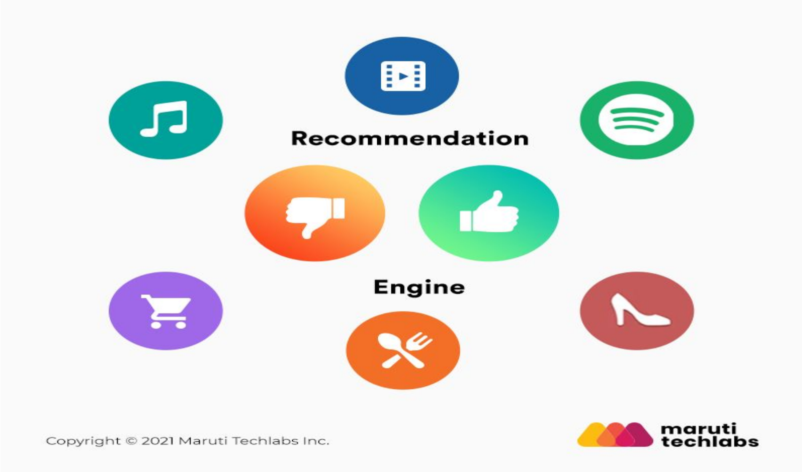
Capstone Project
Movie Recommendation System



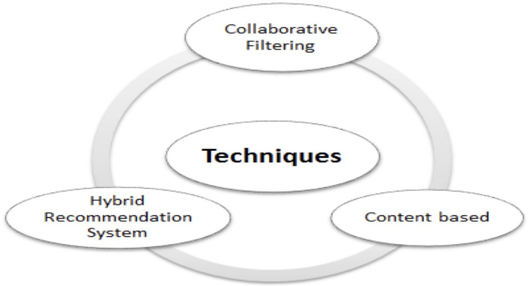
Basazin Belhu
11/23/2021

Outline

- Introduction
- Problem Statements
- Methods
- Datasets
- Data Exploration and Visualization
- Model Development and Evaluations/Tests
- Conclusion and Future Work

Introduction

- We like things that are closely similar or we may have some common things with people. This habits can be useful for recommending new products.
- Recommendation is web based algorithm that recommends products for the customers based on their past interaction.
- Recommendation increases sales and customers satisfaction
- Three techniques of Recommendation System:
  1. Content Based
  2. Collaborative Based
  3. Hybride: Not Covered



https://iopscience.iop.org/article/10.1088/1742-6596/1717/1/012002/pdf

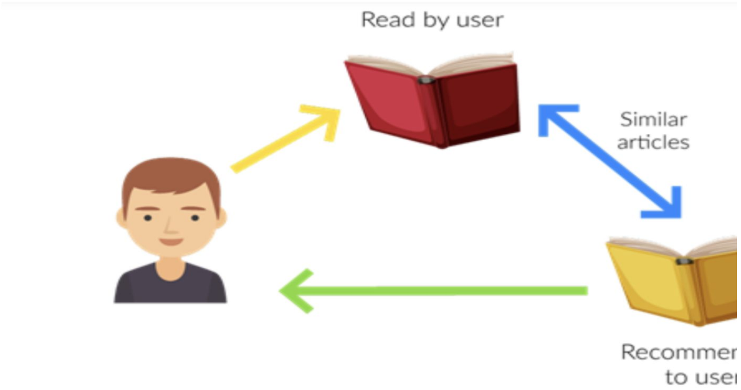
Problem Statements

- On the internet and cloud world, we have a lot of preferences on the website/Netflix account/ to watch, but it is challenging to select the best one for ourselves before we watch the whole movie. This is one of the challenges for having too many options for the customers.
- The data scientist team has all customers' past information and wants to create the best recommendation for the customers. The new entertainment industry has just hired me. They want to create an algorithm that recommends different movies for customers to increases their satisfaction with our startup services.

Methods

Content Based Filtering

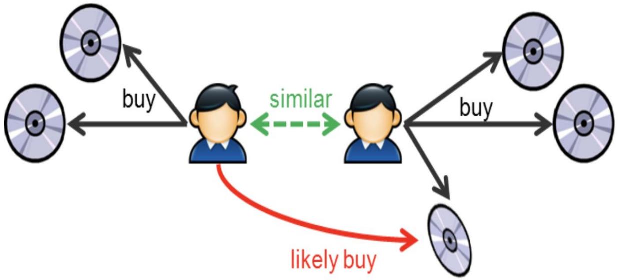
- Based on the product similarity
- Based on the description of the item



Source: https://recosenselabs.com/blog/content-recommendations-media-websites

Collaborative Filtering

- Based previous customers interaction with the products.



Datasets

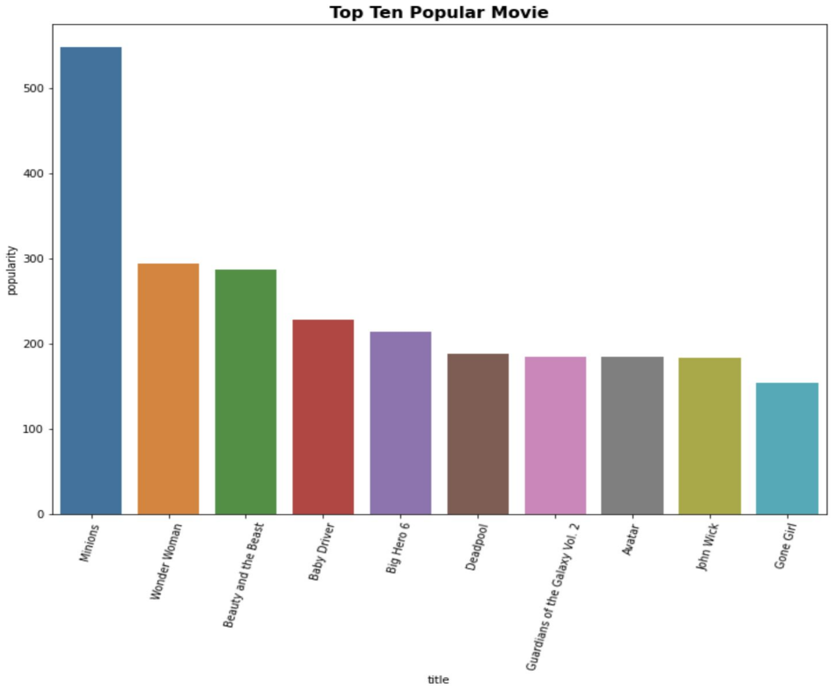
Movies Metadata from kaggle

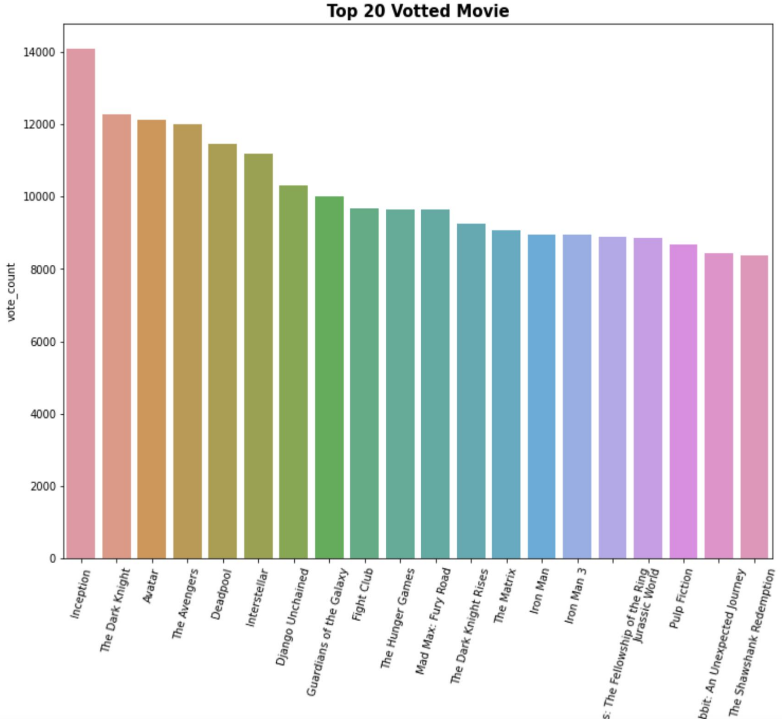
- Has over 45, 000 movies metadata
- For content based filtering

MoviesLens:

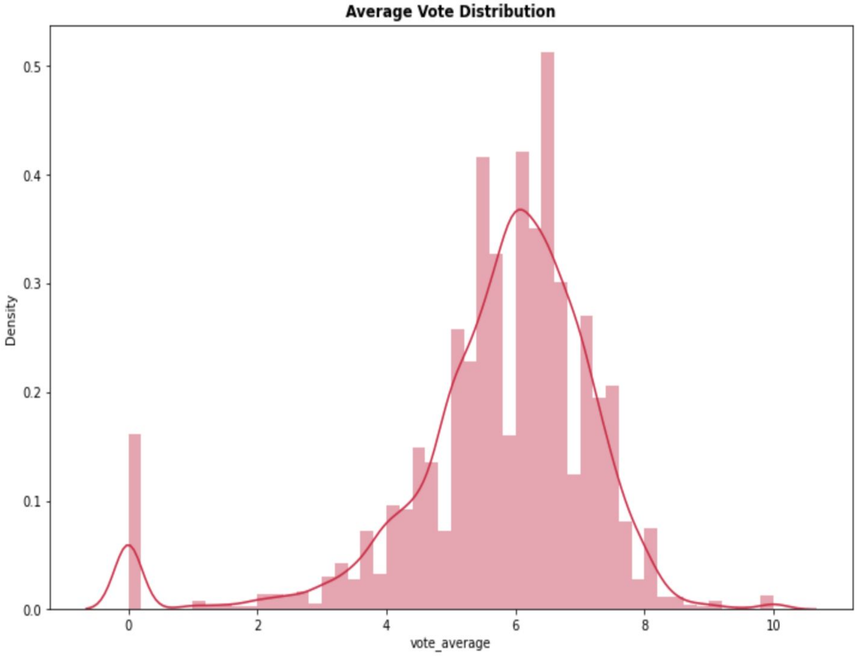
- 1000,000 ratings and 3,600 tag applications applied to 9,000 movies by 600 users. Last updated 9/2018.
- For Collaborative Filtering

Data Exploration and Visualization

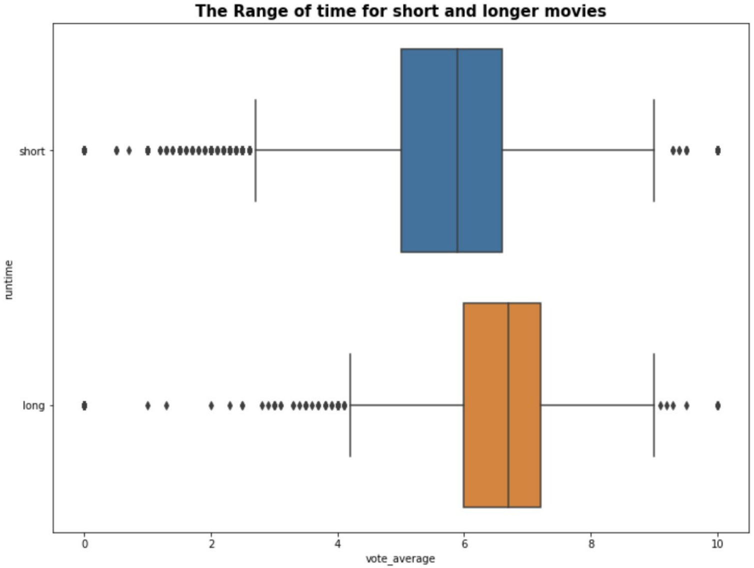




The Average Vote Count



Movie duration



Model Development and Evaluation/ Tests

Content Based Filtering:

- NLP Processing
- TF-IDF
- Get_close_match

Text Analysis

Testing The Model

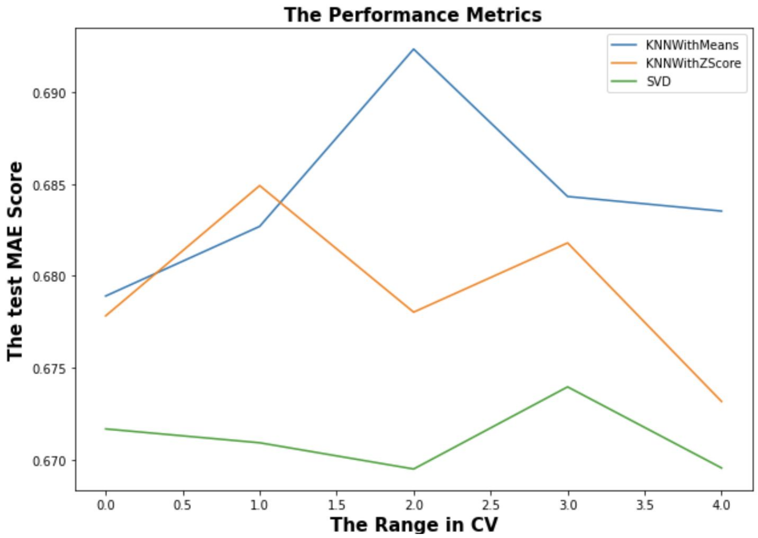
Recommendation for THE DEAD POOL

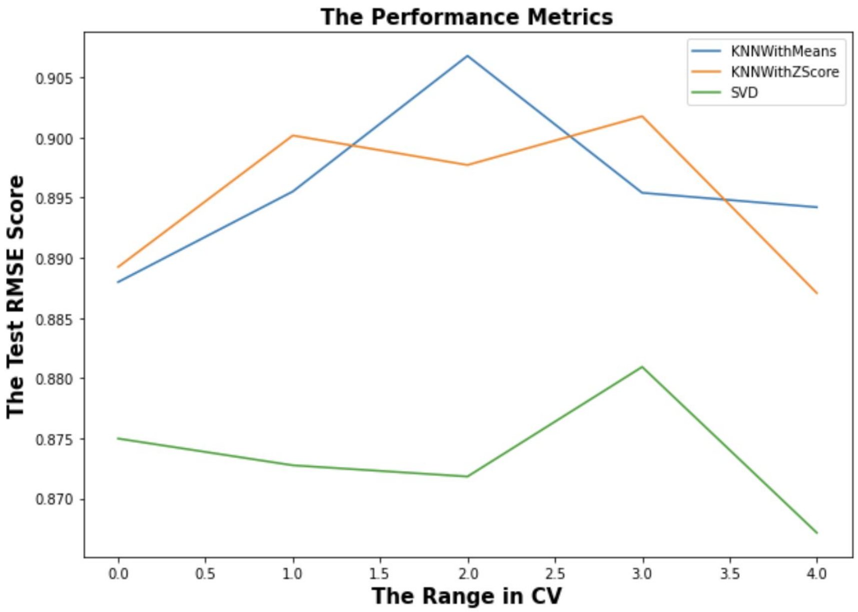
	Movies	Similarity_coef
0	The Enforcer	0.372507
1	Sudden Impact	0.352133
2	Dirty Harry	0.318082
3	Magnum Force	0.268430
4	Let's Get Harry	0.250055
5	Luv	0.241683
6	Harry Potter and the Philosopher's Stone	0.215975
7	Harry Potter and the Goblet of Fire	0.214676
8	Bullet to Beijing	0.211203
9	Deconstructing Harry	0.207683
10	Harry Potter and the Order of the Phoenix	0.205586
11	Harry Potter and the Prisoner of Azkaban	0.201263
12	Wild About Harry	0.199905
13	Harry Potter and the Deathly Hallows: Part 2	0.198425

Collaborative Filtering

User Based Collaborative:

- Predicting the rating of a new item based on customers other items rating
- Use the different algorithms
- Singular Value Decomposition (SVD)





Testing the SVD

```
Predictions for user Id:  7
```

	Movies	predicted_rating
0	Shawshank Redemption, The (1994)	4.314931
1	Dark Knight, The (2008)	4.242832
2	Memento (2000)	4.240722
3	Dark Knight Rises, The (2012)	4.215771
4	Rear Window (1954)	4.184042
5	Casablanca (1942)	4.170300
6	Alien (1979)	4.156911
7	Maltese Falcon, The (1941)	4.147477
8	Princess Bride, The (1987)	4.127199
9	Groundhog Day (1993)	4.101420

```
user_prediction("23", rec_n, movie_id_title_map)
```

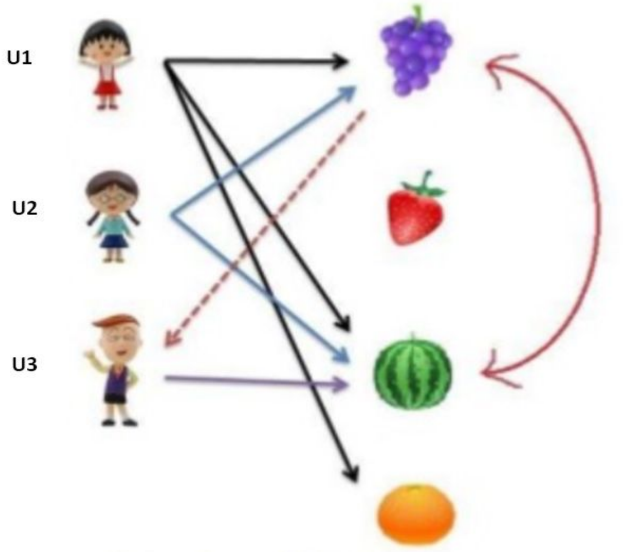
```
Predictions for user Id:  23
```

	Movies	predicted_rating
0	Lord of the Rings: The Fellowship of the Ring, The (2001)	4.162995
1	Three Billboards Outside Ebbing, Missouri (2017)	4.153661
2	Philadelphia Story, The (1940)	4.152012
3	To Catch a Thief (1955)	4.140362
4	3:10 to Yuma (2007)	4.129838
5	Boogie Nights (1997)	4.122272
6	Lawrence of Arabia (1962)	4.120965
7	Whiplash (2014)	4.101586
8	Inglourious Basterds (2009)	4.081907
9	Boondock Saints, The (2000)	4.075175

Item Based Collaborative Filtering

- Suggest products based customers past preference.
- NearestNeighbors(KNN)
  - Select most popular class from the neighbors.

```
model_knn = NearestNeighbors(metric = 'cosine', algorithm = 'brute')
model_knn.fit(full_movie_data_matrix)
```



Item-based filtering

Testing

Recommendation for SERGEANT YORK (1941) :

	Movies	Similarity_coef
0	High Society (1956)	0.387758
1	Detour (1945)	0.370734
2	Captain Blood (1935)	0.331035
3	National Velvet (1944)	0.328544
4	Gunga Din (1939)	0.289588
5	Goodbye, Mr. Chips (1939)	0.259383
6	Grey Zone, The (2001)	0.259383
7	Catch a Fire (2006)	0.252497
8	Fury (1936)	0.202213
9	Ten Commandments, The (1956)	0.111975
10	Shootist, The (1976)	0.106500
11	Lorenzo's Oil (1992)	0.106500
12	Seven Brides for Seven Brothers (1954)	0.106500
13	Country Girl, The (1954)	0.000000

Recommendation for TIMER (2009) :

	Movies	Similarity_coef
0	Lost in Austen (2008)	0.247423
1	Air I Breathe, The (2007)	0.247423
2	Stoning of Soraya M., The (2008)	0.247423
3	Outsourced (2006)	0.247423
4	In Hell (2003)	0.247423
5	Nothing to Lose (1994)	0.247423
6	American Psycho II: All American Girl (2002)	0.247423
7	Gloomy Sunday (Ein Lied von Liebe und Tod) (1999)	0.247423
8	What's Your Number? (2011)	0.247423
9	Patriot, The (1998)	0.247423
10	American Drug War: The Last White Hope (2007)	0.247423
11	Zombie Strippers! (2008)	0.247423
12	Bigger, Stronger, Faster* (2008)	0.118504
13	Bella (2006)	0.000000

Conclusions & Future Works

Conclusions:

- Content and collaborative filtering movie recommender system with python.
- The cosine similarity used for content-based filtering,
- Matrix Factorization, and K Nearest neighborhood for collaborative filtering methods.
- Data science flow for model development.

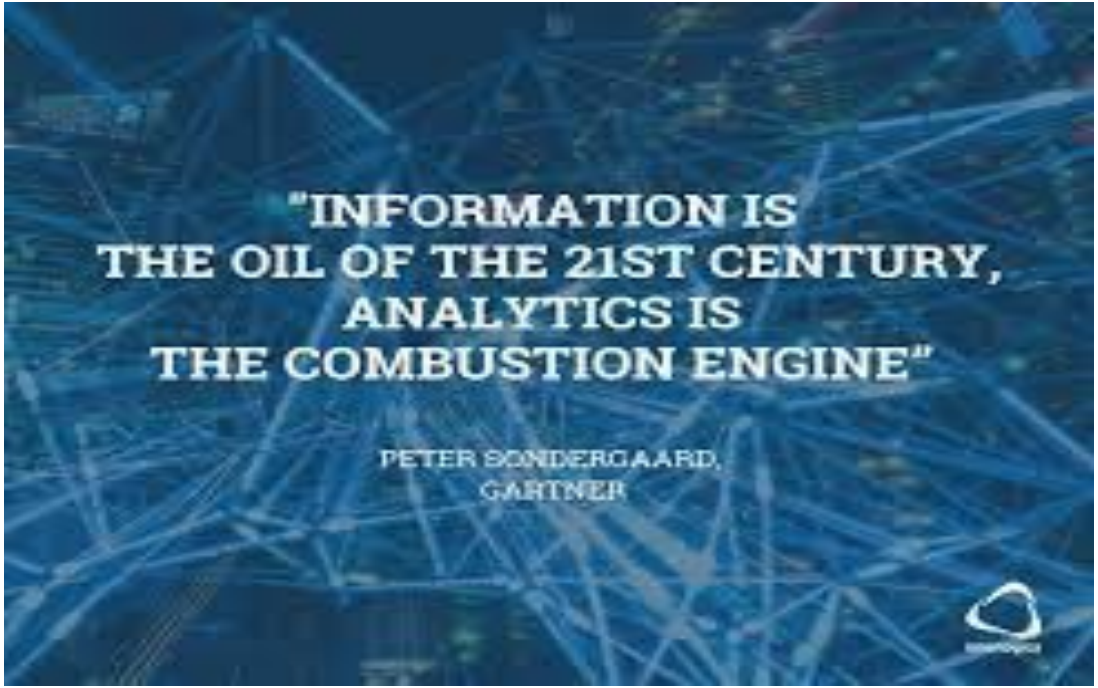
Future Works:

- Add more features such as, age, profession, gender
- Increase the data size to millions
- Create a new movie weight for recommendation

References

1. https://medium.com/@cfpinela/recommender-systems-user-based-and-item-based-collaborative-filtering-5d5f375a127f
2. https://towardsdatascience.com/how-did-we-build-book-recommender-systems-in-an-hour-part-2-k-nearest-neighbors-and-matrix-c04b3c2ef55c
3. https://medium.com/@cfpinela/recommender-systems-user-based-and-item-based-collaborative-filtering-5d5f375a127f

Thank You


"INFORMATION IS
THE OIL OF THE 21ST CENTURY,
ANALYTICS IS
THE COMBUSTION ENGINE"
PETER SONDERGAARD,
GARTNER