Movie Recommendation

Qiao Qianqian
Deng Wenlong
Luo Yaxiong
Wang Pei

Content

Data Acquisition

Data Exploration

Recommendation Engine Based on Movie

Recommendation Engine Based on User

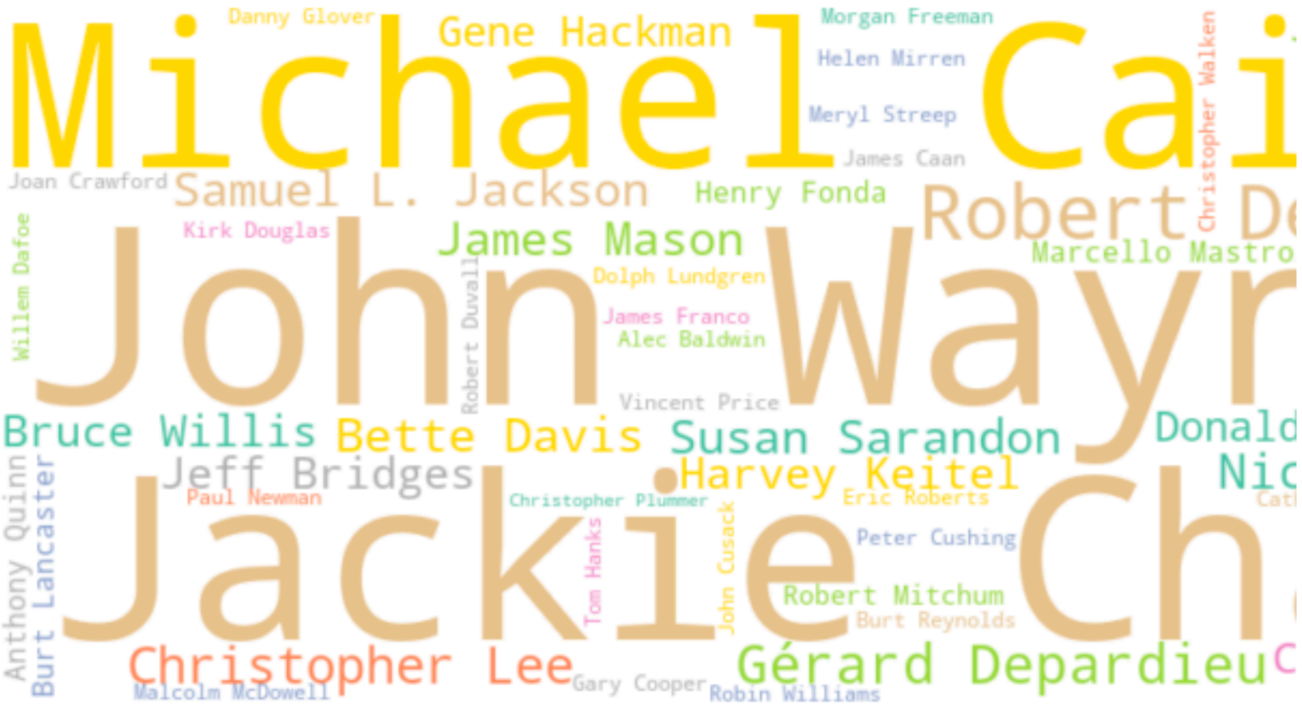
Data Acquisition

- Get useful information

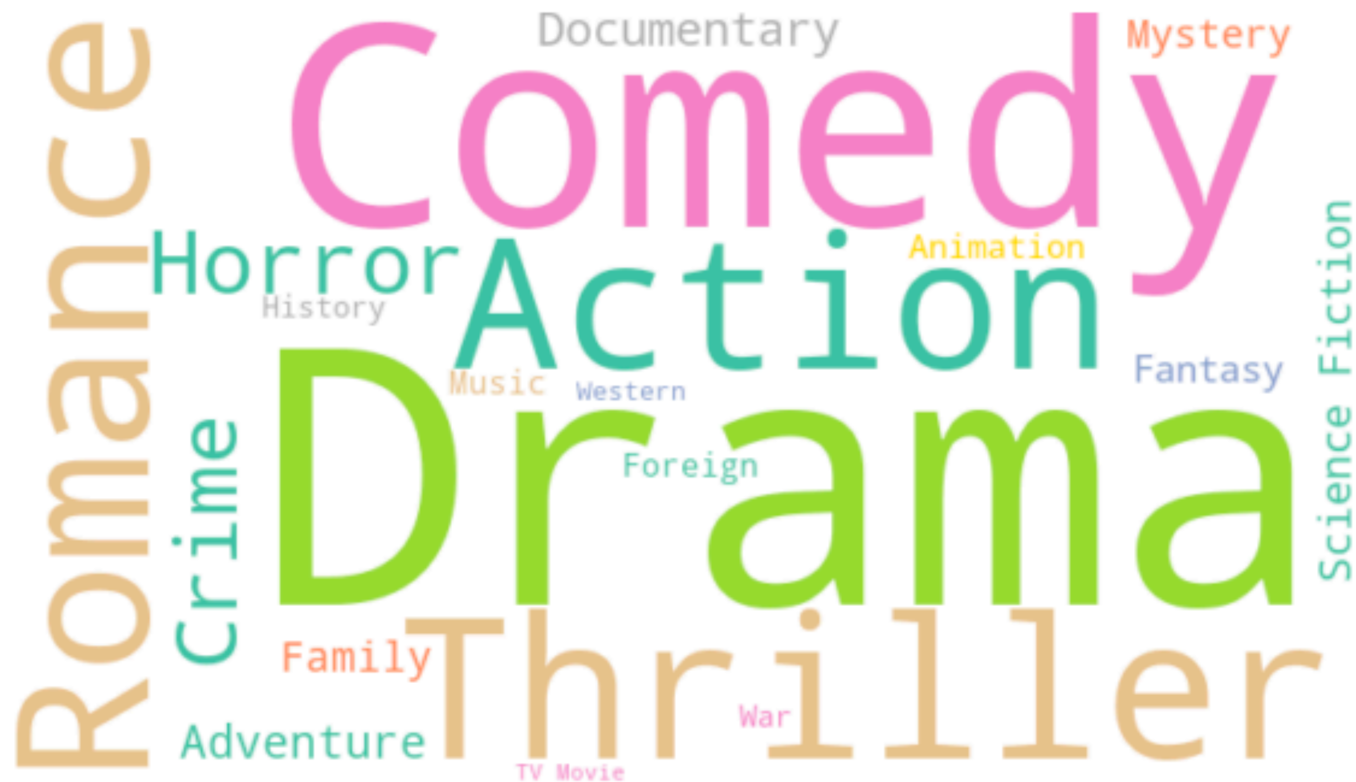
```
print(type(credits.iloc[0]['crew']))
crew_data = ast.literal_eval(credits.iloc[0]['crew'])
print(type(crew_data))
print(type(crew_data[0]))
for i in crew_data:
    if i['job']=='Director':
        print(i['name'])
```

```
<class 'str'>
<class 'list'>
<class 'dict'>
John Lasseter
```

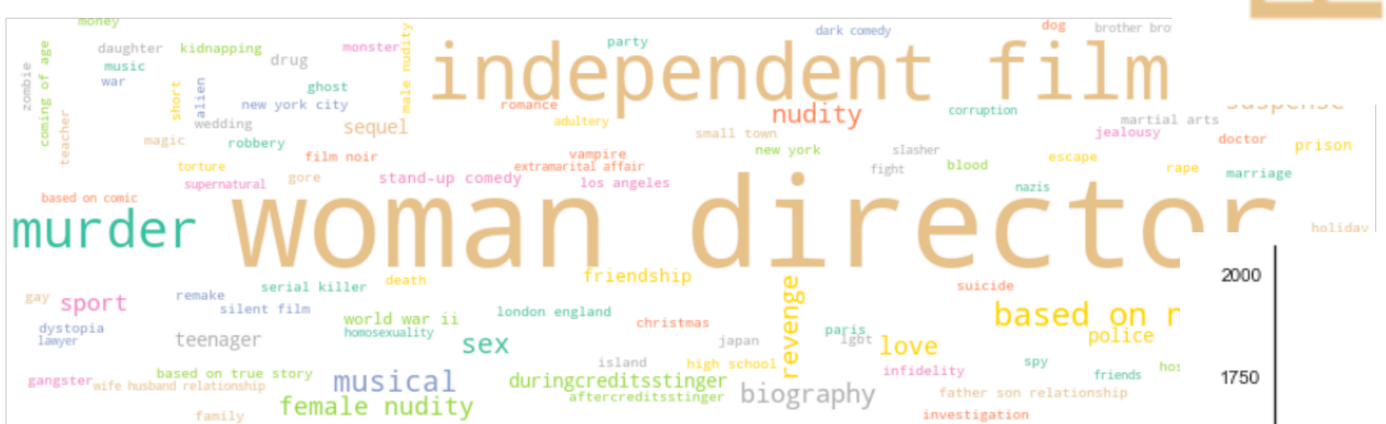
- Observation of data
  - Delete low-frequency words
  - Use word cloud to see frequency



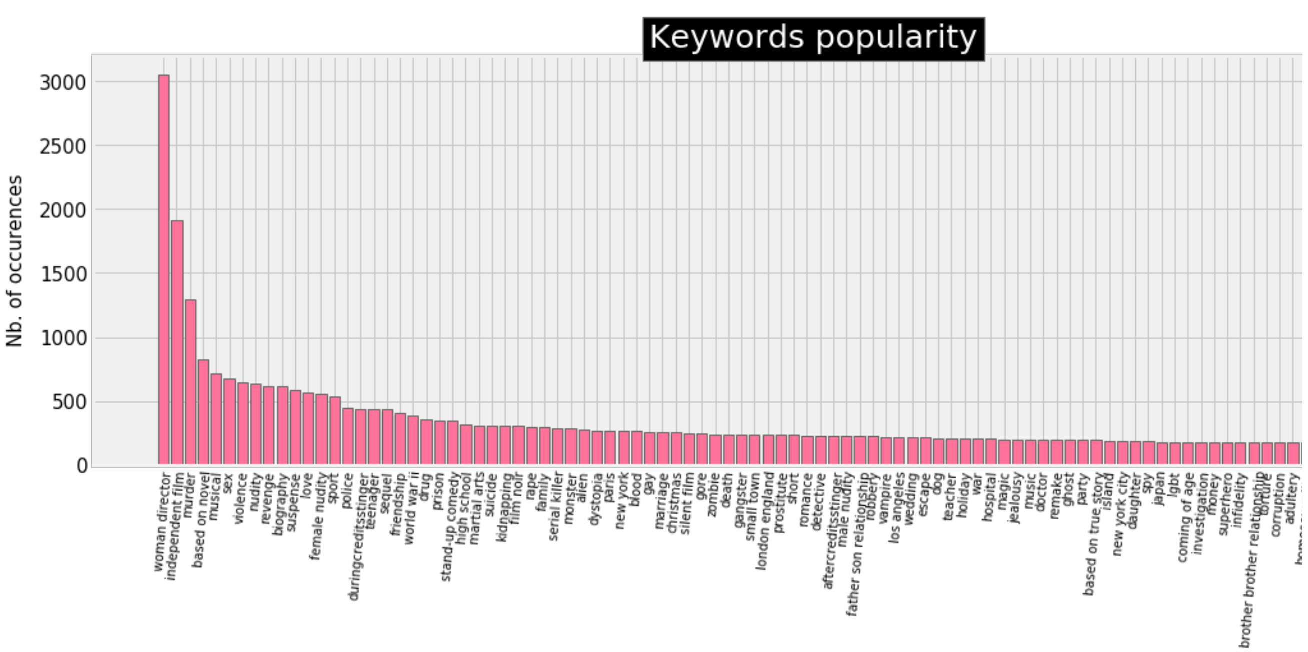
Michael Cai
Danny Glover
Gene Hackman
Morgan Freeman
Helen Mirren
Meryl Streep
James Caan
Christopher Walken
Joan Crawford
Samuel L. Jackson
Henry Fonda
Robert De
Kirk Douglas
James Mason
Marcello Mastro
Willem Dafoe
John Wayn
Dolph Lundgren
Robert Duvall
James Franco
Alec Baldwin
Vincent Price
Bruce Willis
Bette Davis
Susan Sarandon
Donald
Jeff Bridges
Harvey Keitel
Anthony Quinn
Burt Lancaster
Paul Newman
Christopher Plummer
Eric Roberts
Tom Hanks
John Cusack
Peter Cushing
Jackie Ch
Robert Mitchum
Burt Reynolds
Christopher Lee
Gary Cooper
Gérard Depardieu
Malcolm McDowell
Robin Williams



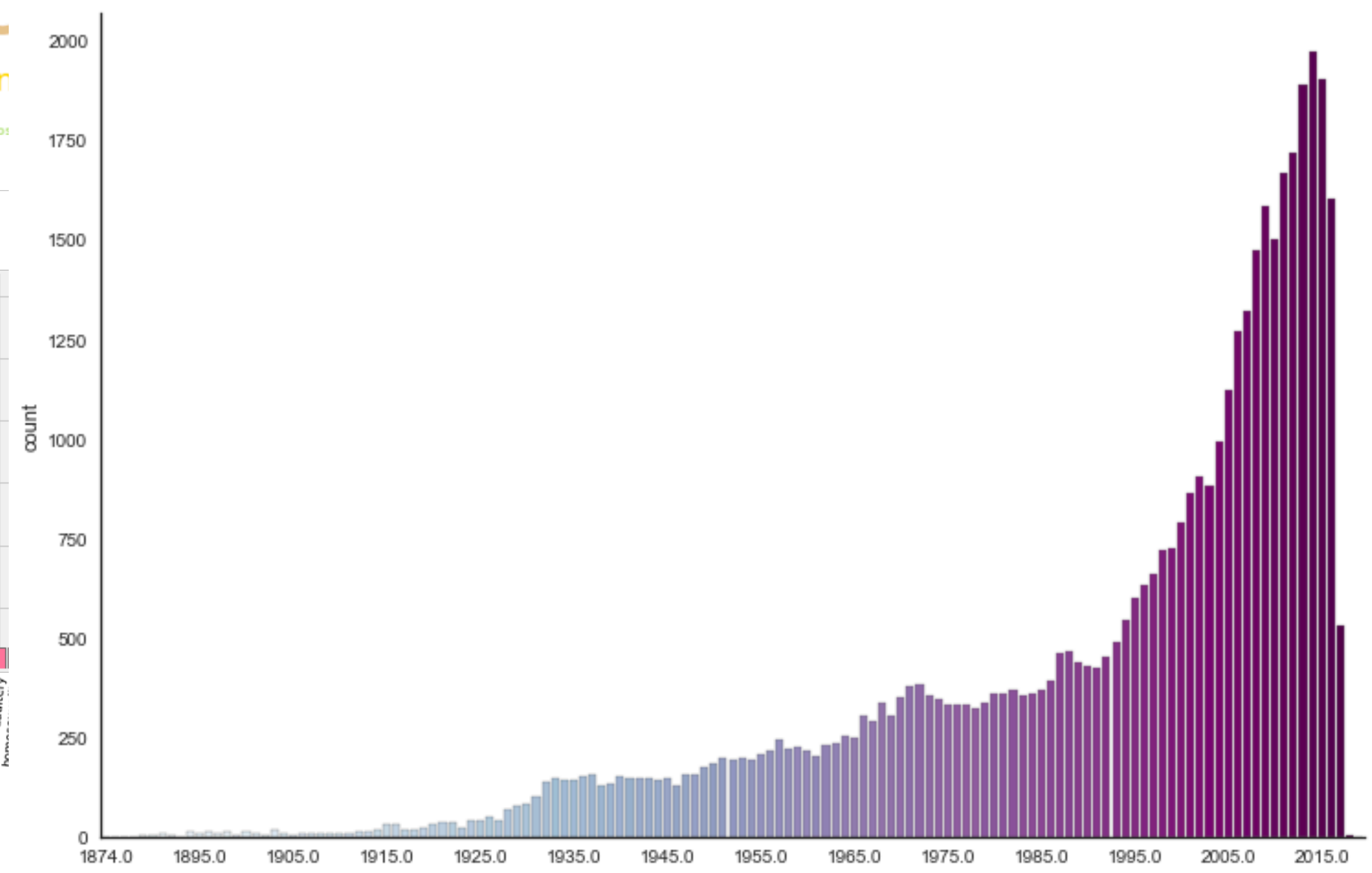
Documentary
Mystery
Comedy
Romance
Horror
Animation
Action
History
Science Fiction
Fantasy
Music
Western
Drama
Crime
Foreign
Family
Thriller
War
Adventure
TV Movie



independent film
woman director
murder
nudity
sequel
based on comic
stand-up comedy
film noir
vampire
extramarital affair
los angeles
supernatural
serial killer
friendship
revenge
based on n
police
sport
teenager
sex
love
musical
duringcreditsstinger
aftercreditsstinger
biography
female nudity
christmas
suicide
infidelity
father son relationship
investigation
prison



Keywords popularity
Nb. of occurences
0
500
1000
1500
2000
2500
3000
woman director
independent film
murder
based on novel
musical
sex
violence
nudity
revenge
biography
suspense
love
female nudity
sport
police
duringcreditsstinger
teenager
sequel
friendship
world war ii
drug
prison



count
0
250
500
750
1000
1250
1500
1750
2000
1874.0
1895.0
1905.0
1915.0
1925.0
1935.0
1945.0
1955.0
1965.0
1975.0
1985.0
1995.0
2005.0
2015.0
title_year

Data Acquisition

Keywords Cleaning

- -Word roots（Natural Language Toolkit package)

  root word holds the most basic meaning of any word

  replace sentence keywords with single word

```
hotel {'resort hotel', 'hotel guests', 'hotel', 'hotel manager', 'haunted hotel', 'luxury hotel', 'hotel suite'} 7
witch {'witch hunt', 'witch', 'witches', 'witch hunter'} 4
africa {'north africa', 'africa', 'south africa', 'cape town south africa', 'northern africa'} 5
subway {'subway train', 'subway station', 'subway', 'new york subway'} 4
```

- -Synonyms (wordnet.synsets module)

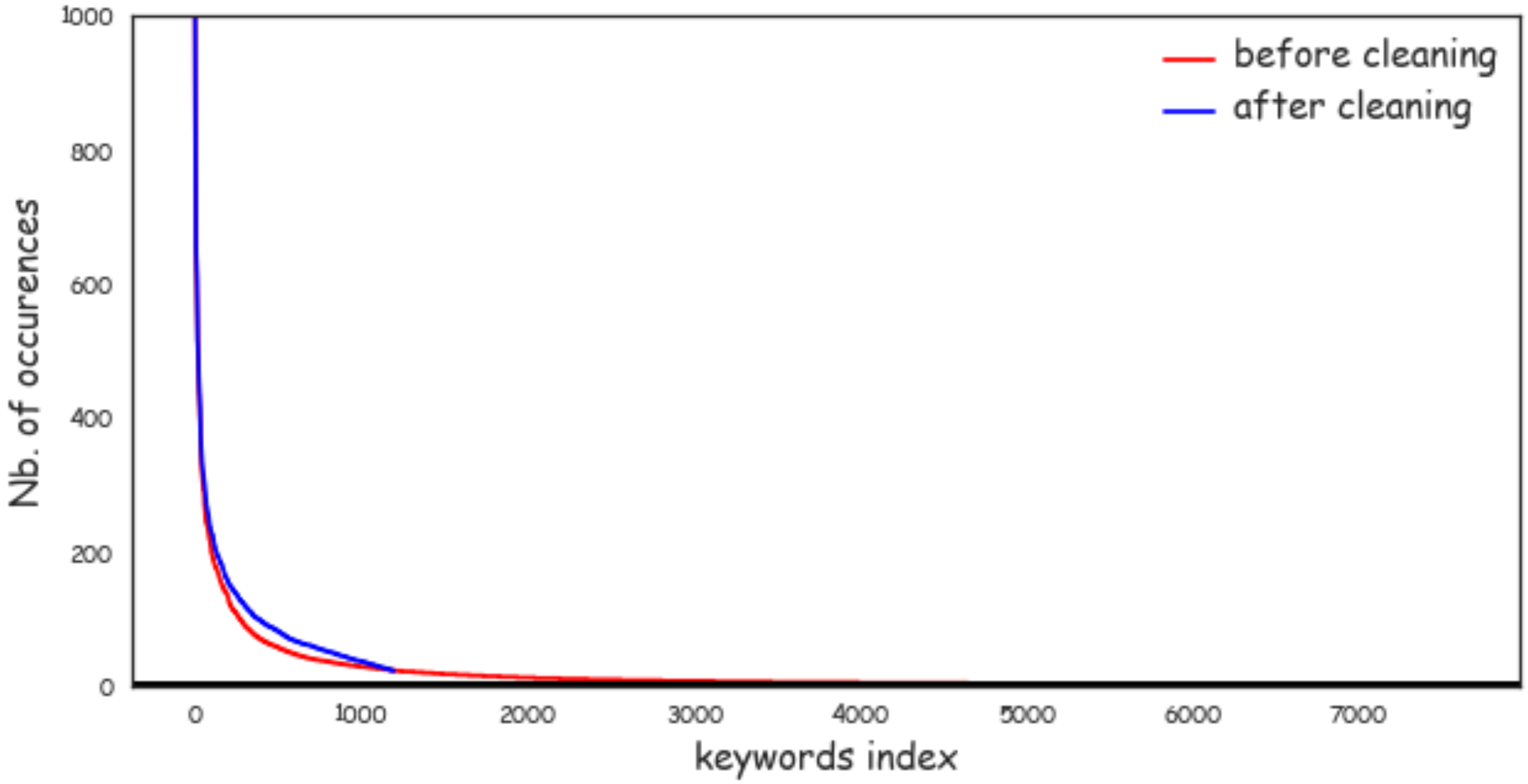
  find synonym

  Replace by high-frequency words

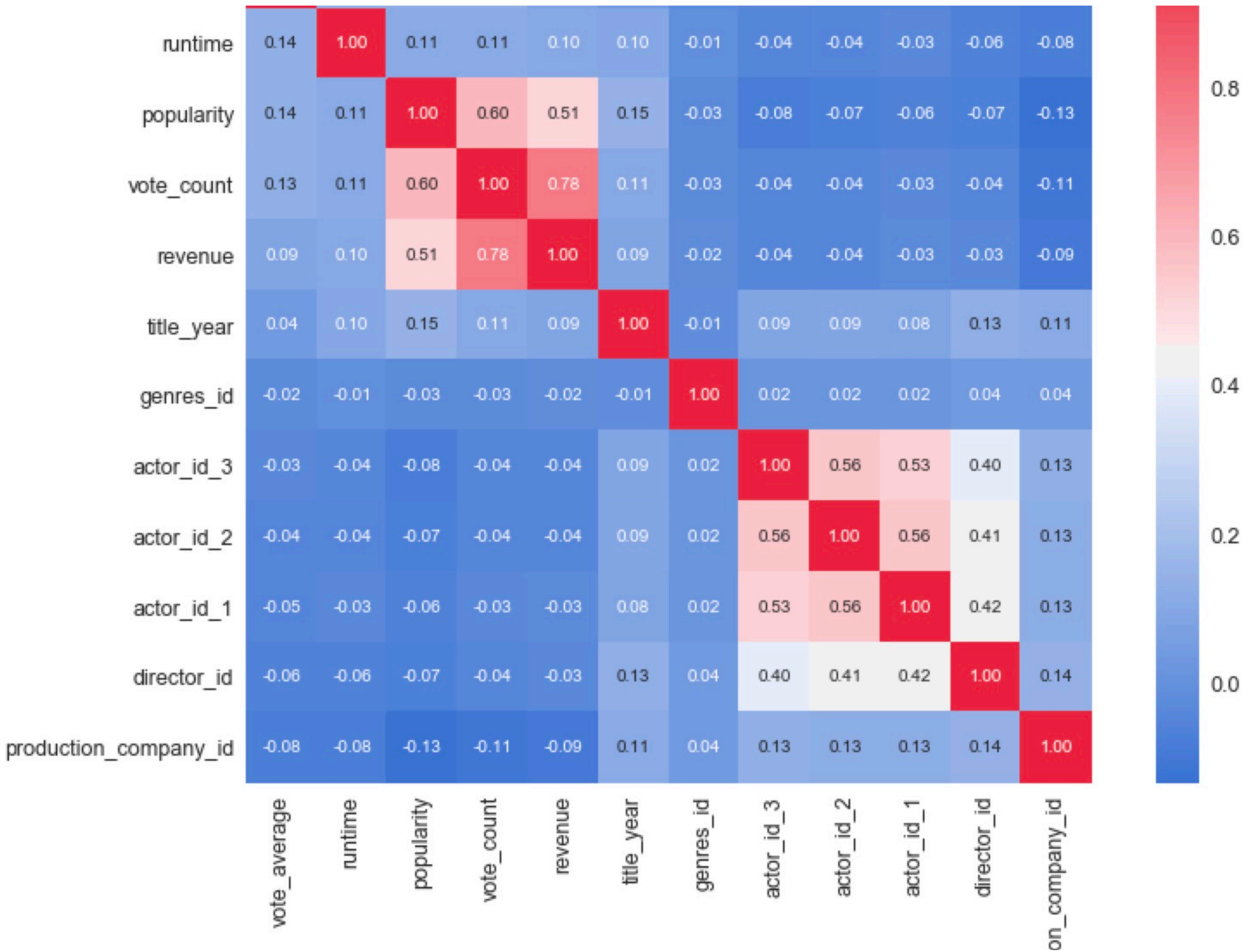
```
rebirth          -> reincarnation
nirvana          -> heaven
seal             -> navy seal
enchantment      -> spell
oldtimer         -> veteran
```

-Word vector

we use the spacy package to calculate the similarity of word vector within words.
(which based on word2vec algorithm)



Here we set nm_keep as 1200, From the picture we can see we have cut a lot keywords .



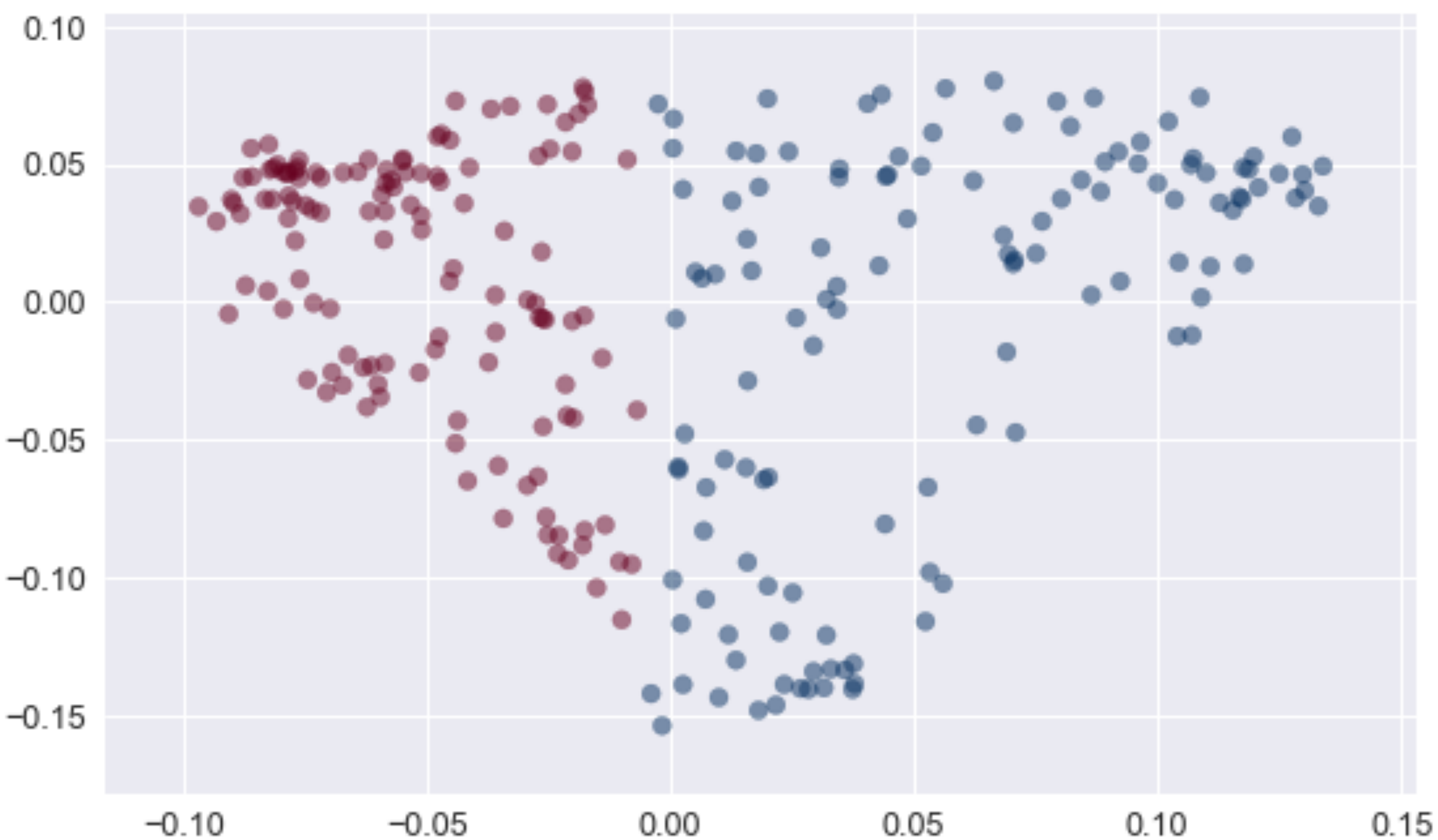
runtime
popularity
vote_count
revenue
title_year
genres_id
actor_id_3
actor_id_2
actor_id_1
director_id
production_company_id
vote_average
on_company_id
0.8
0.6
0.4
0.2
0.0

Data Exploration

Movies Clustering

Laplacian

Contrast




```
the common genres of Group1 [['Action', 65], ['Adventure', 59], ['Thriller', 54]]
the common genres of Group2 [['Drama', 93], ['Comedy', 44], ['Romance', 31]]
```

Group1 are more exciting whose genres are similar. Group2 are more relaxing whose genres are also similar but have sharp contrast to group1.

Actor Relations

- Genres preference

- Cosine Distance

Missing values

- Title years
- Keywords
- Revenue

```
feature
```

actor_id	Crime	Drama	History	Fantasy	Western	Foreign	Science Fiction	Family	Thriller	War	Mystery	Adventure	TV Movie	Docume
2.0	4	6	1	0	0	1	1	1	6	0	1	5	0	
3.0	3	12	0	0	0	0	0	2	4	0	3	1	0	
5.0	0	8	0	0	3	0	2	1	3	3	1	1	0	
31.0	2	13	0	4	0	0	1	5	3	0	2	2	0	
35.0	2	6	1	1	0	1	1	2	5	0	2	2	0	
40.0	3	12	0	5	0	1	5	2	6	0	3	4	0	
48.0	1	7	0	1	0	1	1	2	1	0	0	1	0	
50.0	1	11	1	1	0	0	0	2	2	2	1	3	0	

```
index = np.where(weights!=0)
```

```
x=index[0][0:5]
print(x)
y=index[1][0:5]
print(y)
```

```
[0 0 1 1 2]
[ 64 246 438 492  97]
```

```
print(df_actor[df_actor['actor_id_1']==actor_id[0]]['actor_name_1'].iloc[0])
df_actor[df_actor['actor_id_1']==actor_id[0]]['title']
```

```
Mark Hamill

256                                Star Wars
1167                       Return of the Jedi
9423                    Comic Book: The Movie
10936                         Corvette Summer
18865    Dante's Inferno: An Animated Epic
20105             Kevin Smith: Burn in Hell
Name: title, dtype: object
```

```
print(df_actor[df_actor['actor_id_1']==actor_id[64]]['actor_name_1'].iloc[0])
df_actor[df_actor['actor_id_1']==actor_id[64]]['title']
```

```
Russell Crowe

517            Romper Stomper
2773         Mystery, Alaska
3456               Gladiator
4865        A Beautiful Mind
5886             Breaking Up
9084             No Way Back
12042            3:10 to Yuma
25193               Bastards
```

Recommendation Engine Based On Movie

Two steps

- Similarity

Determine N (N=30) films with a content similar to the movie that provided by the user according to the Euclidean distance.

- Popularity

Select the 5 most popular films from N films according to a scoring method.

Similarity

- Extract the features of movie, such as director name, actor names, genres, and key words (1200)

movie title	director	actor 1	actor 2	actor 3	keyword 1	keyword 2	genre 1	genre 2	...	genre k
Film 1	a_{11}	a_{12}			...					a_{1q}
...					...					
Film i	a_{i1}	a_{i2}			a_{ij}					a_{iq}
...					...					
Film p	a_{p1}	a_{p2}			...					a_{pq}

- Compare these features of each movie with the features of the selected movie. (aij = 1 or 0)
- Calculate Euclidean distance between every two films.

$$d_{m,n} = \sqrt{\sum_{i=1}^{N} \left(a_{m,i} - a_{n,i}\right)^2}$$

- Select the N(N=30) films which are the closest from selected movie.

Popularity

- Use a scoring method according to 3 criteria:

  The IMDB score, The number of votes, The year of release

- Calculate the score according to the formula:

$$\text{score} = IMDB^2 \times \phi_{\sigma_1,c_1} \times \phi_{\sigma_2,c_2} \qquad \phi_{\sigma,c}(x) \propto \exp\left(-\frac{(x-c)^2}{2\,\sigma^2}\right)$$

  For votes, $\sigma 1=c1=$ maximum number of votes.
  For years, $\sigma 1=20$ and center the gaussian on the title year of the selected film.

- Select the 5 movies with highest score

Making meaningful recommendations

- Issue: the existence of sequel

```
____________________________________________________________________________________________________
Recommendation: Films similar to id=2052 -> title: 'The NeverEnding Story'  genres:'Drama|Family|Fantasy|Adventure'.
This film is about: While hiding from bullies in his school's attic, a young boy discovers the extraordinary land of Fantasia, through a mag
ical book called The Neverending Story. The book tells the tale of Atreyu, a young warrior who, with the help of a luck dragon named Falkor,
must save Fantasia from the destruction of The Nothing.
nº1      -> Harry Potter and the Philosopher's Stone
nº2      -> Harry Potter and the Chamber of Secrets
nº3      -> Harry Potter and the Prisoner of Azkaban
nº4      -> Jumanji
nº5      -> Harry Potter and the Goblet of Fire
```

- Solution: check the degree of similarity of two film titles.

```
____________________________________________________________________________________________________
Recommendation: Films similar to id=2052 -> title: 'The NeverEnding Story'  genres:'Drama|Family|Fantasy|Adventure'.
This film is about: While hiding from bullies in his school's attic, a young boy discovers the extraordinary land of Fantasia, through a mag
ical book called The Neverending Story. The book tells the tale of Atreyu, a young warrior who, with the help of a luck dragon named Falkor,
must save Fantasia from the destruction of The Nothing.
nº1      -> Harry Potter and the Philosopher's Stone
nº2      -> Jumanji
nº3      -> A Little Princess
nº4      -> Clash of the Titans
nº5      -> Ronja Robbersdaughter
```

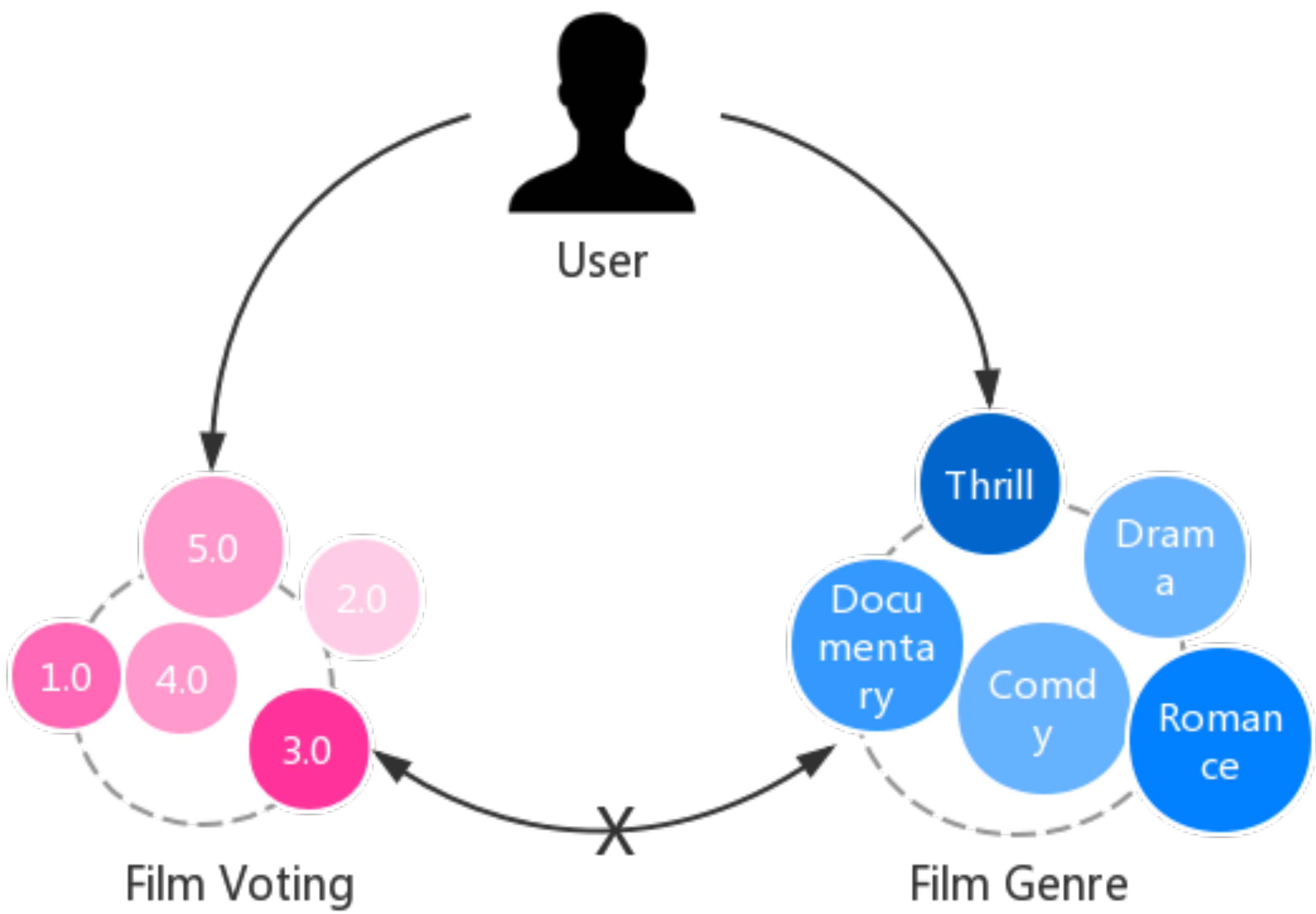
Recommendation: Films similar to id=2052 -> title: 'The NeverEnding Story' genres:'Drama|Family|Fantasy|Adventure'.
This film is about: While hiding from bullies in his school's attic, a young boy discovers the extraordinary land of Fantasia, through a magical book called The Neverending Story. The book tells the tale of Atreyu, a young warrior who, with the help of a luck dragon named Falkor, must save Fantasia from the destruction of The Nothing.

```
nº1     -> Harry Potter and the Philosopher's Stone
nº2     -> Jumanji
nº3     -> A Little Princess
nº4     -> Clash of the Titans
nº5     -> Ronja Robbersdaughter
```

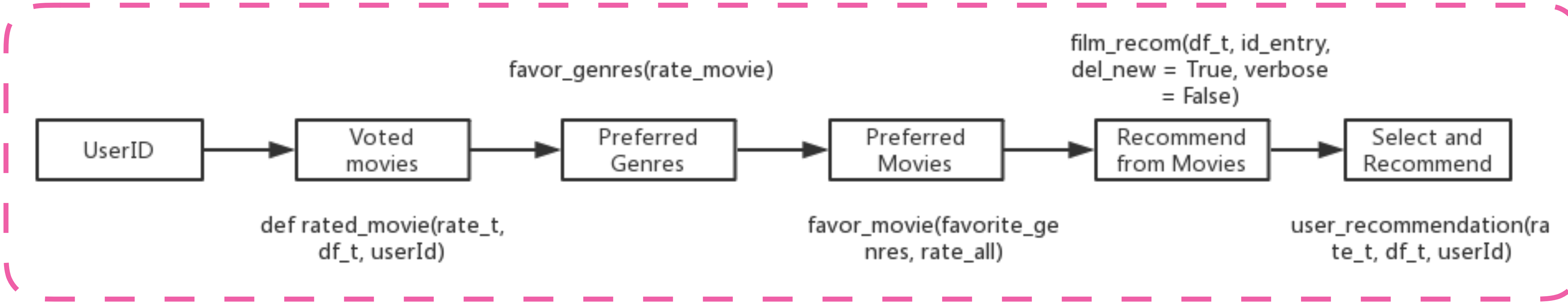
```
[["Harry Potter and the Philosopher's Stone",
  "Harry Potter has lived under the stairs at his aunt and uncle's house his whole life. But on his 11th birthday, he learns he's a powerful wizard -- with a place waiting for him at the Hogwarts School of Witchcraft and Wizardry. As he learns to harness his newfound powers with the help of the school's kindly headmaster, Harry uncovers the truth about his parents' deaths -- and about the villain who's to blame.",
  'Adventure|Fantasy|Family'],
 ['Jumanji',
  "When siblings Judy and Peter discover an enchanted board game that opens the door to a magical world, they unwittingly invite Alan -- an adult who's been trapped inside the game for 26 years -- into their living room. Alan's only hope for freedom is to finish the game, which proves risky as all three find themselves running from giant rhinoceroses, evil monkeys and other terrifying creatures.",
  'Adventure|Fantasy|Family'],
 ['A Little Princess',
  "When her father enlists to fight for the British in WWI, young Sara Crewe goes to New York to attend the same boarding school her late mother attended. She soon clashes with the severe headmistress, Miss Minchin, who attempts to stifle Sara's creativity and sense of self- worth.",
  'Drama|Family|Fantasy'],
 ['Clash of the Titans',
  "To win the right to marry his love, the beautiful princess Andromeda, and fulfil his destiny, Perseus must complete various tasks including taming Pegasus, capturing Medusa's head, and battling the Kraken monster.",
  'Adventure|Fantasy|Family'],
 ['Ronja Robbersdaughter',
  "Ronya lives happily in her father's castle until she comes across a new playmate, Birk, in the nearby dark forest. The two explore the wilderness, braving dangerous Witchbirds and Rump-Gnomes. But when their families find out Birk and Ronja have been playing together, they forbid them to see each other again. Indeed, their fathers are competing robber chieftains and bitter enemies. Now the two spunky children must try to tear down the barriers that have kept their families apart for so long.",
  'Adventure|Drama|Fantasy|Family']]
```

Recommendation Engine Based On User

Consider user’s preference



Recommendation Engine Based On User



- Whole process and functions

- Recommendation results

```
df3=df.copy(deep=True)
rate3=rate.copy(deep=True)
user_recommendation(rate3, df3, 10000)
```

```
It seems that you really like watching Drama and History !
You may like these movies:
nº1       -> Trench of Hope
nº2       -> The Founder

[['Trench of Hope', 8.0], ['The Founder', 7.0]]
```

```
df3=df.copy(deep=True)
rate3=rate.copy(deep=True)
user_recommendation(rate3, df3, 222)
```

```
It seems that you really like watching Drama and Thriller !
You may like these movies:
nº1       -> The Violent Professionals
nº2       -> The Formula
You may like these movies:
nº1       -> The Hunt
nº2       -> This Is England

[['The Violent Professionals', 6.0],
 ['The Formula', 6.0],
 ['The Hunt', 7.9000000000000004],
 ['This Is England', 7.4000000000000004]]
```

Conclusion

- the language of the film was not checked: in fact, this could be important to get sure that the films recommended are in the same language than the one chosen by the user

- some sequels to films may don't share similar titles (e.g. James Bond series)

- if possible, the original data can be separated into two sets, one for training and the other for testing. This can carry out a more direct result of how well the recommendation system works.

- the recommendation engine based on user can also include some other factors, e.x. watching year, preferable language.