
Dataset for movie recommendation system

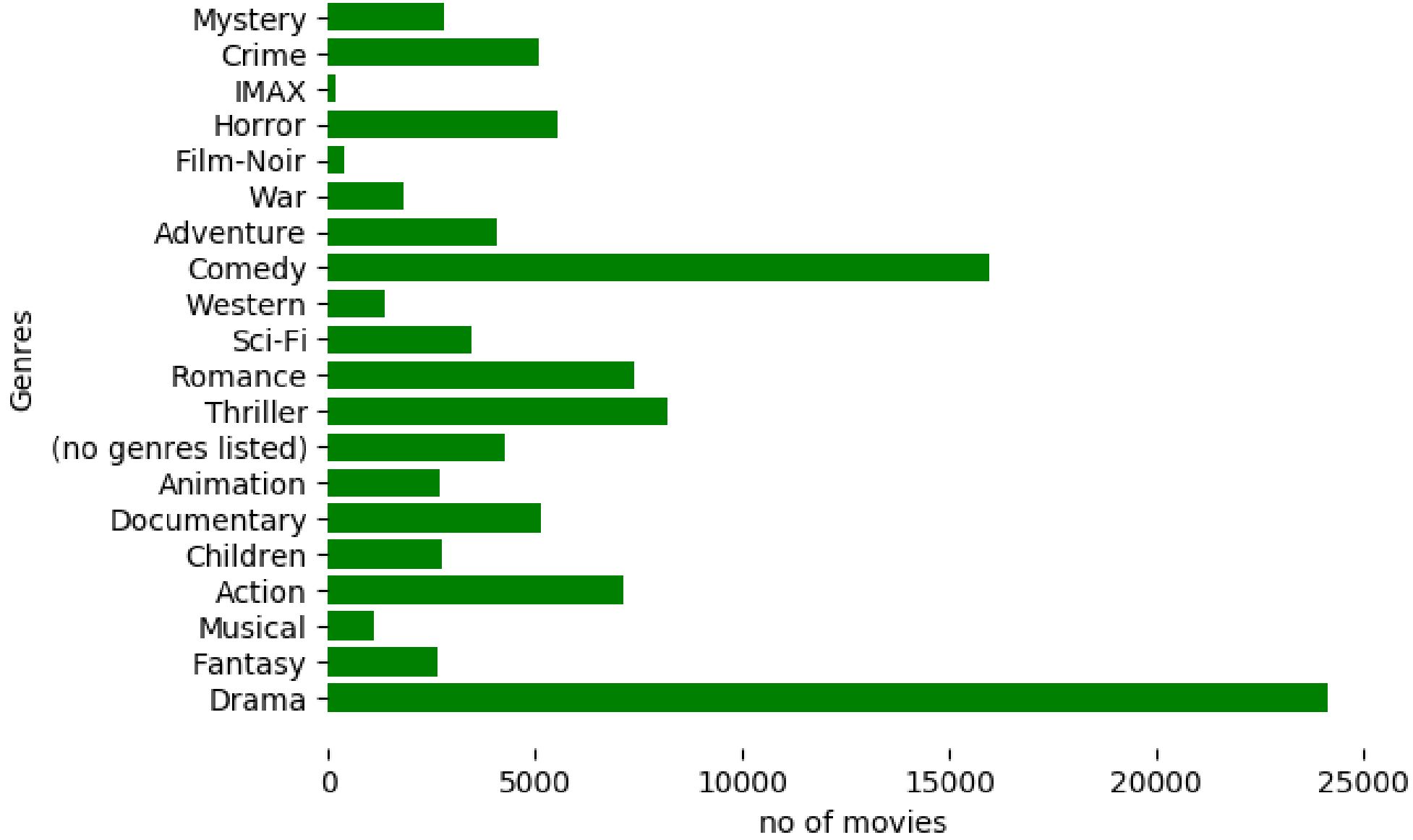
- Open Source data by Group lens
- Size: 1.65GB
- Consists of 4 Main files in CSV format
    - Movies
    - Ratings
    - Tags
    - Link for movies
- In this project Movies, Ratings, Tags are used

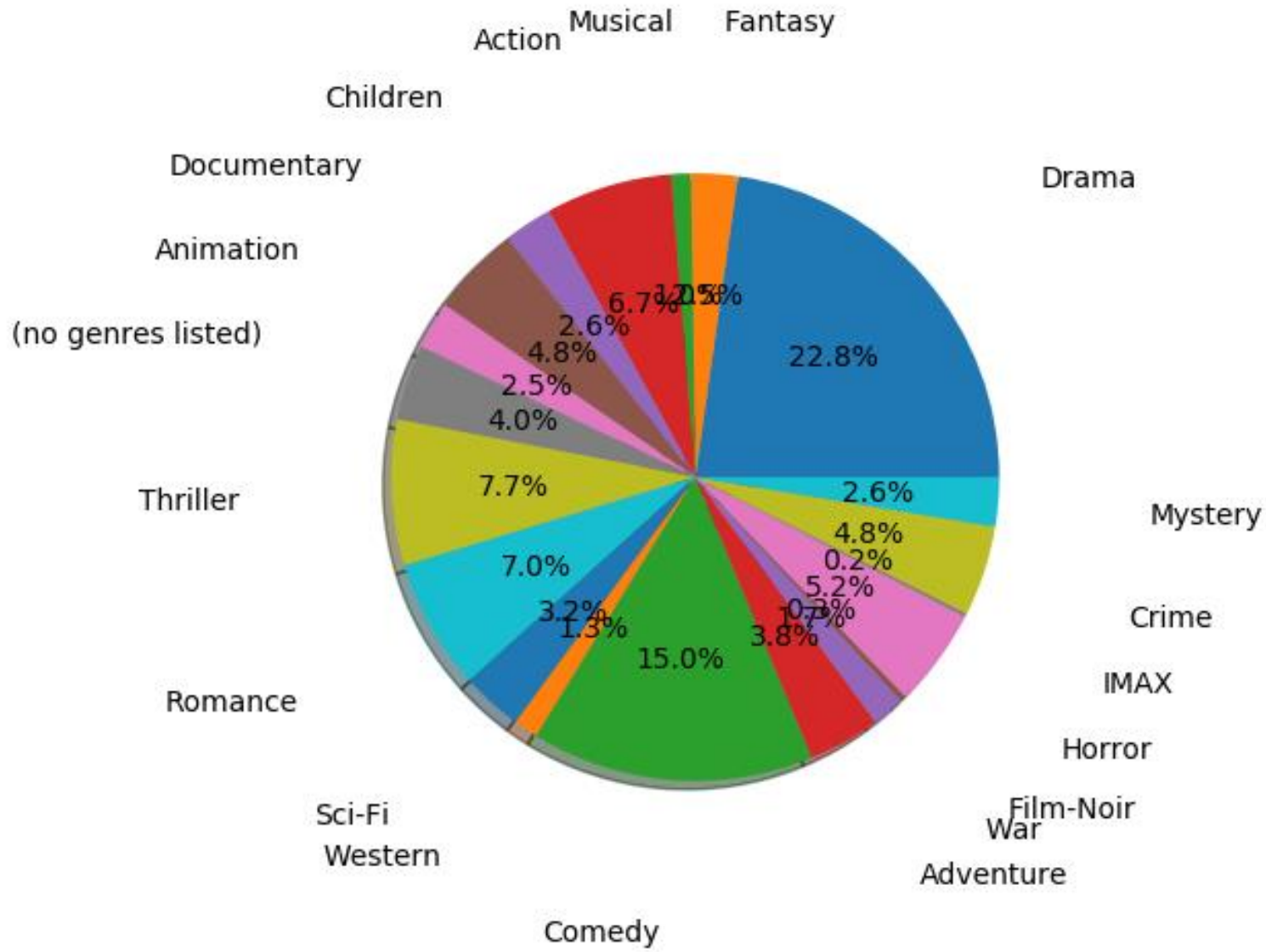
EDA on DATA

- Total No of Users = 283228
- Total No of Movies = 58098
- Total No of Ratings = 27753444
- Total No of Tags = 1108997

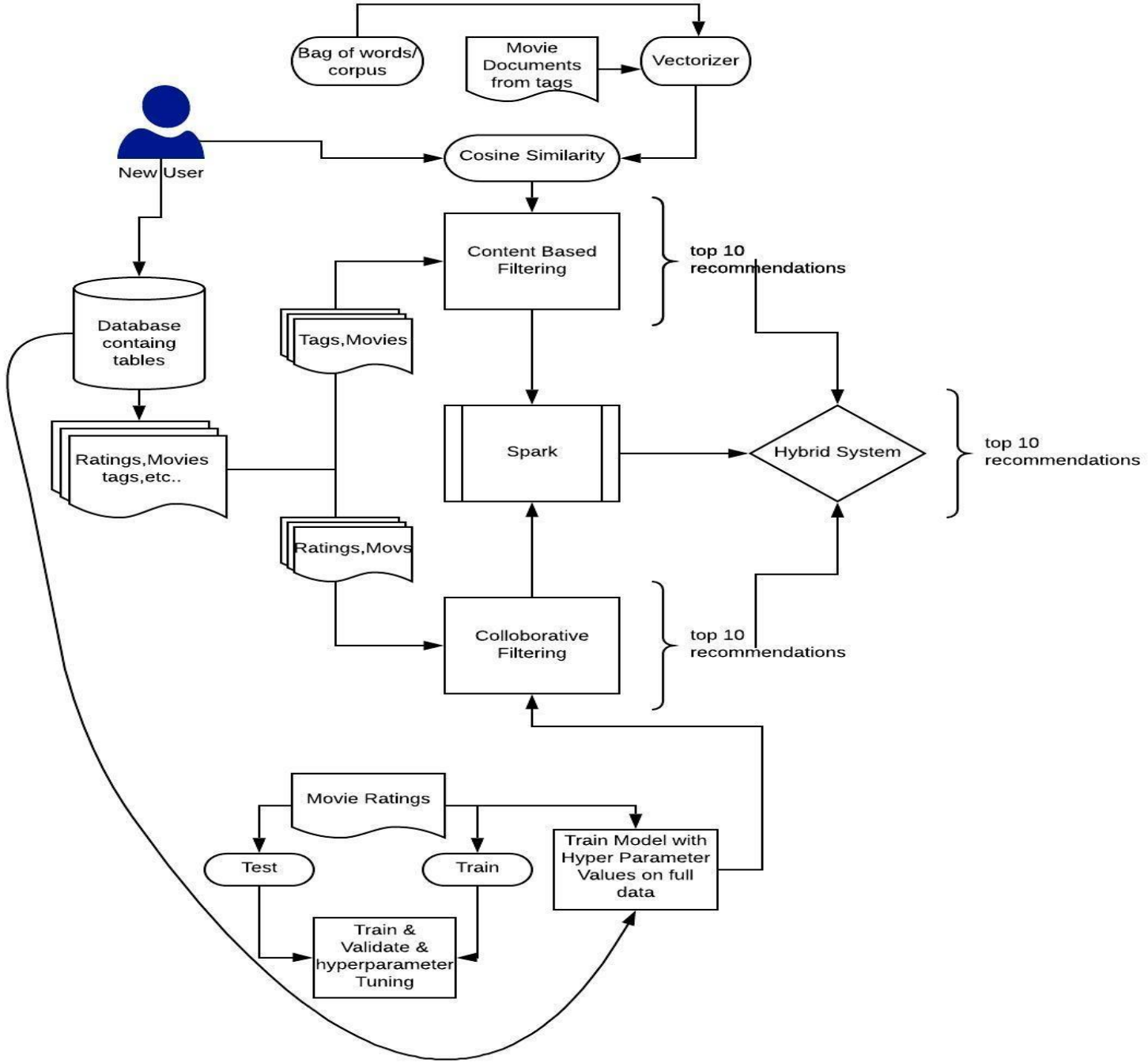



7394710
5515668
4071135
3404360
2373550
1850627
1373419
886233
442388
441354
7000000
6000000
5000000
4000000
3000000
2000000
1000000
0
No of Movies
0.0
0.5
1.0
1.5
2.0
2.5
3.0
3.5
4.0
4.5
5.0
5.5
ratings


Mystery
Crime
IMAX
Horror
Film-Noir
War
Adventure
Comedy
Western
Sci-Fi
Romance
Thriller
(no genres listed)
Animation
Documentary
Children
Action
Musical
Fantasy
Drama
Genres
0
5000
10000
15000
20000
25000
no of movies


Action
Musical
Fantasy
Children
Documentary
Drama
Animation
(no genres listed)
6.7%
2.6%
4.8%
2.5%
4.0%
22.8%
Thriller
7.7%
2.6%
Mystery
4.8%
0.2%
7.0%
5.2%
Crime
3.2%
1.3%
15.0%
3.8%
IMAX
Romance
Horror
Film-Noir
War
Sci-Fi
Western
Adventure
Comedy

Work-Flow


Bag of words/ corpus
Movie Documents from tags
Vectorizer
New User
Cosine Similarity
Content Based Filtering
top 10 recommendations
Database containg tables
Tags,Movies
Ratings,Movies tags,etc..
Spark
Hybrid System
top 10 recommendations
Ratings,Movs
Colloborative Filtering
top 10 recommendations
Movie Ratings
Test
Train
Train Model with Hyper Parameter Values on full data
Train & Validate & hyperparameter Tuning

Content Based (data processing)

movieId	title	genres
1	Toy Story (1995)	Adventure|Animation|Children|Comedy|Fantasy
2	Jumanji (1995)	Adventure|Children|Fantasy
3	Grumpier Old Men (1995)	Comedy|Romance
4	Waiting to Exhale (1995)	Comedy|Drama|Romance
5	Father of the Bride Part II (1995)	Comedy
6	Heat (1995)	Action|Crime|Thriller
7	Sabrina (1995)	Comedy|Romance
8	Tom and Huck (1995)	Adventure|Children
9	Sudden Death (1995)	Action

Showing the first 1000 rows.

userId	movieId	tag	timestamp
14	110	epic	1443148538
14	110	Medieval	1443148532
14	260	sci-fi	1442169410
14	260	space action	1442169421
14	318	imdb top 250	1442615195
14	318	justice	1442615192
14	480	Dinosaurs	1443148563
14	593	psychothriller	1444014286
14	1682	philosophy	1442615158

Showing the first 1000 rows.

movieid_num	document
2	{"data":["fantasy","adapted from:book","animals","bad cgi","based on a boo Actress)","scary","time","time travel","fantasy","Robin Williams","adapted fro travel","game","animals","comedy","fiction","thrill","Dynamic CGI Action","ba game","Children","Fantasy","Robin Williams","Kirsten Dunst","Robin William on a book","board game","Chris Van Allsburg","Baker viu","Robin Williams", game","fantasy","Robin Williams","time travel","Children","kid flick","bad cgi' Williams","time travel","kid flick","bad cgi","kid flick","scary","time travel","adv game","children","fantasy","Kirsten Dunst","Robin Williams","time travel","ar game","family","fantasy","fiction","Robin Williams","scary","time travel","boa
3	{"data":["moldy","old","Ann Margaret","Burgess Meredith","Daryl Hannah","' engraÃ fÂ§ada" "Funniest Movies" "sequel" "old people that is actually funn

Showing the first 622 rows.

words	frequency
epic	757
sci-fi	9400
imdb top 250	1795
philosophy	1430
Pixar	965
dinosaurs	329
classic sci-fi	375
must see	167
overrated	918

Showing the first 1000 rows.

▸ (1) Spark Jobs

```
[['animated', 'buddy movie', 'Cartoon', 'cgi', 'comedy', 'computer animation', 'family', 'friendship', 'kids', 'toy', 'toys', 'adventure', 'animated', 'animation',
'buddy movie', 'children', 'classic', 'clever', 'comedy', 'computer animation', 'Disney', 'family', 'fantasy', 'funny', 'humorous', 'imdb top 250', 'Pixar', 'time tr
avel', 'Tom Hanks', 'toys', 'witty', 'animation', 'cartoon', 'friendship', 'pixar', 'unny', 'adventure', 'animated', 'animation', 'computer animation', 'Disney', 'fu
nny', 'pixar', 'Tom Hanks', 'toys', 'pixar', 'animation', 'children', 'Pixar', 'witty', 'animated', 'animation', 'children', 'comedy', 'fantasy', 'funny', 'humorou
s', 'Pixar', 'time travel', 'animation', 'Pixar', 'animation', 'fun', 'animation', 'Pixar', 'computer animation', 'funny', 'Pixar', 'animation', 'Pixar', 'toys', 'ad
venture', 'children', 'classic', 'computer animation', 'Disney', 'funny', 'Pixar', 'Tim Allen', 'Tom Hanks', 'animation', 'Disney', 'animation', 'children', 'comed
y', 'Pixar', 'Disney', 'Pixar', 'Pixar', 'family', 'friendship', 'pixar', 'Watched', 'Pixar', 'witty', 'animation', 'humorous', 'Pixar', 'time travel', 'Disney', 'fu
nny', 'Pixar', 'time travel', 'children cartoon', 'feel-good', 'funny', 'animation', 'clever', 'friendship', 'funny', 'humorous', 'pixar', 'witty', 'Disney', 'advent
ure', 'animation', 'children', 'comedy', 'animation', 'comedy', 'funny', 'imdb top 250', 'Pixar', 'itaege', 'Tim Allen', 'Tom Hanks', 'fun', 'action figure', 'action
figures', 'Buzz Lightyear', 'CG animation', 'toy', 'toys', 'Woody', 'animation', 'Disney', 'friendship', 'pixar', 'joss whedon', 'animation', 'children', 'Disney',
'unlikely friendships', 'animation', 'comedy', 'kids', 'pixar', 'Tom Hanks', 'Pixar', 'comedy', 'imdb top 250', 'pixar', 'buddy movie', 'computer animation', 'friend
ship', 'Tom Hanks', 'animation', 'Pixar', 'pixar', 'tim allen', 'tom hanks', 'pixar', 'animation', 'cgi', 'Disney', 'family', 'toys', 'adventure', 'animation', 'Disn
ey', 'pixar', 'toys', 'computer animation', 'good cartoon chindren', 'pixar', 'animated', 'cgi', 'comedy', 'children', 'family', 'Pixar', 'Tom Hanks', 'toys', 'witt
y', 'dolls', 'National Film Registry', 'fantasy', 'children', 'computer animation', 'Disney', 'family', 'Pixar', 'very good', 'computer animation', 'Disney', 'fantas
y', 'Pixar', 'toys', 'witty', 'animation', 'cute', 'funny', 'story', 'voice acting', 'witty', 'animation', 'Disney', 'pixar', 'animation', 'pixar', 'Tim Allen', 'Tom
Hanks', 'animation', 'clever', 'Disney', 'pixar', 'pixar', 'childish', 'Pixar', 'computer animation', 'funny', 'humorous', 'Pixar', 'Tom Hanks', 'witty', 'toy', 'toy
s', 'animation', 'children', 'comedy', 'Disney', 'friendship', 'funny', 'pixar', 'Tom Hanks', "Tumey's To See Again", "Tumey's VHS", 'Animation', 'Cartoon', 'Pixar',
'humorous', 'pixar', 'boy', 'boy next door', 'bullying', 'friends', 'friendship', 'jealousy', 'martial arts', 'mission', 'neighborhood', 'new toy', 'pixar', 'rescu
e', 'resourcefulness', 'rivalry', 'toy', 'toy comes to life', 'walkie talkie', 'Disney', 'Pixar', 'Tim Allen', 'Tom Hanks', 'Pixar', 'animation', 'Disney', 'Pixar',
'exciting plot', 'funny lines', 'touching story', 'classic', 'comedy', 'fun', 'funny', 'humorous', 'Pixar', 'rated-G', 'want to see again', 'animation', 'classic',
'comedy', 'computer animation', 'Disney', 'funny', 'humorous', 'Pixar', 'time travel', 'Tom Hanks', 'witty', 'first cgi film', 'Pixar', 'animation', 'friendship', 't
oys', 'children', 'kids and family', 'friendship', 'pixar', 'toys', 'pixar', 'adventure', 'animated', 'animation', 'Cartoon', 'Disney', 'family', 'friendship', 'imdb
top 250', 'pixar', 'toy', 'toys', 'Disney', 'animated', 'animation', 'buddy movie', 'children', 'clever', 'time travel', 'witty', 'animation', 'family', 'Tom Hanks',
'clever', 'clever', 'witty', 'funny', 'Pixar', 'pixar', 'Pixar', 'animation', 'clever', 'friendship', 'funny', 'Tom Hanks', 'witty', 'adventure', 'clever', 'pixar',
'animated', 'animation', 'buddy movie', 'computer animation', 'funny', 'Pixar', 'Tom Hanks', 'adventure', 'animation', 'Disney', 'funny', 'pixar', 'Os dois viram',
'animation', 'comedy', 'Disney', 'Pixar', 'animation', 'computer animation', 'pixar', 'toys', 'children', 'computer animation', 'family', 'humorous', 'time travel',
'Tom Hanks', 'witty', 'adventure', 'animation', 'Disney', 'funny', 'pixar', 'computer animation', 'Disney', 'humorous', 'Pixar', 'animated', 'fun family movie', 'pix
```

Command took 0.33 seconds -- by 9440110838bharath@gmail.com at 4/20/2020, 5:02:02 PM on My Cluster

Content Based (Vectorizing Doc)

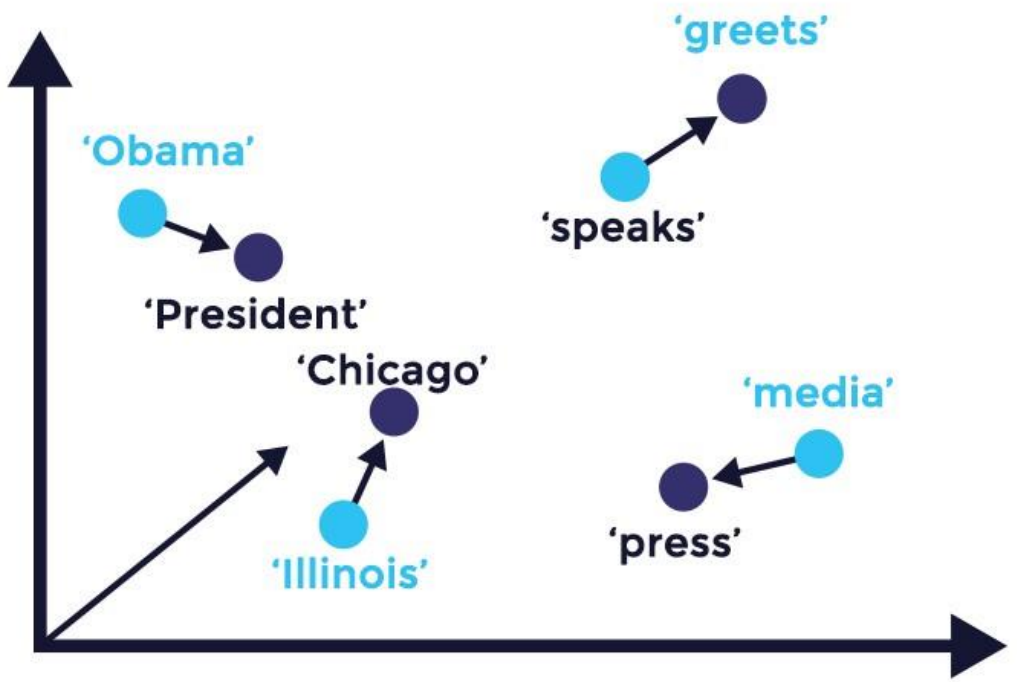
- Bag of words (Corpus)
- TFIDF based on the Corpus

document	_1
{"data":["woman director","road trip","Goth","road trip","teen angst"],"index":0,"maxindex":5}	[0,74672,[5608,42908,46607,54961],[10.33532158156557,2.541775989724065,7.098419134817496,8.3381100217455 1]]
{"data":["woman director","might like","documentary","woman director","female director"],"index":0,"maxindex":5}	[0,74672,[3097,31253,32024,42908],[4.160929453944261,4.316010366396739,5.692580177624635,5.08355197944813]]
{"data":["woman director","marriage","prostitution","unconventional romance"],"index":0,"maxindex":4}	[0,74672,[14670,39515,39993,42908],[4.702919072745081,5.137583335545507,8.171055937082345,2.541775989724065]]
{"data":["woman director","female director"],"index":0,"maxindex":2}	[0,74672,[32024,42908],[5.692580177624635,2.541775989724065]]
{"data":["woman director","cats","declawing","animal documentary","animals","cats","declawing","documentary"],"index":0,"maxindex":8}	[0,74672,[3097,22594,42908,45588,46221,68733],[4.160929453944261,10.042858113983936,2.541775989724065,13.21774181
{"data":["woman director","Rom-Com","Zosia Mamet"],"index":0,"maxindex":3}	[0,74672,[11378,42908,55150],[8.790095145488568,2.541775989724065,9.637393005875772]]
{"data":["woman director","Liv Ullmann"],"index":0,"maxindex":2}	[0,74672,[42908,61661],[2.541775989724065,6.625131430370571]]
{"data":["woman director","Isabel Coixet"],"index":0,"maxindex":2}	[0,74672,[19306,42908],[8.538780717207663,2.541775989724065]]
{"data":["woman director","Allison Berg","Frank Keraudren"],"index":0,"maxindex":3}	[0,74672,[24138,32545,42908],[9.637393005875772,9.637393005875772,2.541775989724065]]
{"data":["woman director","Added","based on a book","Nora Roberts","based on a book","made for TV"],"index":0,"maxindex":6}	[0,74672,[25628,31803,42908,54593,56861],[5.230673758611519,6.92999351252578,2.541775989724065,8.433420201549836,6
{"data":["woman director","19th century","biography","feminism","prison","prison drama","Spain","women"],"index":0,"maxindex":8}	[0,74672,[9339,13636,14498,23977,42908,49032,61769,65045],[3.9083760600410296,5.537508263278056,5.28926792287757 1,
{"data":["woman director"],"index":0,"maxindex":1}	[0,74672,[42908],[2.541775989724065]]
{"data":["woman director"],"index":0,"maxindex":1}	[0,74672,[42908],[2.541775989724065]]
{"data":["woman director"],"index":0,"maxindex":1}	[0,74672,[42908],[2.541775989724065]]
{"data":["woman director"],"index":0,"maxindex":1}	[0,74672,[42908],[2.541775989724065]]
{"data":["woman director"],"index":0,"maxindex":1}	[0,74672,[42908],[2.541775989724065]]
{"data":["woman director"],"index":0,"maxindex":1}	[0,74672,[42908],[2.541775989724065]]
{"data":["woman director"],"index":0,"maxindex":1}	[0,74672,[42908],[2.541775989724065]]

Content Based (cosine-similarity)

$$\text{similarity} = \cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}}$$

Document 1	Document 2
Obama speaks to the media in Illinois	The Presidents greets the press in Chicago



Content Based (Result)

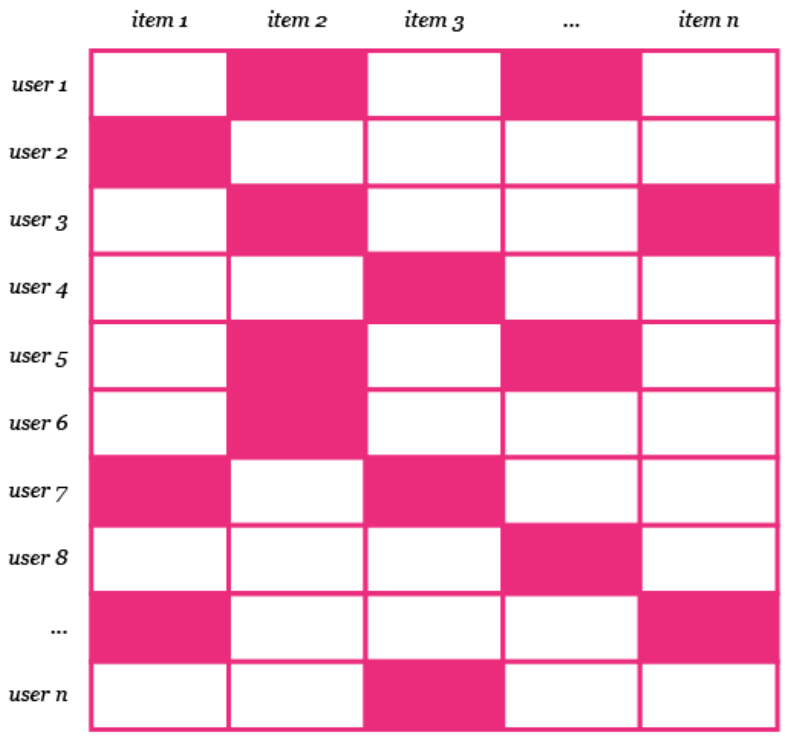
```
new_u_ratings = [
     (0,260,4), # Star Wars (1977)
     (0,1,3), # Toy Story (1995)
     (0,16,3), # Casino (1995)
     (0,25,4), # Leaving Las Vegas (1995)
     (0,32,4), # Twelve Monkeys (a.k.a. 12 Monkeys) (1995)
     (0,335,1), # Flintstones, The (1994)
     (0,379,1), # Timecop (1994)
     (0,296,3), # Pulp Fiction (1994)
     (0,858,5) , # Godfather, The (1972)
     (0,50,4) # Usual Suspects, The (1995)
    ]
```

```
260
[33493, 5378, 98695, 1210, 1196]
1
[4886, 6377, 78499, 2355, 3114]
16
[8042, 89069, 85190, 120795, 1213]
25
[2505, 4144, 7759, 31785, 2933]
32
[4878, 135859, 157563, 96610, 114935]
296
[18, 6874, 7438, 1729, 1089]
858
[2023, 1466, 151591, 4262, 1221]
50
[1625, 47, 1617, 2762, 97973]
```

movieid_num	document	title	genres	movienum	similarity_score
1089	[great dialogue,...	Reservoir Dogs (1...	Crime|Mystery|Thr...	1089	0.7559349517270009
1213	[crime, dark com...	Goodfellas (1990)	Crime|Drama	1213	0.7567710590313329
1221	[boring, Mafia, ...	Godfather: Part I...	Crime|Drama	1221	0.7208180882462243
1729	[dark comedy, Qu...	Jackie Brown (1997)	Crime|Drama|Thriller	1729	0.6631345862127239
2355	[animation, Disn...	Bug's Life, A (1998)	Adventure|Animati...	2355	0.6634401708197005
3114	[Pixar, sequel b...	Toy Story 2 (1999)	Adventure|Animati...	3114	0.7522343935343732
6377	[Pixar, funny, P...	Finding Nemo (2003)	Adventure|Animati...	6377	0.6474311477254641
6874	[Kick-Butt Women...	Kill Bill: Vol. 1...	Action|Crime|Thri...	6874	0.6208954589449202
7438	[martial arts, m...	Kill Bill: Vol. 2...	Action|Drama|Thri...	7438	0.6292780262334008
78499	[boring, overrat...	Toy Story 3 (2010)	Adventure|Animati...	78499	0.6574993605321294

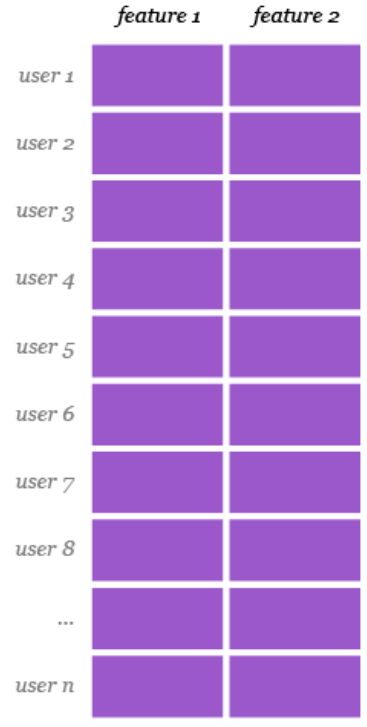
Collaborative Filtering (ALS)

User Movie Rating Matrix



R

$\approx$



U

Users Latent Factors

x

Movies Latent Factors

V

Collaborative Filtering (Data and Training the Model)

$$loss = L = \sum_{u=user} \sum_{i=item} (R_{ui} - X_u Y_i) + \lambda(||X||^2 + ||Y||^2)$$

ALS Loss Function

$$\frac{\partial L}{\partial \mathbf{y}_i} = -2\sum_i (r_{iu} - \mathbf{y}_i^\intercal \cdot \mathbf{x}_u)\mathbf{x}_u^\intercal + 2\lambda_y \mathbf{y}_i^\intercal$$

$$\frac{\partial L}{\partial \mathbf{x}_u} = -2\sum_i (r_{ui} - \mathbf{x}_u^\intercal \cdot \mathbf{y}_i)\mathbf{y}_i^\intercal + 2\lambda_x \mathbf{x}_u^\intercal$$

```
the sample data look like in RDD
Out[5]: [(1, 1, 4.0), (1, 3, 4.0), (1, 6, 4.0)]

Command took 0.30 seconds -- by 9440110838bharath@gmail.com
```

▸ (1) Spark Jobs

```
movies RDD Look like Below
Out[7]: [(1, 'Toy Story (1995)'),
 (2, 'Jumanji (1995)'),
 (3, 'Grumpier Old Men (1995)')]

Command took 0.57 seconds -- by 9440110838bharath@gmail.com
```

Collaborative Filtering (Data and Training the Model)

Train Test Split

```
#train_test_split the data of ratings
train, test = read_ratings_1m_data.randomSplit([7, 3], seed=0)
#train.take(3)
test.take(3)
```

Hyper Parameters

```
seed = 5
iterations = 10
reg_parameter = 0.1
rank = 8
```

Mean Square Error of ALS Model

```
▸ (6) Spark Jobs
Root Mean Square Error
0.8162060462702274

Command took 14.48 minutes -- by 9440110838bharath@gmail.com
```

Collaborative Filtering (Results)

New User with Movies watched and rated

```
new_u_ratings = [
     (0,260,4), # Star Wars (1977)
     (0,1,3), # Toy Story (1995)
     (0,16,3), # Casino (1995)
     (0,25,4), # Leaving Las Vegas (1995)
     (0,32,4), # Twelve Monkeys (a.k.a. 12 Monkeys) (1995)
     (0,335,1), # Flintstones, The (1994)
     (0,379,1), # Timecop (1994)
     (0,296,3), # Pulp Fiction (1994)
     (0,858,5) , # Godfather, The (1972)
     (0,50,4) # Usual Suspects, The (1995)
    ]
```

Recommended movies

```
Out[26]: [(109370, 'Magic & Bird: A Courtship of Rivals (2010)', 4.205458047054151, 27),
 (4171, "Long Night's Journey Into Day (2000)", 4.034671592120159, 35),
 (858, '"Godfather', 3.9401285962269315, 60904),
 (139086,
  'Small Potatoes - Who Killed the USFL? (2009)',
  3.9250156017837905,
  26),
 (171495, 'Cosmos', 3.9183881373494582, 157),
 (79677, '"Two Escobars', 3.9133231390548304, 78),
 (1221, '"Godfather: Part II', 3.8746254985856208, 38875),
 (173351, 'Wow! A Talking Fish! (1983)', 3.8728742513819707, 47),
 (108043, 'Milius (2013)', 3.8611850190449717, 44),
 (160317, 'Roots (2016)', 3.860586669993372, 28),
 (116002, 'History of the Eagles (2013)', 3.8520613377611097, 32),
 (91762, '"Last Lions', 3.8439770709231045, 38),
 (70186,
  'Heimat - A Chronicle of Germany (Heimat - Eine deutsche Chronik) (1984)',
  3.8351099378886944,
  35),
 (296, 'Pulp Fiction (1994)', 3.801915117453738, 92406),
 (100553, 'Frozen Planet (2011)', 3.793612933978338, 402),
 (182615, 'The Landlord (2007)', 3.7694492637625405, 26),
 (165239, 'Supersonic (2016)', 3.7613698898885604, 51),
 (128981,
  'Music for One Apartment and Six Drummers (2001)',
  3.758114602698226,
  31),
```

Failure of Content and Collaborative Filtering

Collaborative Filtering

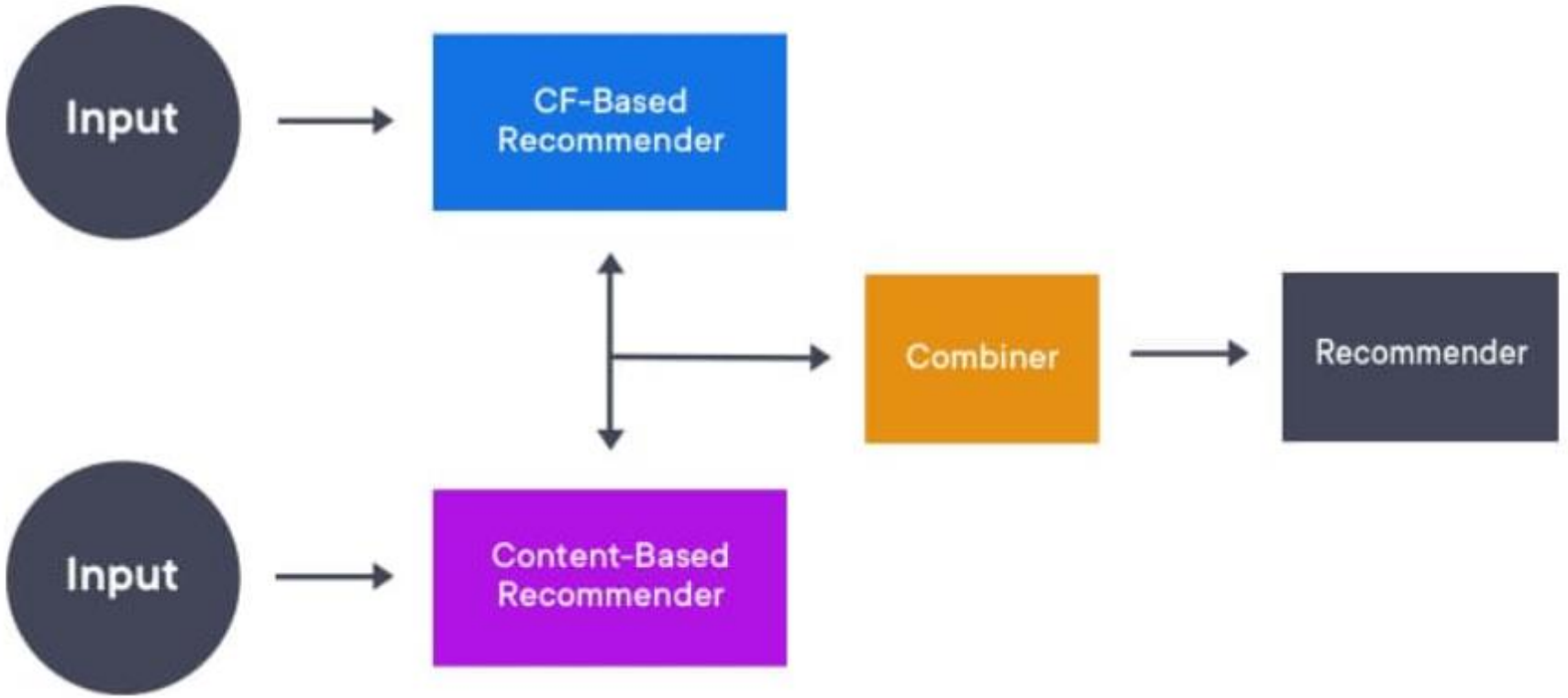
movieId	title	genres	movie_num
109370	Magic & Bird: A C...	Documentary	109370
4171	Long Night's Jour...	Documentary	4171
858	Godfather, The (1...	Crime|Drama	858
139086	Small Potatoes - ...	Documentary	139086
171495	Cosmos	(no genres listed)	171495

Content Based

movieId	title	genres	movienum
1089	Reservoir Dogs (1...	Crime|Mystery|Thr...	1089
1213	Goodfellas (1990)	Crime|Drama	1213
1221	Godfather: Part I...	Crime|Drama	1221
2355	Bug's Life, A (1998)	Adventure|Animati...	2355
3114	Toy Story 2 (1999)	Adventure|Animati...	3114

Hybrid Recommended System (Weighted factor)

- Weighted Ratio for Content and Collaborative approaches
- Ensemble Learning



Hybrid Recommended System (Result)

New User Movies

```
new_u_ratings = [
     (0,260,4), # Star Wars (1977)
     (0,1,3), # Toy Story (1995)
     (0,16,3), # Casino (1995)
     (0,25,4), # Leaving Las Vegas (1995)
     (0,32,4), # Twelve Monkeys (a.k.a. 12 Monkeys) (1995)
     (0,335,1), # Flintstones, The (1994)
     (0,379,1), # Timecop (1994)
     (0,296,3), # Pulp Fiction (1994)
     (0,858,5) , # Godfather, The (1972)
     (0,50,4) # Usual Suspects, The (1995)
    ]
```

Content Based

movienum	similarity_score
1213	0.7567710590313329
1089	0.7559349517270009
3114	0.7522343935343732
1221	0.7208180882462243
2355	0.6634401708197005

Collaborative

movie_num	moviename	rat	count
109370	Magic & Bird: A C...	4.205458047054151	27
4171	Long Night's Jour...	4.034671592120159	35
858	"Godfather	3.9401285962269315	60904
139086	Small Potatoes - ...	3.9250156017837905	26
171495	Cosmos	3.9183881373494582	157

Final Hybrid Result of Content And Collaborative Filtering

movieId	title	genres	movienum
1089	Reservoir Dogs (1...	Crime|Mystery|Thr...	1089
1213	Goodfellas (1990)	Crime|Drama	1213
1221	Godfather: Part I...	Crime|Drama	1221
2355	Bug's Life, A (1998)	Adventure|Animati...	2355
3114	Toy Story 2 (1999)	Adventure|Animati...	3114
109370	Magic & Bird: A C...	Documentary	109370
4171	Long Night's Jour...	Documentary	4171
858	Godfather, The (1...	Crime|Drama	858
139086	Small Potatoes - ...	Documentary	139086
171495	Cosmos	(no genres listed)	171495

THANK YOU