In [4]:
```
import os
from tqdm import tqdm
import pandas as pd
import numpy as np

from sklearn.model_selection import train_test_split
from sklearn.metrics import mean_squared_error

import warnings
warnings.filterwarnings('ignore')
```

MovieLens 데이터셋 불러오기

In [2]:
```
root_path = os.getcwd()
path = os.path.join(root_path, 'data/ml-latest-small/')
```

In [3]:
```
ratings_df = pd.read_csv(os.path.join(path, 'ratings.csv'), encoding='utf-8')
tags_df = pd.read_csv(os.path.join(path, 'tags.csv'), encoding='utf-8')
movies_df = pd.read_csv(os.path.join(path, 'movies.csv'), index_col='movieId', encodi
```

In [6]:
```
tags_df.head()
```

Out[6]:

	userId	movieId	tag	timestamp
0	2	60756	funny	1445714994
1	2	60756	Highly quotable	1445714996
2	2	60756	will ferrell	1445714992
3	2	89774	Boxing story	1445715207
4	2	89774	MMA	1445715200

In [7]:
```
movies_df.head()
```

Out[7]:

	title	genres
movieId		
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

Genres 를 이용한 movie representation

In [35]:
```
total_count = len(movies_df.index)
total_genres = list(set([genre for sublist in list(map(lambda x: x.split('|'), movie
total_genres
```

Out[35]:
```
['Mystery',
 'Documentary',
 '(no genres listed)',
 'Fantasy',
 'Thriller',
 'Sci-Fi',
 'Children',
 'War',
 'Romance',
 'Western',
 'IMAX',
 'Crime',
 'Adventure',
 'Drama',
 'Musical',
 'Action',
 'Horror',
 'Comedy',
 'Film-Noir',
 'Animation']
```

In [40]:
```
genre_count = dict.fromkeys(total_genres)

for each_genre_list in movies_df['genres']:
    for genre in each_genre_list.split('|'):
        if genre_count[genre] == None:
            genre_count[genre] = 1
        else:
            genre_count[genre] = genre_count[genre] + 1
```

In [41]:
```
genre_count
```

Out[41]:
```
{'Mystery': 573,
 'Documentary': 440,
 '(no genres listed)': 34,
 'Fantasy': 779,
 'Thriller': 1894,
 'Sci-Fi': 980,
 'Children': 664,
 'War': 382,
 'Romance': 1596,
 'Western': 167,
 'IMAX': 158,
 'Crime': 1199,
 'Adventure': 1263,
 'Drama': 4361,
 'Musical': 334,
 'Action': 1828,
 'Horror': 978,
 'Comedy': 3756,
 'Film-Noir': 87,
 'Animation': 611}
```

In [42]:
```
for each_genre in genre_count:
    genre_count[each_genre] = np.log10(total_count/genre_count[each_genre])

genre_count
```

Out[42]:
```
{'Mystery': 1.2304935032683613,
 'Documentary': 1.3451954487495636,
 '(no genres listed)': 2.457169208193496,
 'Fantasy': 1.0971106675631865,
 'Thriller': 0.7112681505684965,
 'Sci-Fi': 0.9974220495432563,
```

```
 'Children': 1.1664800458677336,
 'War': 1.4065847623240424,
 'Romance': 0.7856152382210405,
 'Western': 1.7659316540881678,
 'IMAX': 1.7899910382813284,
 'Crime': 0.9098289421369025,
 'Adventure': 0.8872447746804204,
 'Drama': 0.3490620385623247,
 'Musical': 1.4649016584241867,
 'Action': 0.7266719338379385,
 'Horror': 0.9983092704481497,
 'Comedy': 0.4139225416416778,
 'Film-Noir': 2.0491288726171324,
 'Animation': 1.2026069149931968}
```

genre 를 이용한 Movie representation 생성

In [43]:

```
genre_representation = pd.DataFrame(columns=sorted(total_genres), index=movies_df.ind

for index, each_row in tqdm(movies_df.iterrows()):
    dict_temp = {i: genre_count[i] for i in each_row['genres'].split('|')}
    row_to_add = pd.DataFrame(dict_temp, index=[index])
    genre_representation.update(row_to_add)
```

```
9742it [00:31, 312.03it/s]
```

In [44]:

```
genre_representation
```

Out[44]:

movieId	(no genres listed)	Action	Adventure	Animation	Children	Comedy	Crime	Documentary	Dram
1	NaN	NaN	0.887245	1.202607	1.16648	0.413923	NaN	NaN	Na
2	NaN	NaN	0.887245	NaN	1.16648	NaN	NaN	NaN	Na
3	NaN	NaN	NaN	NaN	NaN	0.413923	NaN	NaN	Na
4	NaN	NaN	NaN	NaN	NaN	0.413923	NaN	NaN	0.34906
5	NaN	NaN	NaN	NaN	NaN	0.413923	NaN	NaN	Na
...	...	...	...	...	...	...	...	...	
193581	NaN	0.726672	NaN	1.202607	NaN	0.413923	NaN	NaN	Na
193583	NaN	NaN	NaN	1.202607	NaN	0.413923	NaN	NaN	Na
193585	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	0.34906
193587	NaN	0.726672	NaN	1.202607	NaN	NaN	NaN	NaN	Na
193609	NaN	NaN	NaN	NaN	NaN	0.413923	NaN	NaN	Na

9742 rows × 20 columns

Tag를 이용한 Movie representation 생성

In [46]:

```
tags_df.head()
```

Out[46]:

	userId	movieId	tag	timestamp
0	2	60756	funny	1445714994
1	2	60756	Highly quotable	1445714996
2	2	60756	will ferrell	1445714992
3	2	89774	Boxing story	1445715207
4	2	89774	MMA	1445715200

In [49]:

```
movies_df.loc[60756] # Comedy -> funny : Reasonable !
```

Out[49]:

```
title      Step Brothers (2008)
genres                   Comedy
Name: 60756, dtype: object
```

In [59]:

```
# get unique tag
tag_column = list(map(lambda x: x.split(','), tags_df['tag']))
unique_tags = list(set(list(map(lambda x:x.strip(), list([tag for sublist in tag_col
print(len(unique_tags))
```

```
1589
```

In [62]:

```
### Compute IDF for tag
total_movie_count = len(set(tags_df['movieId']))
# key: tag, value: number of movies with such tag
tag_count_dict = dict.fromkeys(unique_tags)

for each_movie_tag_list in tags_df['tag']:
    for tag in each_movie_tag_list.split(','):
        if tag_count_dict[tag.strip()] == None:
            tag_count_dict[tag.strip()] = 1
        else:
            tag_count_dict[tag.strip()] += 1

tag_idf = dict()
for each_tag in tag_count_dict:
    tag_idf[each_tag] = np.log10(total_movie_count / tag_count_dict[each_tag])

tag_idf
```

Out[62]:

```
{'Sean Connery': 3.196452541703389,
 'Tarantino': 2.7193312869837265,
 'Stoner Movie': 3.196452541703389,
 'Jude Law': 2.895422546039408,
 'stupid': 3.196452541703389,
 'unusual': 3.196452541703389,
 'Hungary': 3.196452541703389,
 'Boxing story': 3.196452541703389,
 'black-and-white': 3.196452541703389,
 'lawn mower': 3.196452541703389,
 'masculinity': 3.196452541703389,
 'terminal illness': 2.7193312869837265,
 'insomnia': 3.196452541703389,
 'wonderwoman': 3.196452541703389,
 'missing children': 3.196452541703389,
 'Everything you want is here': 3.196452541703389,
 'dumpster diving': 3.196452541703389,
 'family': 2.351354501689132,
 'wapendrama': 3.196452541703389,
 'iconic': 3.196452541703389,
```

```
 'celebrity fetishism': 3.196452541703389,
 'sentimental': 2.895422546039408,
 'intense': 2.5943925503754266,
 'homosexuality': 3.196452541703389,
 'stupid but funny': 3.196452541703389,
 'King Arthur': 2.5943925503754266,
 'scenic': 3.196452541703389,
 'wizards': 3.196452541703389,
 'Tolstoy': 3.196452541703389,
 'Arthur C. Clarke': 3.196452541703389,
 'jack nicholson': 2.895422546039408,
 'Visually appealing': 3.196452541703389,
 'nerds': 3.196452541703389,
 'Hugh Jackman': 2.895422546039408,
 'Bill Murray': 3.196452541703389,
 'adventure': 2.2933625547114453,
 'free speech': 3.196452541703389,
 'based on a TV show': 2.5943925503754266,
 'Police': 3.196452541703389,
 'fighting': 3.196452541703389,
 'depressing': 2.5943925503754266,
 'innovative': 3.196452541703389,
 'marijuana': 3.196452541703389,
 'imagination': 2.895422546039408,
 'cult': 2.895422546039408,
 '"artsy"': 3.196452541703389,
 'relaxing': 3.196452541703389,
 'Dogs': 3.196452541703389,
 'virtual reality': 3.196452541703389,
 'Indonesia': 3.196452541703389,
 'space': 2.050324506025151,
 'Henry Darger': 3.196452541703389,
 'Simon Pegg': 3.196452541703389,
 'money': 3.196452541703389,
 'Al Pacino': 2.4974825373673704,
 'unpredictable': 3.196452541703389,
 'Sci-Fi': 3.196452541703389,
 'sports': 2.4974825373673704,
 'start of a beautiful friendship': 3.196452541703389,
 'big wave': 3.196452541703389,
 'Poor plot development': 3.196452541703389,
 'surfing': 2.895422546039408,
 'show business': 2.5943925503754266,
 'Rachel Weisz': 2.7193312869837265,
 'mathematics': 2.7193312869837265,
 'Medieval': 3.196452541703389,
 'characters': 2.895422546039408,
 'Recap': 3.196452541703389,
 'boksdrama': 3.196452541703389,
 'irony': 3.196452541703389,
 'fast paced': 3.196452541703389,
 'deafness': 3.196452541703389,
 'moody': 2.895422546039408,
 'basketball': 2.5943925503754266,
 'Istanbul': 3.196452541703389,
 'Maggie Gyllenhaal': 3.196452541703389,
 'loneliness': 2.5943925503754266,
 'spiders': 3.196452541703389,
 'TERRORISM': 3.196452541703389,
 'Conan': 3.196452541703389,
 'fantasy world': 2.7193312869837265,
 'jungle': 3.196452541703389,
 'Rita Hayworth can dance!': 3.196452541703389,
 'KIDNAPPING': 3.196452541703389,
 'challenging': 3.196452541703389,
 'claymation': 2.895422546039408,
 'Amtrak': 3.196452541703389,
 'rebellion': 3.196452541703389,
 'cerebral': 2.4183012913197452,
```

```
 'financial crisis': 3.196452541703389,
 'CGI': 3.196452541703389,
 'Bette Davis': 3.196452541703389,
 'dreams': 3.196452541703389,
 'daniel radcliffe': 3.196452541703389,
 '1980s': 2.895422546039408,
 'heroin': 2.895422546039408,
 'FIGHTING THE SYSTEM': 3.196452541703389,
 'World War II': 2.2422100322640643,
 'Rachel McAdams': 3.196452541703389,
 'unlikely hero': 3.196452541703389,
 'r:some violence': 3.196452541703389,
 'tricky': 3.196452541703389,
 'pixar': 2.895422546039408,
 'Star Trek': 2.895422546039408,
 'science fiction': 3.196452541703389,
 'Captain Kirk': 3.196452541703389,
 'Dickens': 2.4974825373673704,
 'bromantic': 3.196452541703389,
 'wine': 3.196452541703389,
 'wistful': 3.196452541703389,
 'schizophrenia': 2.895422546039408,
 'new society': 3.196452541703389,
 'opera': 3.196452541703389,
 'England': 2.196452541703389,
 'harry potter': 3.196452541703389,
 'Dark': 3.196452541703389,
 'Lou Gehrig': 3.196452541703389,
 'adorable': 3.196452541703389,
 'freedom': 3.196452541703389,
 'con artists': 3.196452541703389,
 'intimate': 3.196452541703389,
 'Chile': 3.196452541703389,
 'new york': 2.895422546039408,
 'Judaism': 2.4974825373673704,
 'Huey Long': 3.196452541703389,
 'colorful': 3.196452541703389,
 'intelligent sci-fi': 3.196452541703389,
 'hallucinatory': 2.2422100322640643,
 'Charlotte Bronte': 3.196452541703389,
 'needed more autobots': 3.196452541703389,
 'confrontational': 2.895422546039408,
 'monologue': 3.196452541703389,
 'courtroom drama': 2.895422546039408,
 'predictable': 2.351354501689132,
 'DEPRESSING': 3.196452541703389,
 'Mindfuck': 3.196452541703389,
 'Olympics': 2.895422546039408,
 'Studio Ghibli': 2.895422546039408,
 'Teen movie': 3.196452541703389,
 'bad writing': 3.196452541703389,
 'emma thompson': 3.196452541703389,
 'tension building': 3.196452541703389,
 'heroic bloodshed': 3.196452541703389,
 'psychological': 2.155059856545164,
 'Turkey': 3.196452541703389,
 'E. M. Forster': 3.196452541703389,
 'Vulgar': 3.196452541703389,
 'Bad story': 3.196452541703389,
 'imdb top 250': 2.155059856545164,
 'big top': 3.196452541703389,
 'Holy Grail': 3.196452541703389,
 'coma': 2.895422546039408,
 'Hepburn and Tracy': 2.4974825373673704,
 'indiana jones': 2.895422546039408,
 'quirky romantic': 3.196452541703389,
 'seen more than once': 2.895422546039408,
 'Dull': 3.196452541703389,
 'humor': 2.895422546039408,
```

```
 'POW': 2.895422546039408,
 'Rogue': 3.196452541703389,
 'David Fincher': 3.196452541703389,
 'star wars': 3.196452541703389,
 'Amish': 2.895422546039408,
 'Katzanzakis': 3.196452541703389,
 'fast-paced': 3.196452541703389,
 'unintelligent': 3.196452541703389,
 'pigs': 3.196452541703389,
 'diabetes': 3.196452541703389,
 'r:sustained strong stylized violence': 3.196452541703389,
 'first was much better': 3.196452541703389,
 'awesome': 2.895422546039408,
 'Brittany Murphy': 3.196452541703389,
 'Christopher Lloyd': 2.895422546039408,
 'Well Done': 3.196452541703389,
 'drugs & music': 3.196452541703389,
 'symbolism': 2.895422546039408,
 'weather forecaster': 3.196452541703389,
 'Bad writing': 3.196452541703389,
 'deadpan': 3.196452541703389,
 'Justin Timberlake': 3.196452541703389,
 'android(s)/cyborg(s)': 3.196452541703389,
 'great screenplay': 3.196452541703389,
 'twins': 2.4183012913197452,
 'hilarious': 2.7193312869837265,
 'procedural': 3.196452541703389,
 'cult classic': 3.196452541703389,
 'gun tactics': 3.196452541703389,
 'Enterprise': 3.196452541703389,
 'alter ego': 3.196452541703389,
 'beautifully filmed': 3.196452541703389,
 'jazz': 2.895422546039408,
 'classic': 2.155059856545164,
 'Quirky': 3.196452541703389,
 'Roger Avary': 3.196452541703389,
 'travolta': 3.196452541703389,
 'real estate': 3.196452541703389,
 'goofy': 2.895422546039408,
 'lies': 3.196452541703389,
 'great cinematography': 3.196452541703389,
 'anger': 3.196452541703389,
 'Mystery': 2.7193312869837265,
 'abortion': 3.196452541703389,
 'alternate universe': 2.895422546039408,
 'drugs': 2.196452541703389,
 'cate blanchett': 3.196452541703389,
 'creepy': 2.2422100322640643,
 'good dialogue': 2.4974825373673704,
 'martial arts': 2.2933625547114453,
 'Ghosts': 3.196452541703389,
 'gintama': 3.196452541703389,
 'mafia': 2.7193312869837265,
 'diner': 3.196452541703389,
 'British': 3.196452541703389,
 'political right versus left': 3.196452541703389,
 'shark': 3.196452541703389,
 'Epic': 3.196452541703389,
 'Francis Ford Coppola': 3.196452541703389,
 'submarine': 2.895422546039408,
 'mythology': 3.196452541703389,
 'Afghanistan': 3.196452541703389,
 'costume drama': 3.196452541703389,
 'r:strong language': 3.196452541703389,
 'AIDs': 2.895422546039408,
 'tragedy': 3.196452541703389,
 'vertriloquism': 3.196452541703389,
 'butler': 3.196452541703389,
 'scandal': 2.895422546039408,
```

```
'reciprocal spectator': 2.895422546039408,
'classic sci-fi': 2.7193312869837265,
'nonlinear storyline': 3.196452541703389,
'trippy': 2.895422546039408,
'Harper Lee': 3.196452541703389,
'gruesome': 3.196452541703389,
'matchmaker': 3.196452541703389,
'sofia coppola': 3.196452541703389,
'prostitution': 2.351354501689132,
'Great villain': 3.196452541703389,
'disappointing': 3.196452541703389,
'undercover cop': 3.196452541703389,
'nonlinear': 2.7193312869837265,
'Christmas': 2.2933625547114453,
'futuristic': 2.895422546039408,
'military': 2.2422100322640643,
'Jaime Pressly': 3.196452541703389,
'setting:space/space ship': 3.196452541703389,
'Neil Patrick Harris': 3.196452541703389,
'blindness': 2.5943925503754266,
'ransom': 3.196452541703389,
'2D animation': 3.196452541703389,
'milkshake': 3.196452541703389,
'ryan reynolds': 3.196452541703389,
'SNL': 3.196452541703389,
'HORRIBLE ACTING': 3.196452541703389,
'Nick Hornby': 2.895422546039408,
'alternate endings': 2.895422546039408,
'Guns': 3.196452541703389,
'lion': 3.196452541703389,
'brothers': 3.196452541703389,
'meditative': 2.895422546039408,
'stylish': 2.4974825373673704,
'violence in america': 2.895422546039408,
'black humor': 3.196452541703389,
'satirical': 2.895422546039408,
'organized crime': 2.4974825373673704,
'Chuck Palahniuk': 3.196452541703389,
'cameo:Whoopi Goldberg': 3.196452541703389,
'unexplained': 3.196452541703389,
'haunting': 3.196452541703389,
'terrorism': 2.4974825373673704,
'silly': 2.7193312869837265,
'humour': 3.196452541703389,
'Dwayne Johnson': 3.196452541703389,
'DC Comics': 3.196452541703389,
'Peta Wilson': 3.196452541703389,
'gold': 3.196452541703389,
'irreverent': 2.7193312869837265,
'great ending': 2.895422546039408,
'bad language': 3.196452541703389,
'carnival': 3.196452541703389,
'No DVD at Netflix': 3.196452541703389,
'horror': 2.5943925503754266,
'original': 2.895422546039408,
'psychology': 1.8742332469694698,
'character development': 3.196452541703389,
'Thor': 3.196452541703389,
'great humor': 3.196452541703389,
'General Motors': 3.196452541703389,
'biography': 3.196452541703389,
'camp': 3.196452541703389,
'gentle': 3.196452541703389,
'notable soundtrack': 2.895422546039408,
'Gangs': 3.196452541703389,
'Morrow': 3.196452541703389,
'photography': 2.7193312869837265,
'Alfred Hitchcock': 2.5943925503754266,
'last man on earth': 2.895422546039408,
```

```
'Seann William Scott': 2.895422546039408,
'longing': 3.196452541703389,
'men in drag': 2.4183012913197452,
'Beautiful': 2.7193312869837265,
'remaster': 3.196452541703389,
'Mark Wahlberg': 3.196452541703389,
'It was melodramatic and kind of dumb': 3.196452541703389,
'mindfuck': 2.050324506025151,
'Nerd': 3.196452541703389,
'e-mail': 3.196452541703389,
'father-son relationship': 3.196452541703389,
'dating': 2.895422546039408,
'justice': 3.196452541703389,
'John Grisham': 2.5943925503754266,
'fucked up': 3.196452541703389,
'Disaster': 3.196452541703389,
'ships': 3.196452541703389,
'McDonalds': 3.196452541703389,
'out of order': 3.196452541703389,
'weddings': 2.895422546039408,
'amazing artwork': 3.196452541703389,
'stiller': 3.196452541703389,
'bad ass': 3.196452541703389,
'audience intelligence underestimated': 2.895422546039408,
'tense': 2.2422100322640643,
'test tag': 3.196452541703389,
'Suspense': 3.196452541703389,
'music': 1.9923325590474643,
'bromance': 2.895422546039408,
'stone age': 3.196452541703389,
'dance marathon': 3.196452541703389,
'Henry James': 2.895422546039408,
'dinosaurs': 2.895422546039408,
'MacBeth': 3.196452541703389,
'Existential': 3.196452541703389,
'confusing ending': 3.196452541703389,
'pizza beer': 3.196452541703389,
'preacher': 3.196452541703389,
'true story': 2.5943925503754266,
'Family': 3.196452541703389,
'Josh Brolin': 3.196452541703389,
'Borg': 3.196452541703389,
'amazing': 3.196452541703389,
"Eugene O'Neill": 3.196452541703389,
'biking': 3.196452541703389,
'aging': 3.196452541703389,
'Christopher Nolan': 2.7193312869837265,
'postmodern': 3.196452541703389,
'interracial marriage': 3.196452541703389,
'Lonesome Polecat': 3.196452541703389,
'apes': 3.196452541703389,
'annoying': 3.196452541703389,
'r:disturbing violent images': 3.196452541703389,
'Luc Besson': 3.196452541703389,
'Mount Rushmore': 3.196452541703389,
'Journalism': 3.196452541703389,
'retro': 2.895422546039408,
'meryl streep': 3.196452541703389,
'Sexual Humor': 3.196452541703389,
'jay and silent bob': 3.196452541703389,
'Johnny Cash': 3.196452541703389,
'mice': 3.196452541703389,
'golfing': 3.196452541703389,
'Animal movie': 2.351354501689132,
'Mark Ruffalo': 2.895422546039408,
'revolutionary': 3.196452541703389,
'boxing': 2.155059856545164,
'very funny': 2.895422546039408,
'purposefulness': 3.196452541703389,
```

```
'beautiful cinematography': 3.196452541703389,
'James Stewart': 2.895422546039408,
'Dialogue': 3.196452541703389,
'Matt Damon': 3.196452541703389,
'human rights': 3.196452541703389,
'India': 2.196452541703389,
'daniel craig': 3.196452541703389,
'Insanity': 3.196452541703389,
'Sinbad': 3.196452541703389,
'planes': 3.196452541703389,
'L.A.': 2.895422546039408,
'nocturnal': 3.196452541703389,
'Ben Stiller': 2.5943925503754266,
'Norman Bates': 3.196452541703389,
'amnesia': 2.4974825373673704,
'younger men': 3.196452541703389,
'Robert De Niro': 2.7193312869837265,
'non-linear': 2.895422546039408,
'Dumas': 3.196452541703389,
'death': 2.4183012913197452,
'lack of development': 3.196452541703389,
'best comedy': 3.196452541703389,
'based on a book': 2.4974825373673704,
'bad music': 3.196452541703389,
'spies': 3.196452541703389,
'Chris Klein': 3.196452541703389,
'intertwining storylines': 3.196452541703389,
'Japan': 2.7193312869837265,
'MMA': 3.196452541703389,
'Romans': 3.196452541703389,
'action choreography': 3.196452541703389,
'pulp': 3.196452541703389,
'lieutenant dan': 3.196452541703389,
'bloody': 2.7193312869837265,
'space epic': 3.196452541703389,
'mystery': 2.4183012913197452,
'figure skating': 2.895422546039408,
'Tom Hanks': 2.5943925503754266,
'zither': 3.196452541703389,
'teenagers': 3.196452541703389,
'Anne Boleyn': 3.196452541703389,
'Great Visuals': 3.196452541703389,
'Hilary Swank': 3.196452541703389,
'Nuclear disaster': 3.196452541703389,
'unnecessary sequel': 3.196452541703389,
'Renee Zellweger': 3.196452541703389,
'holocaust': 2.895422546039408,
'multiple storylines': 2.7193312869837265,
'a dingo ate my baby': 3.196452541703389,
'Action': 2.5943925503754266,
'Michael Bay': 2.895422546039408,
'Russia': 3.196452541703389,
'Inigo Montoya': 3.196452541703389,
'cartoon': 3.196452541703389,
'oldie but goodie': 3.196452541703389,
'uplifting': 3.196452541703389,
'cyberpunk': 2.895422546039408,
'police': 2.351354501689132,
'voyeurism': 3.196452541703389,
'neo-noir': 2.895422546039408,
'Robert Downey Jr.': 2.895422546039408,
'Politics': 3.196452541703389,
'murder': 2.2422100322640643,
'Aardman': 2.5943925503754266,
'cool': 3.196452541703389,
'gun-fu': 3.196452541703389,
'autism': 3.196452541703389,
'Salieri': 3.196452541703389,
'suburbia': 3.196452541703389,
```

```
 'ocean': 3.196452541703389,
 'sci-fi': 1.8742332469694698,
 'ex-con': 3.196452541703389,
 'faerie tale': 3.196452541703389,
 'britpop': 3.196452541703389,
 'Beatles': 3.196452541703389,
 'American propaganda': 3.196452541703389,
 'mind-blowing': 3.196452541703389,
 'memory': 2.5943925503754266,
 'bad': 2.2933625547114453,
 'Holocaust': 2.155059856545164,
 'Disney': 1.8347247056857963,
 'whimsical': 2.7193312869837265,
 'casino': 3.196452541703389,
 'doctors': 2.895422546039408,
 'post-apocalyptic': 2.4974825373673704,
 'Jesse Eisenberg': 3.196452541703389,
 'humorous': 2.4974825373673704,
 'bubba gump shrimp': 3.196452541703389,
 'black and white': 2.4183012913197452,
 'nightclub': 2.7193312869837265,
 'Michael Cera': 2.7193312869837265,
 'Wolverine': 3.196452541703389,
 'bad dialogue': 3.196452541703389,
 'television': 2.4183012913197452,
 'TOGA': 3.196452541703389,
 'quick cuts': 3.196452541703389,
 'acting': 3.196452541703389,
 'organised crime': 3.196452541703389,
 'Nabokov': 3.196452541703389,
 'Crude humor': 3.196452541703389,
 'BEST PICTURE': 3.196452541703389,
 'blood splatters': 3.196452541703389,
 '1900s': 3.196452541703389,
 'assassin-in-training (scene)': 3.196452541703389,
 'Oscar Wilde': 3.196452541703389,
 'deaf': 3.196452541703389,
 'Anne Hathaway': 2.895422546039408,
 'big budget': 2.895422546039408,
 'circus': 2.5943925503754266,
 'crude humor': 2.7193312869837265,
 'paranoia': 2.4974825373673704,
 'kung fu': 3.196452541703389,
 'emotional': 2.082509189396552,
 'Cole Porter': 3.196452541703389,
 'immortality': 3.196452541703389,
 'Andy Kaufman': 3.196452541703389,
 'fugitive': 2.895422546039408,
 'prequel': 3.196452541703389,
 'Steve Carell': 2.7193312869837265,
 'claims to be true': 3.196452541703389,
 'Eric Bana': 3.196452541703389,
 'con men': 3.196452541703389,
 'wry': 3.196452541703389,
 'AS Byatt': 3.196452541703389,
 'whales': 3.196452541703389,
 'original plot': 2.895422546039408,
 'Hawkeye': 3.196452541703389,
 'fatherhood': 2.5943925503754266,
 'casual violence': 3.196452541703389,
 'tension': 3.196452541703389,
 'lord of the rings': 3.196452541703389,
 'Moving': 3.196452541703389,
 'conspiracy': 3.196452541703389,
 'secret society': 2.895422546039408,
 'reflective': 2.895422546039408,
 'superb soundtrack': 3.196452541703389,
 'directorial debut': 3.196452541703389,
 'societal criticism': 3.196452541703389,
```

```
 'artsy': 2.895422546039408,
 'rape': 2.895422546039408,
 'Empire State Building': 3.196452541703389,
 'British gangster': 3.196452541703389,
 'big name actors': 3.196452541703389,
 'truth': 3.196452541703389,
 'violence': 2.2933625547114453,
 'game': 3.196452541703389,
 'adoption': 2.895422546039408,
 'elegant': 3.196452541703389,
 'shipwreck': 3.196452541703389,
 'children': 2.7193312869837265,
 'space action': 2.895422546039408,
 'Atmospheric': 2.4974825373673704,
 'Italy': 2.895422546039408,
 'Jared Leto': 2.895422546039408,
 'Christina Ricci': 2.895422546039408,
 'Sundance award winner': 3.196452541703389,
 'Mrs. DeWinter': 3.196452541703389,
 'dreamlike': 2.050324506025151,
 'happpiness': 3.196452541703389,
 'gambling': 2.4974825373673704,
 'Thanos': 3.196452541703389,
 'camels': 3.196452541703389,
 'Butler': 3.196452541703389,
 'Marx brothers': 2.895422546039408,
 'christmas': 2.5943925503754266,
 'love story': 2.7193312869837265,
 'random': 3.196452541703389,
 'Nicolas Cage': 3.196452541703389,
 'beat poetry': 3.196452541703389,
 'corruption': 2.4974825373673704,
 'Western': 3.196452541703389,
 'mobsters': 3.196452541703389,
 'secrets': 3.196452541703389,
 'screwball': 2.5943925503754266,
 'disability': 2.351354501689132,
 'homeless': 2.5943925503754266,
 'cheating': 3.196452541703389,
 'thought-provoking': 1.816241299991783,
 'Doc Ock': 3.196452541703389,
 'anti-Semitism': 2.7193312869837265,
 'jon hamm': 3.196452541703389,
 'insanity': 2.895422546039408,
 'ballet': 2.895422546039408,
 'morality': 2.895422546039408,
 'Twist Ending': 3.196452541703389,
 'widows/widowers': 2.895422546039408,
 'intellectual': 2.895422546039408,
 'Broadway': 3.196452541703389,
 'Screwball': 2.7193312869837265,
 'hitman': 2.7193312869837265,
 'sisterhood': 3.196452541703389,
 'anti-war': 3.196452541703389,
 'ironic': 2.895422546039408,
 'Savannah': 3.196452541703389,
 'tear jerker': 2.895422546039408,
 'movies about movies': 3.196452541703389,
 'Ichabod Crane': 3.196452541703389,
 'Harrison Ford': 3.196452541703389,
 'enigmatic': 2.5943925503754266,
 'race': 2.2933625547114453,
 'Emilia Clarke': 3.196452541703389,
 'earnest': 3.196452541703389,
 'Motivational': 3.196452541703389,
 'bears': 3.196452541703389,
 'zombies': 2.4183012913197452,
 'parenthood': 3.196452541703389,
 'bruce willis': 3.196452541703389,
```

```
 'PTSD': 3.196452541703389,
 'intelligent': 2.351354501689132,
 'insightful': 3.196452541703389,
 'psychiatrist': 2.895422546039408,
 'President': 3.196452541703389,
 'Russell Crowe': 3.196452541703389,
 'great performances': 3.196452541703389,
 'aliens': 2.0203612826477078,
 'overcomplicated': 3.196452541703389,
 'post apocalyptic': 3.196452541703389,
 'dialogue': 2.895422546039408,
 'kids': 3.196452541703389,
 'rich guy - poor girl': 3.196452541703389,
 'Mexico': 3.196452541703389,
 'E.M. Forster': 2.895422546039408,
 'good soundtrack': 2.7193312869837265,
 'Keanu Reeves': 2.895422546039408,
 'football': 2.7193312869837265,
 'movies': 2.5943925503754266,
 'philosopical': 3.196452541703389,
 'Kurt Russell': 3.196452541703389,
 'jackie chan': 3.196452541703389,
 'black humour': 3.196452541703389,
 'interracial romance': 3.196452541703389,
 'old': 3.196452541703389,
 'Navy': 3.196452541703389,
 'mind-bending': 3.196452541703389,
 'time travel': 1.9923325590474643,
 'Cold War': 2.7193312869837265,
 'Uma Thurman': 3.196452541703389,
 'Revenge': 3.196452541703389,
 'sweet': 2.7193312869837265,
 'ancient Rome': 3.196452541703389,
 'storytelling': 3.196452541703389,
 'birds': 2.5943925503754266,
 'cliche characters': 3.196452541703389,
 'Documentary': 3.196452541703389,
 'eerie': 2.895422546039408,
 'good writing': 3.196452541703389,
 'Pearl S Buck': 3.196452541703389,
 'Casey Affleck': 3.196452541703389,
 'espionage': 2.895422546039408,
 'abstract': 3.196452541703389,
 'lovely': 3.196452541703389,
 'twisted': 3.196452541703389,
 'philosophical': 2.196452541703389,
 'lyrical': 2.895422546039408,
 'disturbing': 2.1172712956557644,
 'In Your Eyes': 3.196452541703389,
 'video': 3.196452541703389,
 'poetic': 2.895422546039408,
 'Hearst': 3.196452541703389,
 'obsession': 2.7193312869837265,
 'religion': 1.854029860881183,
 'saint': 3.196452541703389,
 'western': 3.196452541703389,
 'Black comedy': 3.196452541703389,
 'exciting': 3.196452541703389,
 'baseball': 2.4974825373673704,
 'bad script': 2.895422546039408,
 'Made me cry': 3.196452541703389,
 'moon': 2.895422546039408,
 'bizzare': 3.196452541703389,
 'Shakespeare': 2.1172712956557644,
 'Native Americans': 3.196452541703389,
 'slick': 2.895422546039408,
 'unconventional': 2.895422546039408,
 'contemplative': 3.196452541703389,
 'imaginary friend': 3.196452541703389,
```

```
 'interesting scenario': 3.196452541703389,
 'video games': 3.196452541703389,
 'Mila Kunis': 3.196452541703389,
 'spying': 2.7193312869837265,
 'Einstein': 3.196452541703389,
 'social commentary': 2.351354501689132,
 'revenge': 2.2933625547114453,
 'special effects': 3.196452541703389,
 'orphans': 2.7193312869837265,
 'superhero': 1.816241299991783,
 'seen at the cinema': 3.196452541703389,
 'dark humor': 2.351354501689132,
 'pageant': 3.196452541703389,
 'Emma': 3.196452541703389,
 'system holism': 3.196452541703389,
 'Backwards. memory': 3.196452541703389,
 'geeky': 3.196452541703389,
 'fairy tales': 2.895422546039408,
 'golden watch': 3.196452541703389,
 'sequel': 2.2933625547114453,
 'white guilt': 2.895422546039408,
 'violent': 2.5943925503754266,
 'Mental Hospital': 3.196452541703389,
 'Kevin Costner': 3.196452541703389,
 'ridiculous': 2.895422546039408,
 'coke': 3.196452541703389,
 'Suspenseful': 3.196452541703389,
 'rug': 3.196452541703389,
 '1990s': 3.196452541703389,
 'Jason Biggs': 3.196452541703389,
 'slow paced': 3.196452541703389,
 'guns': 2.895422546039408,
 'goretastic': 3.196452541703389,
 'death penalty': 2.4974825373673704,
 'Civil War': 2.4974825373673704,
 'ogres': 3.196452541703389,
 'apocalypse': 2.895422546039408,
 'Robin Williams': 2.7193312869837265,
 'Amazing Cinematography': 3.196452541703389,
 'Humour': 3.196452541703389,
 'beautiful scenery': 2.895422546039408,
 'Favelas': 3.196452541703389,
 'harsh': 2.895422546039408,
 'falling': 3.196452541703389,
 'school': 2.895422546039408,
 'David Bowie': 3.196452541703389,
 'blood': 3.196452541703389,
 'austere': 3.196452541703389,
 'Cold': 3.196452541703389,
 'bombs': 3.196452541703389,
 'romantic': 2.895422546039408,
 'big corporations': 3.196452541703389,
 'Stephen King': 2.1172712956557644,
 'nonlinear narrative': 3.196452541703389,
 'dance': 2.7193312869837265,
 'strange': 3.196452541703389,
 'remade': 2.4183012913197452,
 'robbery': 3.196452541703389,
 'skiing': 3.196452541703389,
 "Palme d'Or": 3.196452541703389,
 'I am your father': 3.196452541703389,
 'south park': 3.196452541703389,
 'infertility': 3.196452541703389,
 'Adrien Brody': 3.196452541703389,
 'immigration': 3.196452541703389,
 'rap': 3.196452541703389,
 'dc comics': 2.895422546039408,
 'pool': 3.196452541703389,
 'disjointed timeline': 3.196452541703389,
```

```
'Hammett': 3.196452541703389,
'Watergate': 3.196452541703389,
'Sustainability': 3.196452541703389,
'introspection': 3.196452541703389,
'Up series': 3.196452541703389,
'lack of story': 3.196452541703389,
'Arnold Schwarzenegger': 2.895422546039408,
'suspense': 1.8954225460394079,
'end of the world': 3.196452541703389,
'Metaphorical': 3.196452541703389,
'thought provoking': 3.196452541703389,
'allegorical': 3.196452541703389,
'cruel characters': 3.196452541703389,
'royalty': 3.196452541703389,
'adult humor': 3.196452541703389,
'controversial': 2.5943925503754266,
'really bad': 3.196452541703389,
'Astaire and Rogers': 2.4183012913197452,
'Cerebral': 3.196452541703389,
'Harvey Keitel': 3.196452541703389,
'scifi': 3.196452541703389,
'cult film': 2.351354501689132,
'Gambling': 3.196452541703389,
'menacing': 3.196452541703389,
'Grace': 3.196452541703389,
'off-beat comedy': 2.7193312869837265,
'Tom Clancy': 2.895422546039408,
'computer': 3.196452541703389,
'reunion': 2.895422546039408,
'interwoven storylines': 3.196452541703389,
'war': 2.895422546039408,
'New York': 2.4183012913197452,
'creativity': 3.196452541703389,
'cross dressing': 2.4974825373673704,
'historical': 3.196452541703389,
'bad science': 3.196452541703389,
'nonlinear timeline': 3.196452541703389,
'sword fight': 3.196452541703389,
'survival': 2.5943925503754266,
'scary': 2.895422546039408,
'Philip Seymour Hoffman': 3.196452541703389,
'Marion Cotillard': 3.196452541703389,
'multiple personalities': 3.196452541703389,
'episodic': 3.196452541703389,
'Chris Evans': 2.895422546039408,
'Lloyd Dobbler': 3.196452541703389,
'Nick and Nora Charles': 2.4183012913197452,
'Graham Greene': 2.895422546039408,
'restaurant': 3.196452541703389,
'flood': 3.196452541703389,
'twist ending': 1.9176989407505602,
'Kevin Smith': 3.196452541703389,
'humane': 3.196452541703389,
'helena bonham carter': 2.895422546039408,
'lions': 3.196452541703389,
'killer-as-protagonist': 3.196452541703389,
'Beethoven': 3.196452541703389,
'Wesley Snipes': 3.196452541703389,
'Jane Austen': 2.5943925503754266,
'plastic surgery': 3.196452541703389,
'quotable': 3.196452541703389,
'foul language': 3.196452541703389,
'poignant': 2.4974825373673704,
'Peace Corp': 3.196452541703389,
'Boston': 3.196452541703389,
'Edith Wharton': 2.895422546039408,
'freaks': 2.895422546039408,
'black hole': 3.196452541703389,
'Lolita theme': 3.196452541703389,
```

```
'stranded': 3.196452541703389,
'parrots': 3.196452541703389,
'Jennifer Lawrence': 3.196452541703389,
'singletons': 3.196452541703389,
'Samuel L. Jackson': 2.5943925503754266,
'Horrid characterisation': 3.196452541703389,
'Klingons': 3.196452541703389,
'space opera': 2.4974825373673704,
'achronological': 3.196452541703389,
'Paris': 3.196452541703389,
'feel-good': 2.5943925503754266,
'memory loss': 3.196452541703389,
'Philip K. Dick': 3.196452541703389,
'Myth': 2.895422546039408,
'somber': 3.196452541703389,
'Oscar (Best Cinematography)': 3.196452541703389,
'cynical': 2.895422546039408,
'1970s': 2.7193312869837265,
'S.E. Hinton': 3.196452541703389,
'stupid is as stupid does': 3.196452541703389,
'motherfucker': 3.196452541703389,
'sarcasm': 2.4974825373673704,
'prejudice': 3.196452541703389,
'Ninotchka remake': 3.196452541703389,
'royal with cheese': 3.196452541703389,
'generation X': 2.5943925503754266,
'statue': 3.196452541703389,
'cattle drive': 3.196452541703389,
'Insane': 3.196452541703389,
'Will Ferrell': 2.4974825373673704,
'economics': 3.196452541703389,
'Heroic Bloodshed': 3.196452541703389,
'prodigies': 3.196452541703389,
'virginity': 2.895422546039408,
'nanny': 3.196452541703389,
'soundtrack': 2.5943925503754266,
'Denzel Washington': 3.196452541703389,
'Jason': 2.4974825373673704,
'horses': 2.5943925503754266,
'made me cry': 2.895422546039408,
'Shangri-La': 3.196452541703389,
'suicide': 3.196452541703389,
'Jennifer Connelly': 3.196452541703389,
'hugh jackman': 2.895422546039408,
'teacher': 3.196452541703389,
'shenanigans': 3.196452541703389,
'black comedy': 2.082509189396552,
'Van Gogh': 3.196452541703389,
'Kevin Spacey': 3.196452541703389,
'Rob Zombie': 2.895422546039408,
'gangs': 3.196452541703389,
'Oscar (Best Music - Original Score)': 3.196452541703389,
'freedom of expression': 3.196452541703389,
'R language': 3.196452541703389,
'small time criminals': 3.196452541703389,
'Toto': 3.196452541703389,
'women': 3.196452541703389,
'sad': 2.5943925503754266,
'psychological thriller': 3.196452541703389,
'six-fingered man': 3.196452541703389,
'beautiful': 2.4974825373673704,
'Tom Hardy': 2.895422546039408,
'parody': 2.5943925503754266,
'Notable Nudity': 3.196452541703389,
'marvel': 2.895422546039408,
'oil': 3.196452541703389,
'multiple stories': 3.196452541703389,
'cheeky': 3.196452541703389,
'Rogers and Hammerstein': 2.895422546039408,
```

```
'downbeat': 2.895422546039408,
'symbolic': 3.196452541703389,
'invisibility': 3.196452541703389,
'roald dahl': 3.196452541703389,
'ben stiller': 3.196452541703389,
'mel gibson': 3.196452541703389,
'incest': 3.196452541703389,
'surprise ending': 3.196452541703389,
'touching': 2.4183012913197452,
'painter': 3.196452541703389,
'post-college': 3.196452541703389,
'far fetched': 3.196452541703389,
'Sci-fi': 3.196452541703389,
'understated': 2.895422546039408,
'edward norton': 2.895422546039408,
'Death': 3.196452541703389,
'Music': 3.196452541703389,
'Steven Spielberg': 2.895422546039408,
'dogs': 2.5943925503754266,
'samurai': 2.4974825373673704,
'bad-ass': 3.196452541703389,
'the Force': 3.196452541703389,
'anime': 2.1172712956557644,
'space station': 3.196452541703389,
'slow action': 3.196452541703389,
'Captain America': 3.196452541703389,
'James Fennimore Cooper': 3.196452541703389,
'small towns': 3.196452541703389,
'ben affleck': 3.196452541703389,
'gunfight': 3.196452541703389,
'adolescence': 2.155059856545164,
'the catholic church is the most corrupt organization in history': 3.196452541703389,
'paranoid': 2.5943925503754266,
'purity of essence': 3.196452541703389,
'Colin Farrell': 3.196452541703389,
'mental hospital': 3.196452541703389,
'Siam': 3.196452541703389,
'illusions': 3.196452541703389,
'Sad': 3.196452541703389,
'immigrants': 2.5943925503754266,
'Union': 3.196452541703389,
'sniper': 2.895422546039408,
'Ryan Reynolds': 2.5943925503754266,
'muppets': 3.196452541703389,
'Thrilling': 3.196452541703389,
'class': 2.7193312869837265,
'superman': 3.196452541703389,
'heist': 2.082509189396552,
'singers': 3.196452541703389,
'Bugs Bunny': 3.196452541703389,
'free to download': 3.196452541703389,
'tom hardy': 3.196452541703389,
'Thanksgiving': 3.196452541703389,
'passion': 3.196452541703389,
'great soundtrack': 2.4183012913197452,
'twists & turns': 2.895422546039408,
'love': 2.895422546039408,
'not funny': 3.196452541703389,
'Marvel': 2.895422546039408,
'bank': 3.196452541703389,
'John Malkovich': 3.196452541703389,
'Trey Parker': 3.196452541703389,
'satire': 2.1172712956557644,
'different': 3.196452541703389,
'independent': 3.196452541703389,
'Saturday Night Live': 3.196452541703389,
'Pixar': 2.5943925503754266,
'Oscar (Best Effects - Visual Effects)': 3.196452541703389,
'French': 3.196452541703389,
```

```
 'road trip': 3.196452541703389,
 'comedy': 2.0203612826477078,
 'accident': 3.196452541703389,
 'based on a play': 3.196452541703389,
 'Star Wars': 2.895422546039408,
 'theater': 2.895422546039408,
 'M. Night Shyamalan': 2.895422546039408,
 'Seth Rogen': 2.5943925503754266,
 'inspirational': 2.351354501689132,
 'comic book': 2.155059856545164,
 'Matrix': 3.196452541703389,
 'books': 2.7193312869837265,
 'court': 2.1172712956557644,
 'Dr. Strange': 3.196452541703389,
 'southern US': 3.196452541703389,
 'fantasy': 2.4183012913197452,
 'aggressive': 3.196452541703389,
 'Coen Bros': 3.196452541703389,
 'mediacentralism': 3.196452541703389,
 'Something for everyone in this one... saw it without and plan on seeing it with kid
s!': 3.196452541703389,
 'dark comedy': 1.8742332469694698,
 'Comedy': 2.5943925503754266,
 'r:violence': 3.196452541703389,
 'Alcatraz': 3.196452541703389,
 'bible': 3.196452541703389,
 'Gulf War': 3.196452541703389,
 'narrated': 2.895422546039408,
 'film history': 3.196452541703389,
 'Hemingway': 2.895422546039408,
 'Not Seen': 3.196452541703389,
 'Amy Adams': 3.196452541703389,
 'friendship': 2.4974825373673704,
 'Day and Hudson': 3.196452541703389,
 'vampire': 2.895422546039408,
 'Agatha Christie': 2.895422546039408,
 'food': 2.7193312869837265,
 'long takes': 3.196452541703389,
 'philosophy': 2.4183012913197452,
 'Wizards': 2.5943925503754266,
 'Visually stunning': 2.895422546039408,
 'Halloween': 3.196452541703389,
 'stop looking at me swan': 3.196452541703389,
 'witty': 2.4183012913197452,
 'Mafia': 2.196452541703389,
 'film-noir': 3.196452541703389,
 'teachers': 3.196452541703389,
 'werewolf': 3.196452541703389,
 'space craft': 3.196452541703389,
 'dystopia': 2.4974825373673704,
 'arthouse': 3.196452541703389,
 'Natalie Portman': 3.196452541703389,
 'DC': 3.196452541703389,
 'art house': 3.196452541703389,
 'luke skywalker': 2.895422546039408,
 'sex': 3.196452541703389,
 'blind': 2.7193312869837265,
 'pop culture references': 3.196452541703389,
 'California': 3.196452541703389,
 'Housekeeper': 3.196452541703389,
 'entirely dialogue': 3.196452541703389,
 'racism': 2.2422100322640643,
 'Moses': 3.196452541703389,
 'Howard Hughes': 3.196452541703389,
 'Depressing': 3.196452541703389,
 'Seth MacFarlane': 3.196452541703389,
 'mecha': 3.196452541703389,
 'killer': 3.196452541703389,
 'Dodie Smith': 3.196452541703389,
```

```
 'Academy award (Best Supporting Actress)': 3.196452541703389,
 'meaningless violence': 3.196452541703389,
 'Emma Stone': 3.196452541703389,
 'Creature Feature': 3.196452541703389,
 'spacecraft': 3.196452541703389,
 '2001-like': 3.196452541703389,
 'fairy tale': 3.196452541703389,
 'subway': 3.196452541703389,
 'China': 3.196452541703389,
 'FBI': 2.895422546039408,
 'Peter Pan': 2.895422546039408,
 'magic board game': 3.196452541703389,
 'Ray Bradbury': 2.895422546039408,
 'fast-paced dialogue': 3.196452541703389,
 'Jim Morrison': 3.196452541703389,
 ...}
```

In [65]:

```
# create tag representation
tag_representation = pd.DataFrame(columns=sorted(unique_tags), index=list(set(tags_df

for name, group in tqdm(tags_df.groupby(by='movieId')):
    temp_list = list(map(lambda x:x.split(','), list(group['tag'])))
    temp_tag_list = list(set(list(map(lambda x:x.strip(), list([tag for sublist in t

    dict_temp = {i: tag_idf[i.strip()] for i in temp_tag_list}
    row_to_add = pd.DataFrame(dict_temp, index=[group['movieId'].values[0]])
    tag_representation.update(row_to_add)

tag_representation = tag_representation.sort_index(0)
tag_representation
```

```
100%|██████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████| 1572/1572 [02:55<00:00,  8.96it/s]
```

Out[65]:

	"artsy"	06 Oscar Nominated Best Movie - Animation	1900s	1920s	1950s	1960s	1970s	1980s	1990s	2001-like	...	women
1	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN
2	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN
3	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN
5	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN
7	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN
...	...	...	...	...	...	...	...	...	...	...	...	...
183611	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN
184471	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN
187593	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN
187595	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN
193565	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	NaN	...	NaN

1572 rows × 1589 columns

In [ ]:

```
# tag representation 확인
```

In [66]:

```
movies_df.head()
```

Out[66]:

	title	genres
movieId		
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

In [67]:

```
tag_representation.loc[1].dropna()
```

Out[67]:

```
fun      2.497483
pixar    2.895423
Name: 1, dtype: object
```

In [68]:

```
tag_representation.loc[2].dropna()
```

Out[68]:

```
Robin Williams      2.719331
fantasy             2.418301
game                3.196453
magic board game    3.196453
Name: 2, dtype: object
```

Fianl Movie Representation 생성

In [70]:

```
movie_representation = pd.concat([genre_representation, tag_representation], axis=1).
print(movie_representation.shape)
movie_representation.describe()
```

```
(9742, 1609)
```

Out[70]:

	(no genres listed)	Action	Adventure	Animation	Children	Comedy	Crime
count	9742.000000	9742.000000	9742.000000	9742.000000	9742.000000	9742.000000	9742.000000
mean	0.008576	0.136354	0.115027	0.075425	0.079506	0.159587	0.111978
std	0.144915	0.283726	0.298052	0.291593	0.293989	0.201476	0.298916
min	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
25%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
50%	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
75%	0.000000	0.000000	0.000000	0.000000	0.000000	0.413923	0.000000
max	2.457169	0.726672	0.887245	1.202607	1.166480	0.413923	0.909829

8 rows × 1609 columns

콘텐츠 유사도 평가 (cosine similarity)

In [71]:
```
from sklearn.metrics.pairwise import cosine_similarity

def cos_sim_matrix(a, b):
    cos_sim = cosine_similarity(a, b)
    result_df = pd.DataFrame(data=cos_sim, index=[a.index])
    return result_df
```

In [72]:
```
cs_df = cos_sim_matrix(movie_representation, movie_representation)
cs_df.head()
```

Out[72]:

	0	1	2	3	4	5	6	7	8	9	...	9
1	1.000000	0.124438	0.008403	0.040571	0.011755	0.0	0.016339	0.331122	0.0	0.131794	...	0.064
2	0.124438	1.000000	0.000000	0.000000	0.000000	0.0	0.000000	0.240843	0.0	0.095861	...	0.000
3	0.008403	0.000000	1.000000	0.179391	0.011294	0.0	0.072246	0.000000	0.0	0.000000	...	0.006
4	0.040571	0.000000	0.179391	1.000000	0.054530	0.0	0.348828	0.000000	0.0	0.000000	...	0.031
5	0.011755	0.000000	0.011294	0.054530	1.000000	0.0	0.640342	0.000000	0.0	0.000000	...	0.009

5 rows × 9742 columns

In [81]:
```
cs_df[1].sort_values(ascending=False)
```

Out[81]:
```
2         1.000000
46972     0.322201
158813    0.300850
119655    0.300850
80748     0.300850
            ...
4921      0.000000
4920      0.000000
4919      0.000000
4917      0.000000
193609    0.000000
Name: 1, Length: 9742, dtype: float64
```

유사도 평가 결과

In [85]:
```
print(movies_df.loc[2])
print(movies_df.loc[46972])
print(movies_df.loc[158813])
print(movies_df.loc[80748])
```

```
title                 Jumanji (1995)
genres    Adventure|Children|Fantasy
Name: 2, dtype: object
title     Night at the Museum (2006)
genres    Action|Comedy|Fantasy|IMAX
Name: 46972, dtype: object
title     Alice Through the Looking Glass (2016)
genres                Adventure|Children|Fantasy
Name: 158813, dtype: object
title     Alice in Wonderland (1933)
genres    Adventure|Children|Fantasy
Name: 80748, dtype: object
```

성능평가

In [86]:

```
train_df, test_df = train_test_split(ratings_df, test_size=0.2, random_state=1234)
print(train_df.shape)
print(test_df.shape)
```

```
(80668, 4)
(20168, 4)
```

In [88]:

```
test_userids = list(set(test_df.userId.values))
```

In [93]:

```
result_df = pd.DataFrame()

for user_id in tqdm(test_userids):
    user_record_df = train_df.loc[train_df.userId == int(user_id), :]

    user_sim_df = cs_df.loc[user_record_df['movieId']] # (n, 9742) 차원 : n은 유저가 
    user_ratings_df = user_record_df[['rating']] # (n, 1) 차원
    sim_sum = np.sum(user_sim_df.T.to_numpy(), -1) # (9742, 1) # 유저가 매긴 영화유사

    prediction = np.matmul(user_sim_df.T.to_numpy(), user_ratings_df.to_numpy()).flat

    prediction_df = pd.DataFrame(prediction, index=cs_df.index).reset_index()
    prediction_df.columns = ['movieId', 'pred_rating']
    prediction_df = prediction_df[['movieId', 'pred_rating']][prediction_df.movieId.i

    temp_df = prediction_df.merge(test_df[test_df.userId == user_id], on='movieId')
    result_df = pd.concat([result_df, temp_df], axis=0)
```

```
100%|████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████████| 610/610 [00:06<00:00, 92.27it/s]
```

In [96]:

```
result_df.head(10)
```

Out[96]:

	movieId	pred_rating	userId	rating	timestamp
0	1	4.145652	1	4.0	964982703
1	50	3.650755	1	5.0	964982931
2	216	2.670124	1	5.0	964981208
3	223	2.612844	1	3.0	964980985
4	231	4.215284	1	5.0	964981179
5	235	3.619820	1	4.0	964980908
6	316	4.136756	1	3.0	964982310
7	457	3.218743	1	5.0	964981909
8	543	3.729524	1	4.0	964981179
9	592	4.024728	1	4.0	964982271

In [98]:

```
mse = mean_squared_error(y_true=result_df['rating'].values, y_pred=result_df['pred_ra
rmse = np.sqrt(mse)

print(mse, rmse)
```

```
1.40606646706041 1.1857767357561078
```

In [ ]: