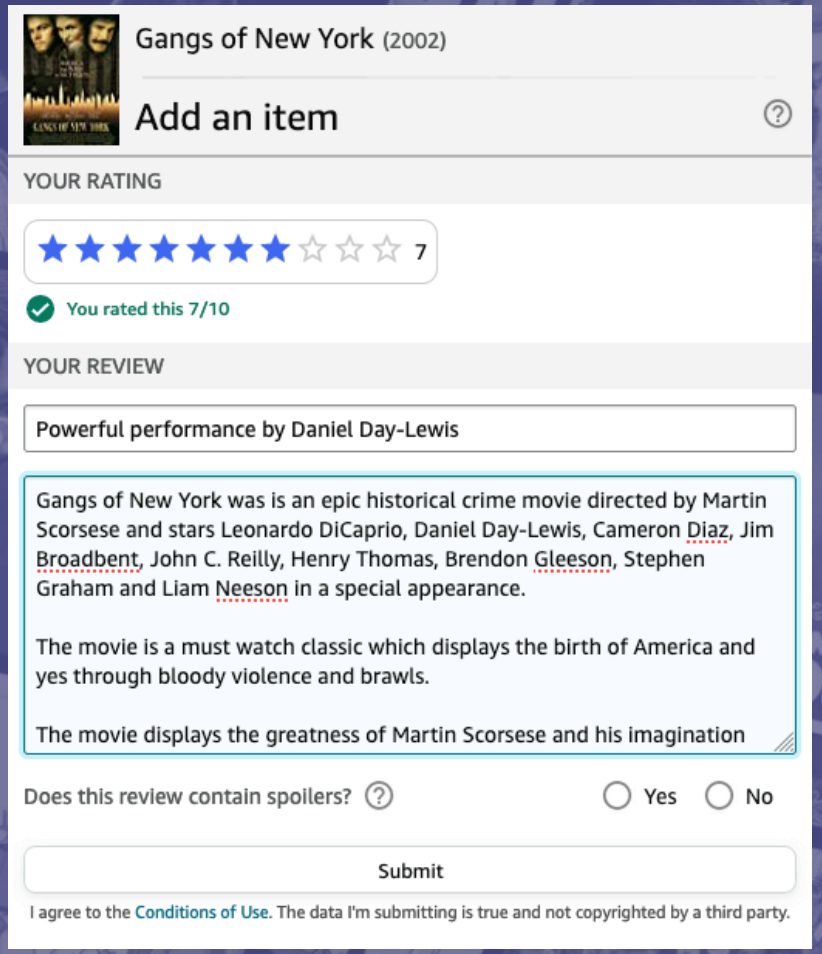
SENTIMENT ANALYSIS ON MOVIE REVIEWS WITH NLP AND CREATING A RECOMMENDATION SYSTEM

AMIN KHOEINI

Data Science Career Track Capstone Project September 2021 Cohort

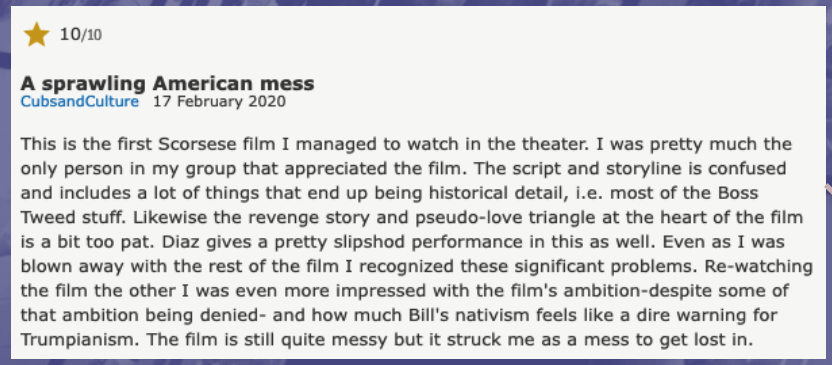
PROBLEM:

- Streaming site tries to make user to engage more with platform.
- Recommending movie, more movie watch more time spends on the platform.
- Netflix uses thumbs up and double thumbs up for this feature and then creates a list specific to the user's liking to watch.
- IMDB as biggest movie database can imply same same recommendation system.
- This means that people have more clicks and spend more time on the website which further suggests more user engagement and ad revenues for IMDB.



GOAL AND UTILITY:

- Make a NLP model predict the sentiment of the user review. User Love the movie or not interested in it.
- In the case that the user love the movie, create a recommendation list for user to watch and probably write more review on the website.

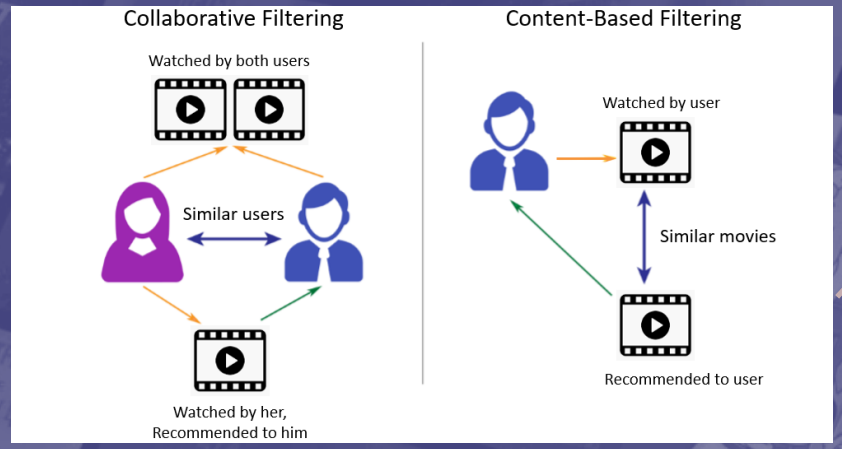


10/10

A sprawling American mess

CubsandCulture 17 February 2020

This is the first Scorsese film I managed to watch in the theater. I was pretty much the only person in my group that appreciated the film. The script and storyline is confused and includes a lot of things that end up being historical detail, i.e. most of the Boss Tweed stuff. Likewise the revenge story and pseudo-love triangle at the heart of the film is a bit too pat. Diaz gives a pretty slipshod performance in this as well. Even as I was blown away with the rest of the film I recognized these significant problems. Re-watching the film the other I was even more impressed with the film's ambition-despite some of that ambition being denied- and how much Bill's nativism feels like a dire warning for Trumpianism. The film is still quite messy but it struck me as a mess to get lost in.



DATA WRANGLING:

Reviews

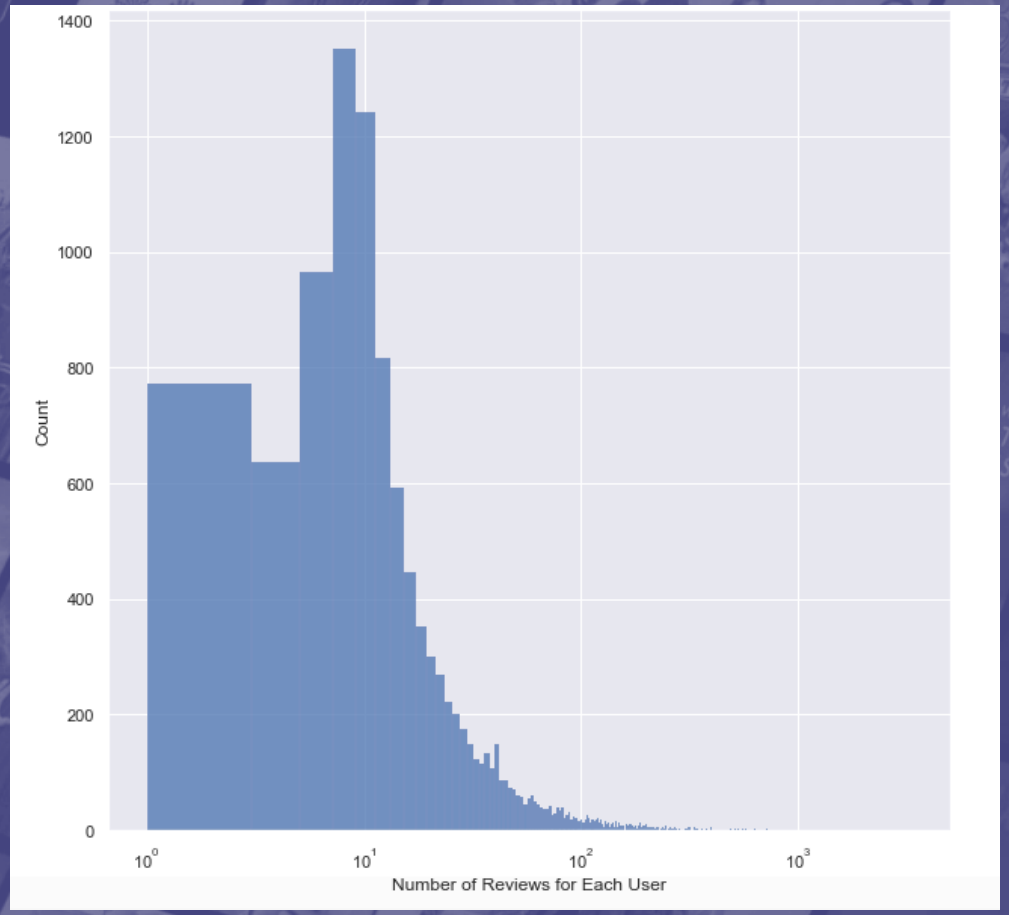
- User ID : created from reviewer user name
- Movie ID : extracted form the IMDB link
- Review: processed, Lemmatize and Stem
- Label: created from the rating
  - Love it and Not Love It

Movie Metadata

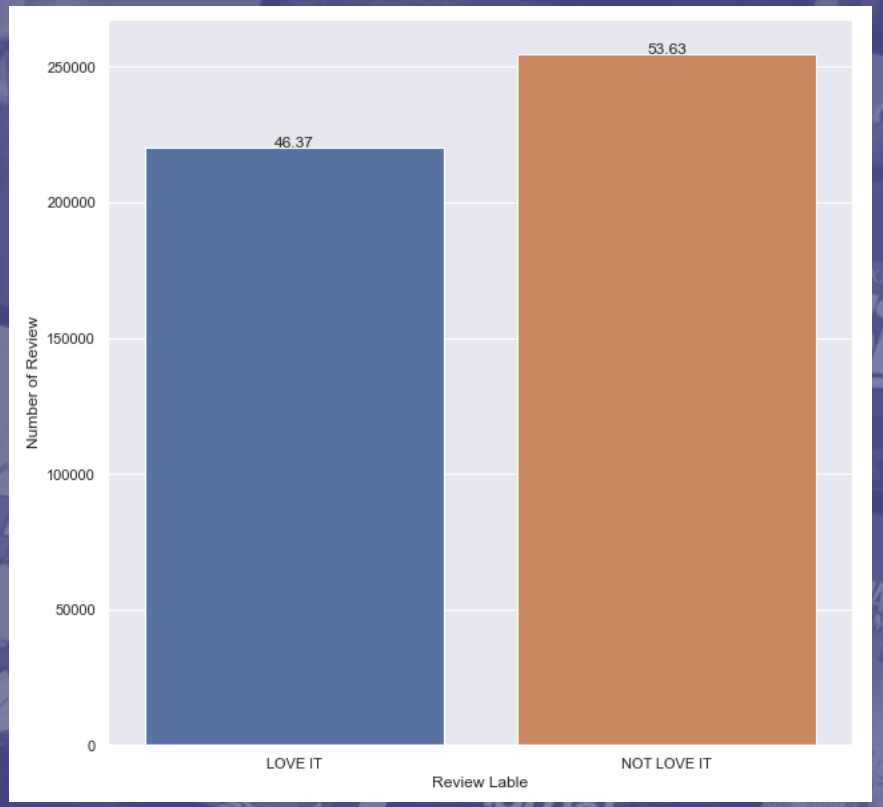
- Movie ID
- Movie Title
- Year
- Director
- Genre
- Actors
- Description

DATA EXPLORATION: LABEL

Most of the user wrote less than 10 review on the website

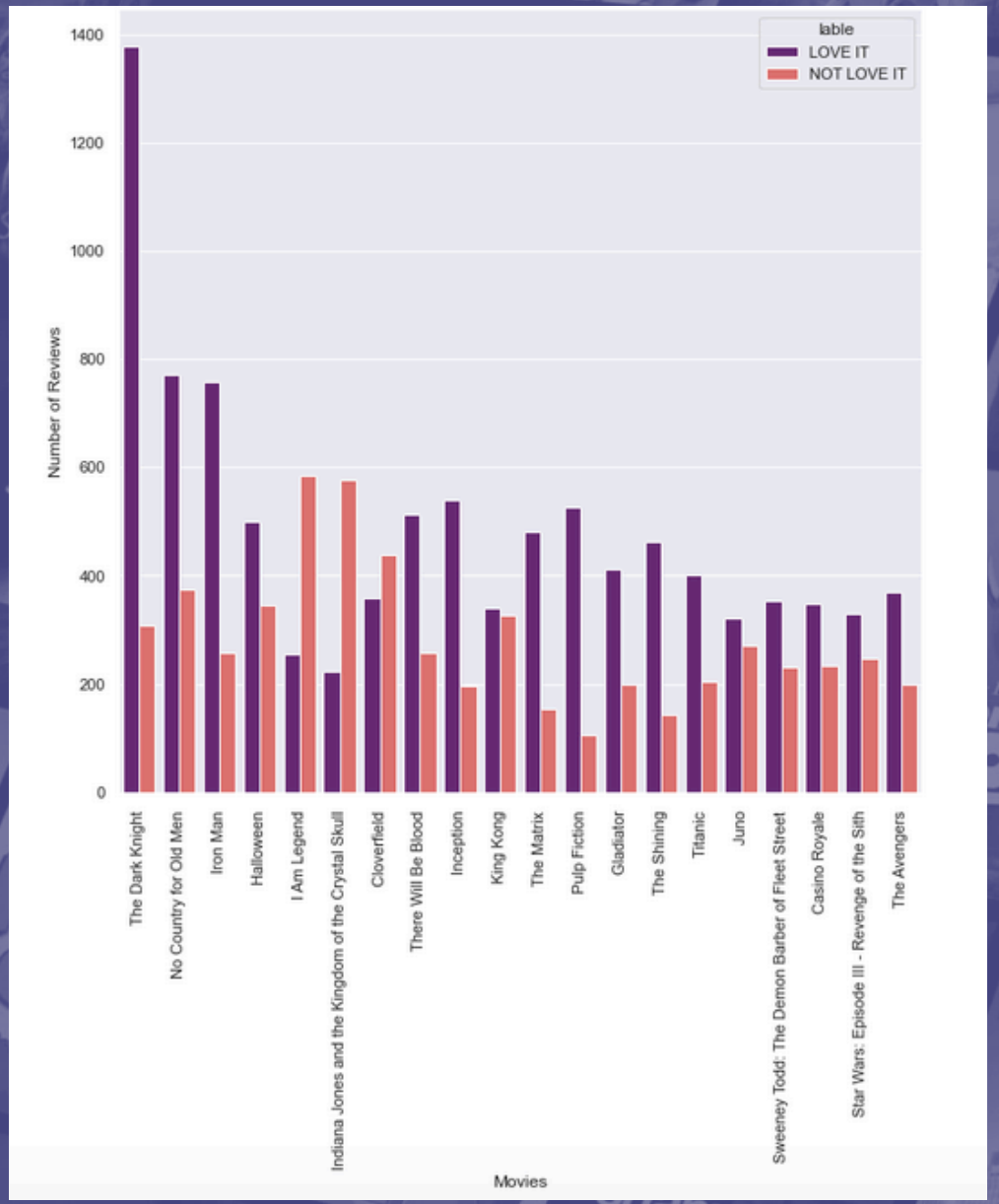


Good balanced distribution of label



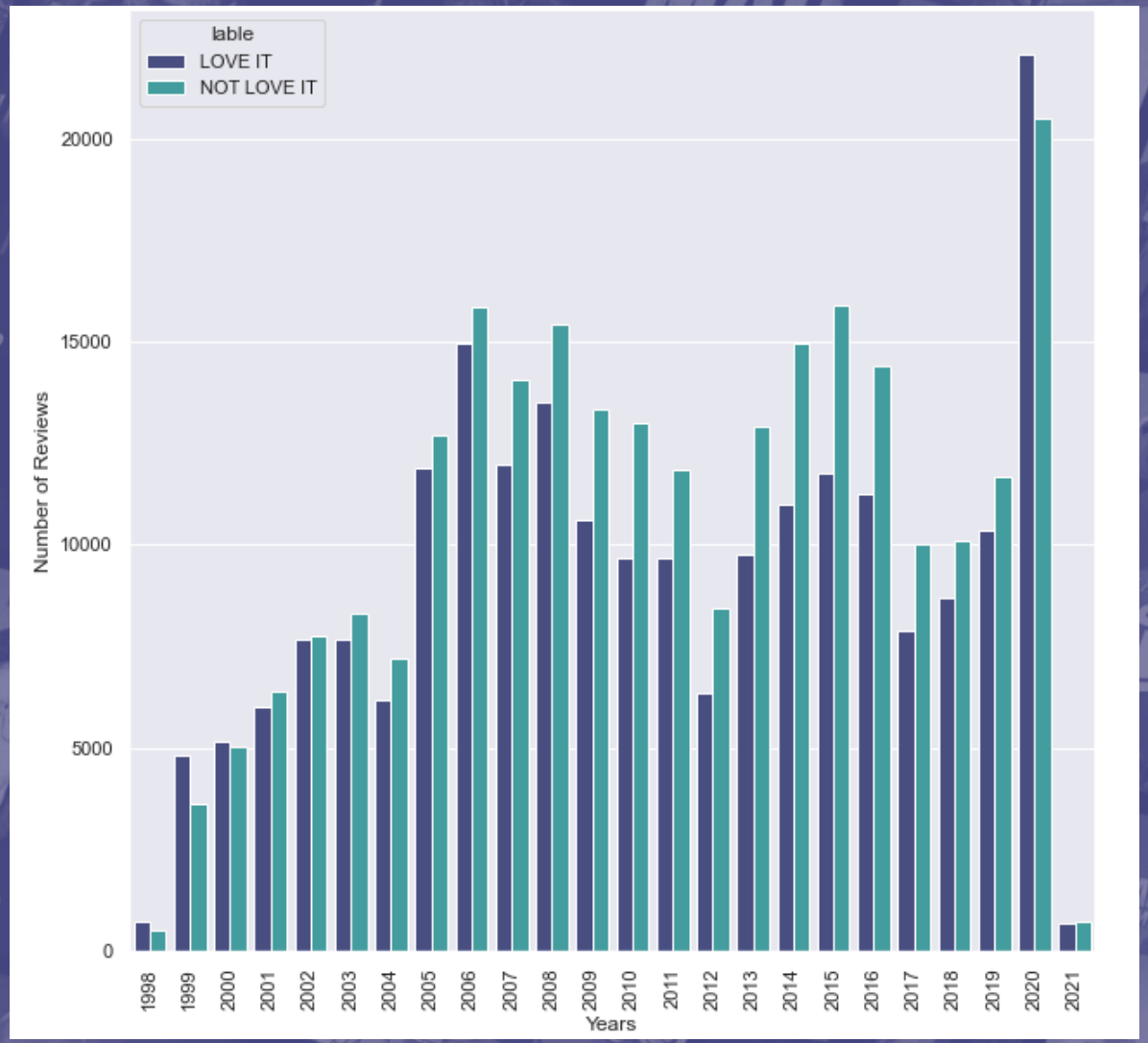
DATA EXPLORATION: MOVIES

- Popular movies have considerably more positive reviews
- there are a large number of negative reviews for I am Legend, Indiana Jones and the Kingdom of the Crystal Skull.
- Indiana Jones was the reboot of the famous series which apparently viewers did not like.



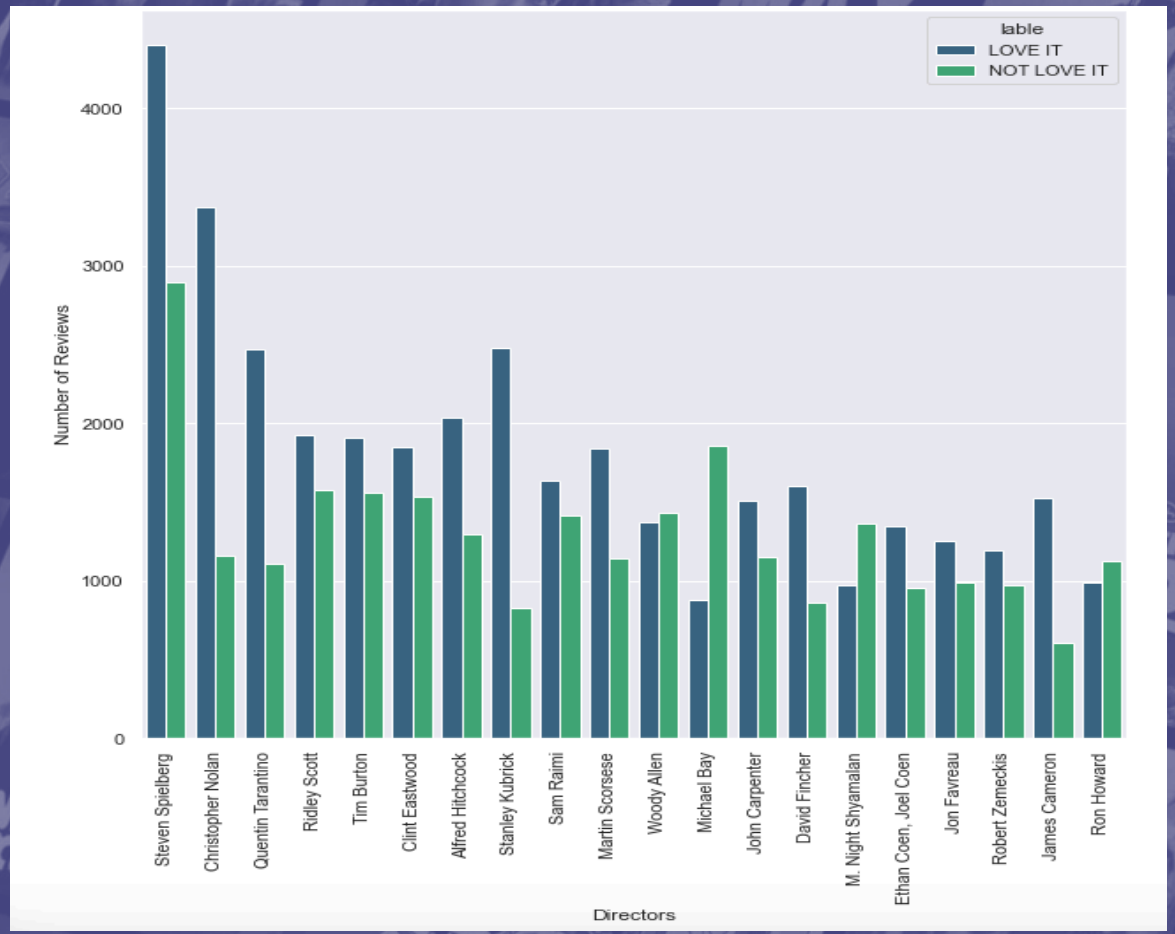
DATA EXPLORATION: YEARS

- Jump in reviews in 2020
- Also positive reviews are more than negative for first time



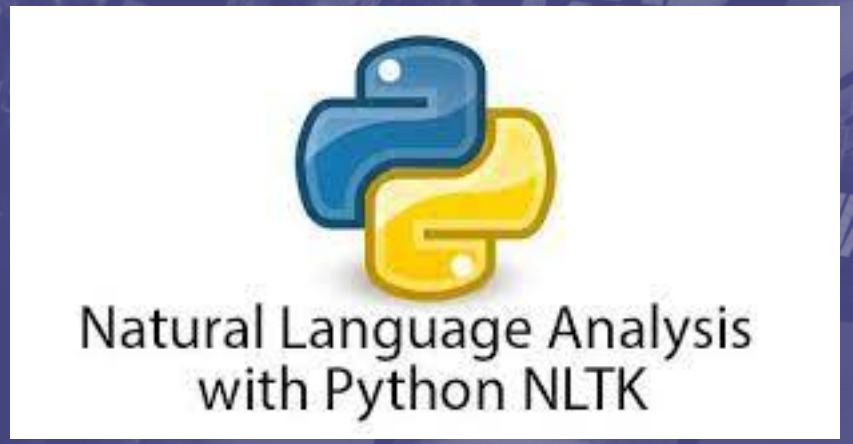
DATA EXPLORATION: DIRECTORS

- Famous directors have the most reviews about them.
- Michael Bay and M. Night Shyamalan are notorious for making a bad blockbuster.
- Woody Allen also has more negative reviews



REVIEW SENTIMENT ANALYSIS:

- NLTK library
- Tokenize the review corpus
- Remove the stop word and punctuation
- Lemmatized using WordNetLemmatizer
- Stem using SnowballStemmer

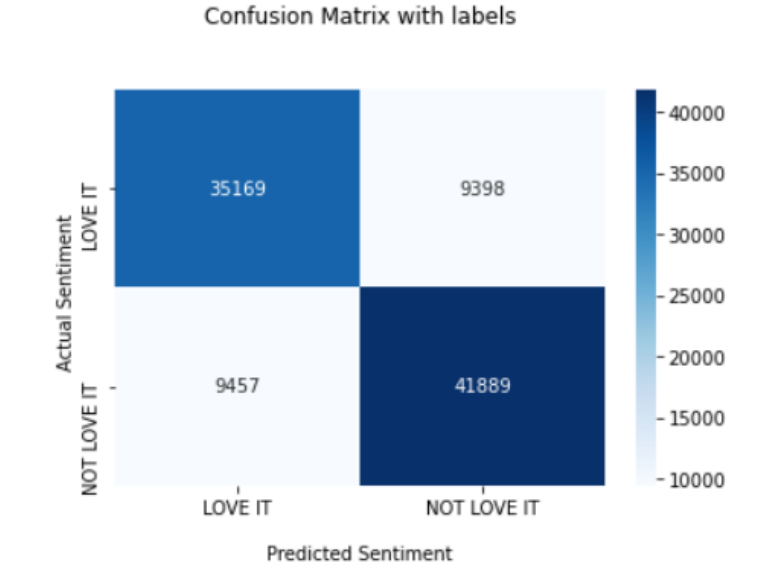


- Sklearn Library
- TFIDF vectorizer
- MultinomialNB, LogisticRegression and Linear SVC

REVIEW SENTIMENT ANALYSIS:

- Liner SVC with hyper parameters (C = 1, loss = hinge, penalty = l2)
- Best performance with 81% accuracy

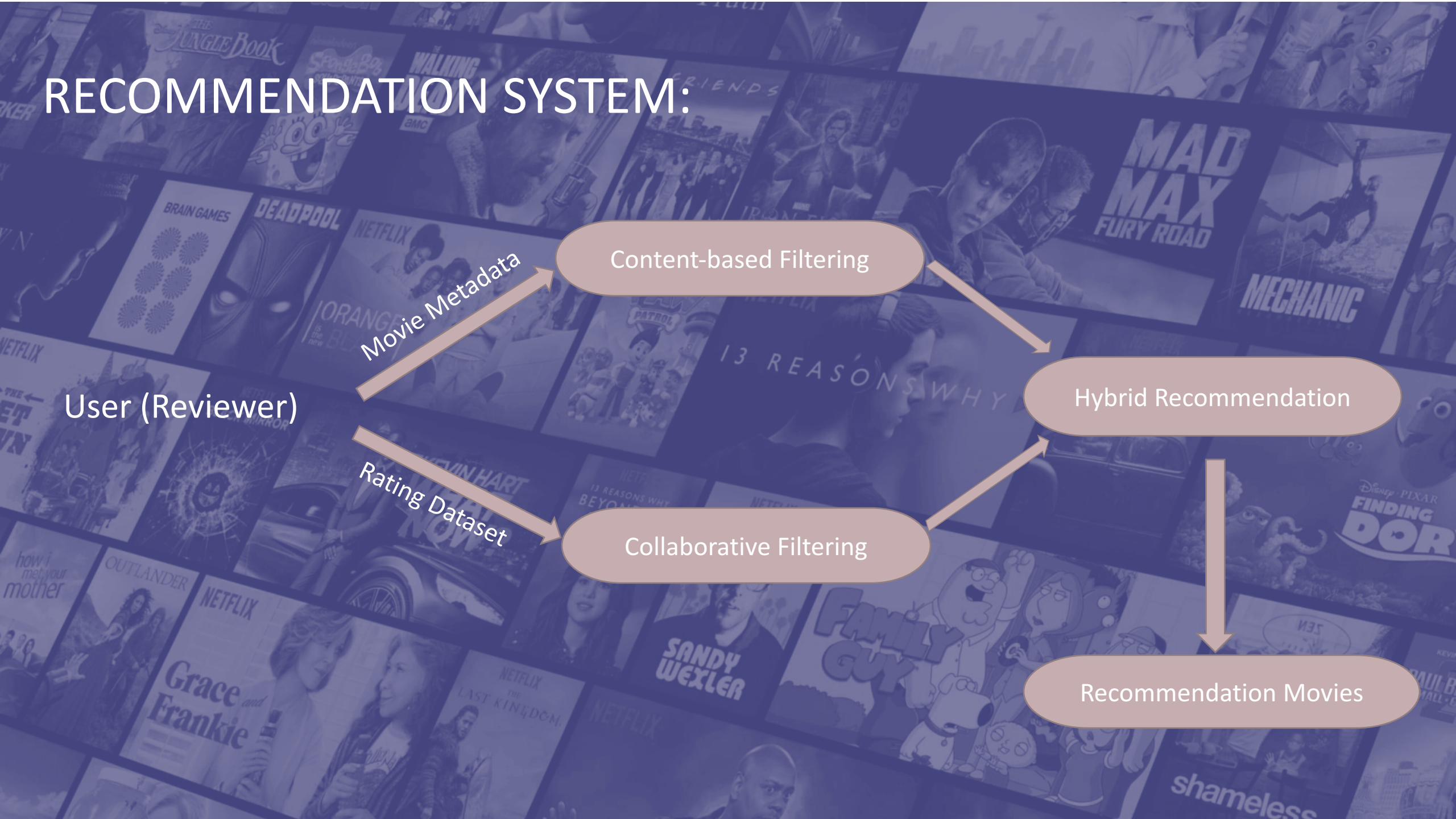
	precision	recall	f1-score	support
LOVE IT	0.79	0.79	0.79	44567
NOT LOVE IT	0.82	0.82	0.82	51346
accuracy			0.80	95913
macro avg	0.80	0.80	0.80	95913
weighted avg	0.80	0.80	0.80	95913




```
review = 'if you want to see a entertaining movie avoid this film by all means'
review = punc_clean(review)
review = lemmatize_text(review)
review = remove_stopword(review)
review = tfidf_vectorizer.transform([review])
lsvc_tfidf.predict(review)
```

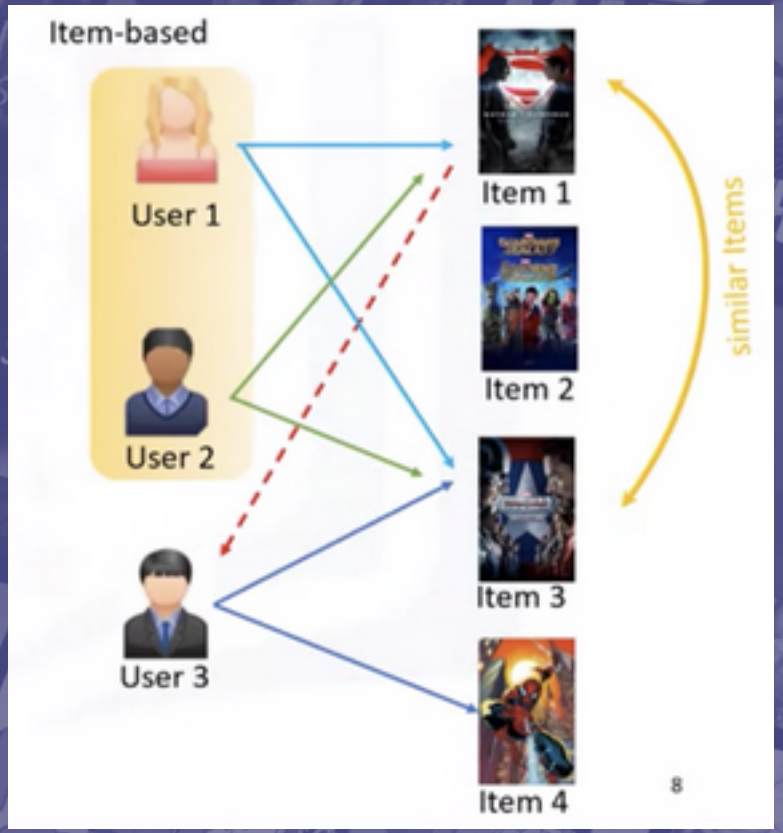
```
array(['NOT LOVE IT'], dtype=object)
```

RECOMMENDATION SYSTEM:



CONTENT-BASE FILTER:

- Using movie properties to make a item based filter
- Create a mix of all the movie data columns
- Called it soup
- Do the Tfidf vectorization
- Use cousin similarity to make a matrix
- Create a list of 30 movies closest to the target.



CONTENT-BASE FILTER:

Gangs of New York
- 2002
- Crime/Drama
- Martin Scorsese
- Leonardo DiCaprio

```
['The Departed (2006)',
 'The Aviator (2004)',
 'The Wolf of Wall Street (2013)',
 'Mean Streets (1973)',
 'The Last Temptation of Christ (1988)',
 'Shutter Island (2010)',
 'After Hours (1985)',
 'Casino (1995)',
 'New York, New York (1977)',
 "Alice Doesn't Live Here Anymore (1974)",
 'Bringing Out the Dead (1999)',
 'Hugo (2011)',
 'Midnight Cowboy (1969)',
 'The Bounty (1984)',
 'The Tailor of Panama (2001)',
 'Dead Man Down (2013)',
 'Catch Me If You Can (2002)',
 'There Will Be Blood (2007)',
 'Public Enemies (2009)',
 'Key Largo (1948)',
 'American Psycho (2000)',
 'London Boulevard (2010)',
 'Kalifornia (1993)',
 'Extremely Loud & Incredibly Close (2011)',
 'The Taking of Pelham 123 (2009)',
 'The Disappearance of Alice Creed (2009)',
 'A Most Violent Year (2014)',
 'Casualties of War (1989)',
 'Calvary (2014)',
 'The Shadow (1994)']
```

COLLABORATIVE FILTERING:

- Using the Surprise library
- Rating Data (User ID, Movie ID, Rating)
- KNNBaseline algorithms has the least RMSE

surprise

A Python scikit for recommender systems.

	test_rmse	fit_time	test_time	Algorithm
0	1.854455	38.425230	2.319381	SVD
1	1.912588	2272.251957	56.917552	SVDpp
2	1.843742	9.421812	46.156351	SlopeOne
3	2.274557	40.471097	2.107328	NMF
4	3.132171	0.577075	2.039930	NormalPredictor
5	1.804845	60.256454	55.050260	KNNBaseline
6	1.935067	69.662538	58.847921	KNNBasic
7	1.865102	64.828813	49.204878	KNNWithMeans
8	1.867365	60.622549	51.948352	KNNWithZScore
9	1.841107	0.421204	1.647150	BaselineOnly

COLLABORATIVE FILTERING:

- Getting the top 20 movie based on the rating for user ID of 2.

The Other (1972)	robert mulligan	1972	drama horror mystery	6.908325
La Grande Illusion (1937)	jean renoir	1937	drama war	6.661664
A Christmas Carol (1951)	brian desmond hurst	1951	drama fantasy	6.655398
It's a Wonderful Life (1946)	frank capra	1946	drama family fantasy	6.633878
The Wizard of Oz (1939)	victor fleming george cukor	1939	adventure family fantasy	6.629090
Glory (1989)	edward zwick	1989	biography drama history	6.598417
Dead Man Walking (1995)	tim robbins	1995	crime drama	6.567237
The Right Stuff (1983)	philip kaufman	1983	adventure biography drama	6.556452
The Spy Who Came in from the Cold (1965)	martin ritt	1965	drama thriller	6.553010
Alive (1993)	frank marshall	1993	adventure biography drama	6.513601
A Streetcar Named Desire (1951)	elia kazan	1951	drama	6.472359
The Long Goodbye (1973)	robert altman	1973	comedy crime drama	6.438745
The Heiress (1949)	william wyler	1949	drama romance	6.437617
The Apostle (1997)	robert duvall	1997	drama	6.432570
Perfect Blue (1997)	satoshi kon	1997	animation crime mystery	6.411650
The Good, the Bad and the Ugly (1966)	sergio leone	1966	western	6.394269
Rare Exports (2010)	jalmari helander	2010	adventure fantasy horror	6.384164
Jaws (1975)	steven spielberg	1975	adventure thriller	6.378790
12 Angry Men (1957)	sidney lumet	1957	crime drama	6.372233
Beauty and the Beast (1946)	jean cocteau rené clément	1946	drama fantasy romance	6.325395

HYBRID RECOMMENDATION:

- First Used the content-base filter to get the 30 movie that closest to the target movie.
- Use the collaborative filter to predict the rating for this list for the target user
- Sort the list based on the predicted rating
- Print the top 10 as the final recommendation

Title	director	year	genre	est
Key Largo (1948)	john huston	1948	action crime drama	5.757901
Midnight Cowboy (1969)	john schlesinger	1969	drama	5.647158
Calvary (2014)	john michael mcdonagh	2014	comedy drama	5.350856
Alice Doesn't Live Here Anymore (1974)	martin scorsese	1974	drama romance	5.337787
The Last Temptation of Christ (1988)	martin scorsese	1988	drama	5.066978
Casino (1995)	martin scorsese	1995	crime drama	5.003399
Casualties of War (1989)	brian de palma	1989	crime drama war	4.890082
The Aviator (2004)	martin scorsese	2004	biography drama	4.810042
After Hours (1985)	martin scorsese	1985	comedy crime drama	4.805793
Catch Me If You Can (2002)	steven spielberg	2002	biography crime drama	4.719803

MODELING:

- The final stage is to combine the NLP model with the Hybrid recommendation.
- Input:
  - User ID
  - Movie Title
  - Review
- Output:
  - In case of negative review, no recommendation
  - In case of positive review print the recommendation list

```
#5/10 review.
user_id= 9877
movie_title= 'Superman IV: The Quest for Peace (1987)'
review = "Superman IV is not nearly as bad as the reviews suggest. The actors try really hard, \
particularly Christopher Reeve, Gene Hackman, and Margot Kidder, to make it work. The movie is watchable \
and the musical score is good. The movie is an improvement over the disappointing Superman III. \
However, Superman IV has major problems. The movie has obviously been cut from its original length \
make it incoherent at times. The special effects are below the standards set in the first two movies \
(even the third movie had decent effects). Maybe if the movie were restored to its original length, \
it would be better. I can only give this movie a 5/10. I wished it were better and hope someday they \
do restore this movie to its original length."


sentiment_recommender(user_id,movie_title,review)
```

```
The user did not love Superman IV: The Quest for Peace (1987) therefor the system has no recommandation based on th
at review.
```

```
#7/10 review.
user_id= 5675
movie_title= 'Superman IV: The Quest for Peace (1987)'
review = "This is the fourth and final Superman film with Christopher Reeve taking on the role, \
where he tries to stop the spread of nuclear weapons and battles Lex Luther (Gene Hackman) and his \
super-powered sidekick, Nuclear Man (Mark Pillow).Though the level of excitement and intrigue of this \
film doesn't match the first two, it is almost on par with the third and is still fun and fast-paced \
with neat special effects that showcase Superman's powers and great action scenes, something that sorely \
lacks in Superman Returns.The nuclear weapons plot I thought was clever and unique to a Superman movie. \
Total disarmament of nuclear weapons may not be feasible in the real world, but Superman's quest for \
peace among all countries is well-intended in the film. Reeve is great as Superman and Gene Hackman gave \
us another thrilling performance as super villain Lex Luther. In addition, the fight scenes between \
Nuclear Man and Superman were awesome - Nuclear Man looked like a force to be reckon with as he possessed \
the same level of powers and same physique as Superman.Overall, I think this film is a fitting ending to \
the Christopher Reeve saga."

sentiment_recommender(user_id,movie_title,review)
```

```
Based on the review, This user loved Superman IV: The Quest for Peace (1987) therefor the system recommand the belo
w list for the user to watch:

 ['Raiders of the Lost Ark (1981)', '2001: A Space Odyssey (1968)', 'Predator (1987)', 'Iron Man (2008)', 'Spider-M
an (2002)', 'Captain America: Civil War (2016)', 'Dead Man (1995)', 'The Ipcress File (1965)', 'Star Trek VI: The U
ndiscovered Country (1991)', 'Hot Shots! (1991)']
```

FUTURE IMPROVEMENTS:

- In NLP model and sentiment analysis, it is optimal to have an accuracy above 90%
- Having more reviews, better hardware to run hyper parameters tuning for even more different algorithms, would help to improve the NLP performance.

- There is only data about 4000 movies in the dataset. Because of that the recommendation engine was limited to those only.
- Having more score and user data created more reliable recommendations for the user.