Netflix Recommendation System

Wenjing Chen, Yan(Jeffrey) Chen, Yang Liu,

Yicheng Ren, Meghan Rokas, Lingyi Zhao

Agenda

- Executive Summary
- Business Use Case
- Data Profile
- Data Cleaning
- Exploratory Data Analysis
- NLP Sentiment Analysis
- Collaborative Filtering
- Clustering
- Content Based
- Challenges & Improvement
- Reference

Executive Summary

Netflix - Overview

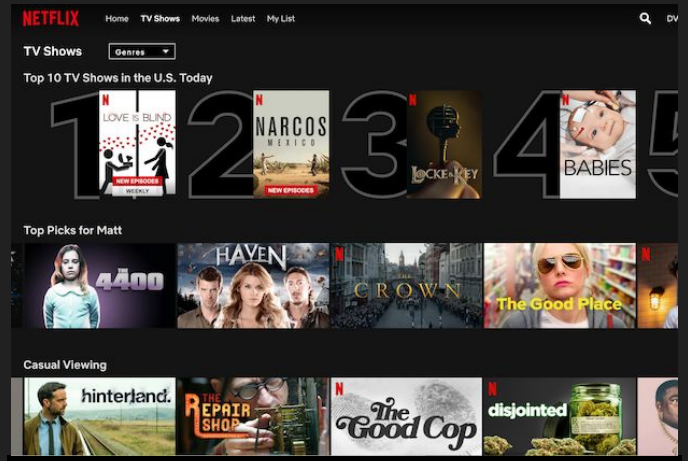
A subscription-based streaming service that allows our members to watch TV shows and movies without commercials on an internet-connected device.

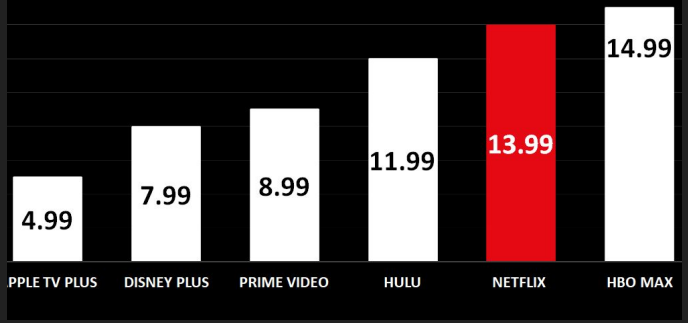
The Problem

Recommendation systems may not efficiently target audiences with high-rating movies that are similar to the ones watched in the past.

The Solutions

Improve Netflix's recommendation system by using data mining skills.





Business Use Case

Personalized recommendation is the key to Netflix's success model.

As strong competitors such as Disney+ and HBO Max join the streaming service war, we want to provide the best content to attract as many customers as we could.

Every year, Netflix invests billions into acquiring and creating new contents to satisfy their customers. Specifically, Netflix has accumulated over 3700 movies and 1900 TV shows in total over the years. No one has the time to watch them all. It is also time-consuming for customers to figure out what will be their favorite ones. Thus, Netflix must present the most interesting content to each customer on their homepage.

Data Profiles

Data Profile

Netflix Prize Dataset

17770 movies, 480189 customers and 100 million customer ratings

```
1:
1488844,3,2005-09-06
822109,5,2005-05-13
885013,4,2005-10-19
30878,4,2005-12-26
823519,3,2004-05-03
893988,3,2005-11-17
124105,4,2004-08-05
1248029,3,2004-04-22
1842128,4,2004-05-09
2238063,3,2005-05-11
1503895,4,2005-05-19
2207774,5,2005-06-06
2590061,3,2004-08-12
2442,3,2004-04-14
543865,4,2004-05-28
1209119,4,2004-03-23
804919,4,2004-06-10
1086807,3,2004-12-28
1711859,4,2005-05-08
372233,5,2005-11-23
1080361,3,2005-03-28
1245640,3,2005-12-19
558634,4,2004-12-14
2165002,4,2004-04-06
1181550,3,2004-02-01
1227322,4,2004-02-06
427928,4,2004-02-26
814701,5,2005-09-29
```

	index	year	title
0	1	2003	Dinosaur Planet
1	2	2004	Isle of Man TT 2004 Review
2	3	1997	Character
3	4	1994	Paula Abdul's Get Up & Dance
4	5	2004	The Rise and Fall of ECW
5	6	1997	Sick
6	7	1992	8 Man
7	8	2004	What the #$*! Do We Know!?
8	9	1991	Class of Nuke 'Em High 2
9	10	2001	Fighter

Data Profile

IMDB Movies

85855 movies with attributes such as title, year, genre, duration, country, language, director, actors, description, and avg_vote (rating)

	imdb_title_id	title	original_title	year	date_published	genre	duration	country	language	director	...	actors	description	avg_vote
0	tt0000009	Miss Jerry	Miss Jerry	1894	1894-10-09	Romance	45	USA	None	Alexander Black	...	Blanche Bayliss, William Courtenay, Chauncey D...	The adventures of a female reporter in the 1890s.	5.9
1	tt0000574	The Story of the Kelly Gang	The Story of the Kelly Gang	1906	1906-12-26	Biography, Crime, Drama	70	Australia	None	Charles Tait	...	Elizabeth Tait, John Tait, Norman Campbell, Be...	True story of notorious Australian outlaw Ned ...	6.1
2	tt0001892	Den sorte drøm	Den sorte drøm	1911	1911-08-19	Drama	53	Denmark, Germany	NaN	Urban Gad	...	Asta Nielsen, Valdemar Psilander, Gunnar Helse...	Two men of high rank are both wooing the beaut...	5.8
3	tt0002101	Cleopatra	Cleopatra	1912	1912-11-13	Drama, History	100	USA	English	Charles L. Gaskill	...	Helen Gardner, Pearl Sindelar, Miss Fielding, ...	The fabled queen of Egypt's affair with Roman ...	5.2

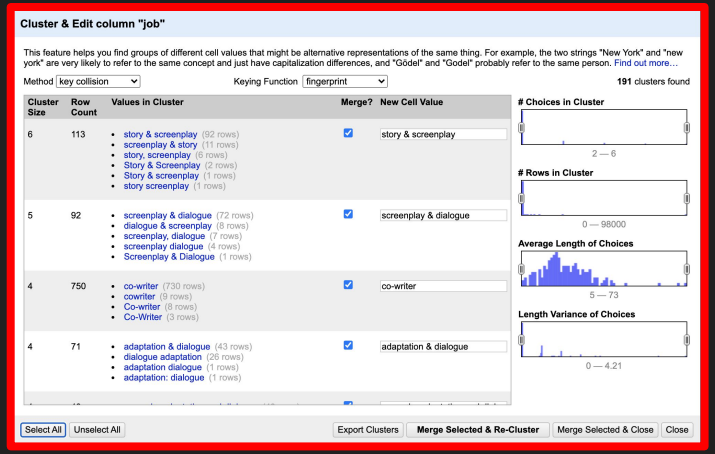
IMDB Reviews

This second IMDB dataset from kaggle that just has 49,582 randomly selected movie reviews. Because our large IMDB dataset does not contain the actual text of the reviews, just ratings, we decided to add in this dataset to provide room for extra analysis.

Data Cleaning

OpenRefine

- Initial cleaning using OpenRefine:
  - spelling errors, genre clustering



Data Cleaning

Python

- Secondary cleaning:
- Fill nulls, combining datasets, formatting data types

```
dfMovie.head()
```

title	title	genre	country	language	director	actors	description
Miss Jerry	[miss, jerry]	[romance]	[usa]	[none]	alexanderblack	[blanchebayliss, williamcourtenay, chaunceydepew]	The adventures of a female reporter in the 1890s.
The Story of the Kelly Gang	[the, story, of, the, kelly, gang]	[biography, crime, drama]	[australia]	[none]	charlestait	[elizabethtait, johntait, normancampbell]	True story of notorious Australian outlaw Ned ...
Den sorte drøm	[den, sorte, drøm]	[drama]	[denmark, germany]	[nan]	urbangad	[astanielsen, valdemarpsilander, gunnarhelseng...	Two men of high rank are both wooing the beaut...
Cleopatra	[cleopatra]	[drama, history]	[usa]	[english]	charlesl.gaskill	[helengardner, pearlsindelar, missfielding]	The fabled queen of Egypt's affair with Roman ...
L'Inferno	[l'inferno]	[adventure, drama, fantasy]	[italy]	[italian]	francescobertolini,adolfopadovan	[salvatorepapa, arturopirovano, giuseppedeligu...	Loosely adapted from Dante's Divine Comedy and...

Data Cleaning

Python Cont.

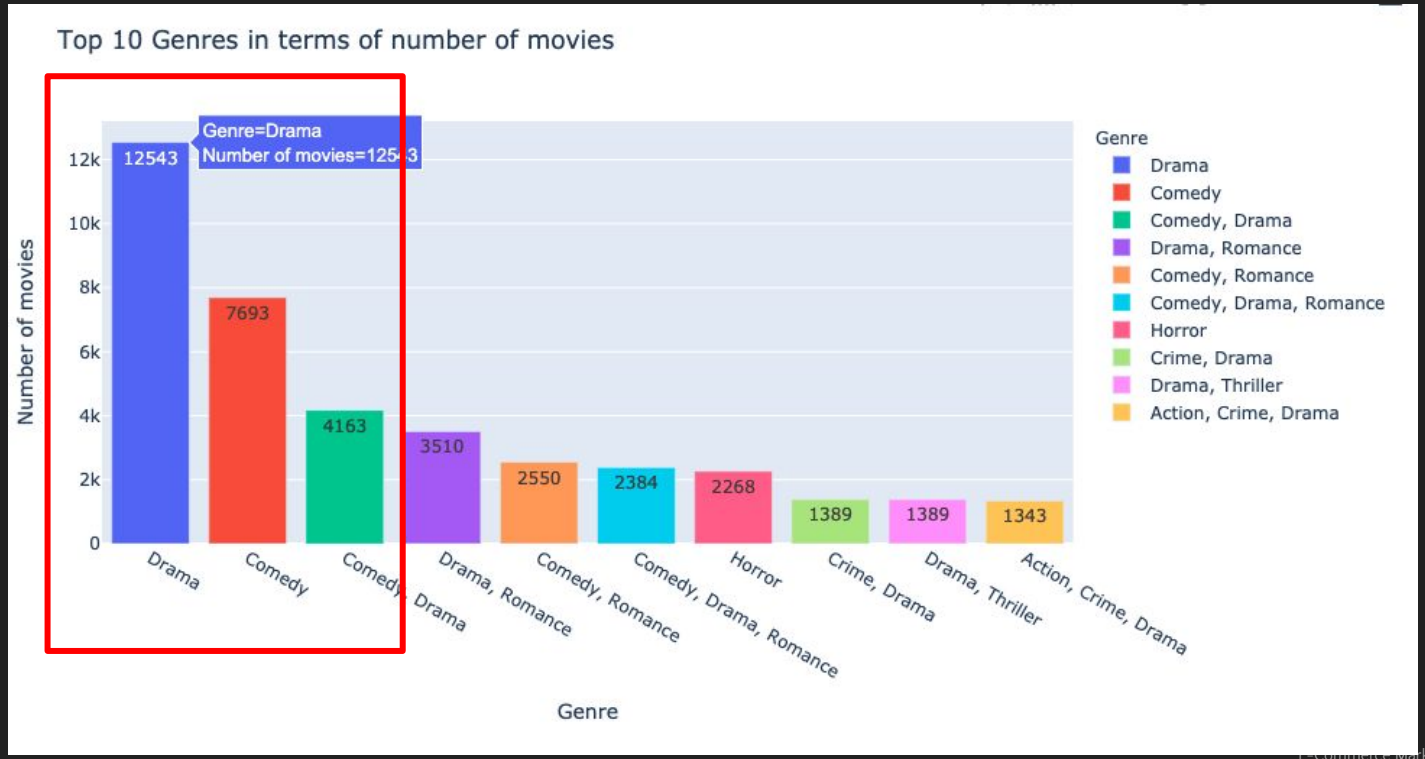
- Combine the four separate txt files to a csv file
- We end up with one file of 100,480,507 rows of data

	movie_id	user_id	rating	date
0	1	1488844	3	2005-09-06
1	1	822109	5	2005-05-13
2	1	885013	4	2005-10-19
3	1	30878	4	2005-12-26
4	1	823519	3	2004-05-03
...	...	...	...	...
100480502	17770	1790158	4	2005-11-01
100480503	17770	1608708	3	2005-07-19
100480504	17770	234275	1	2004-08-07
100480505	17770	255278	4	2004-05-28
100480506	17770	453585	2	2005-03-10

EDA

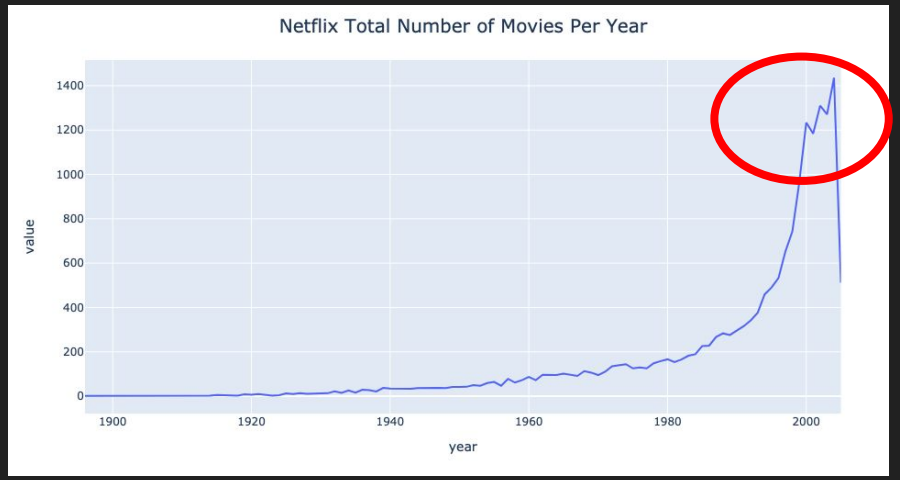
Exploratory Data Analysis

IMDB Dataset

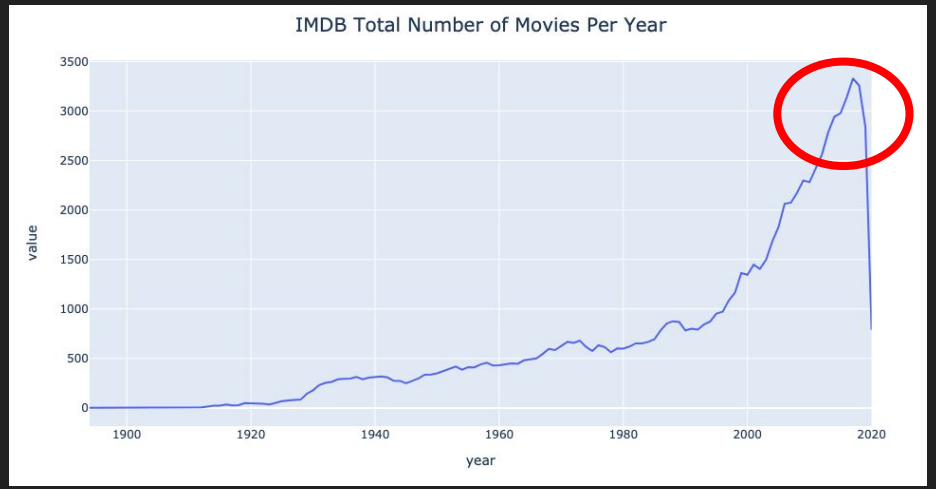


Exploratory Data Analysis

Netflix Prize Dataset

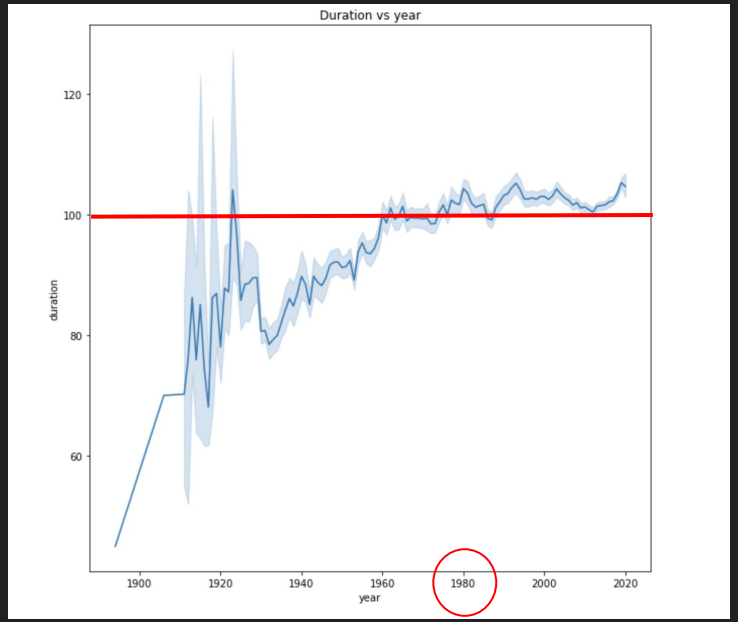


IMDB Dataset

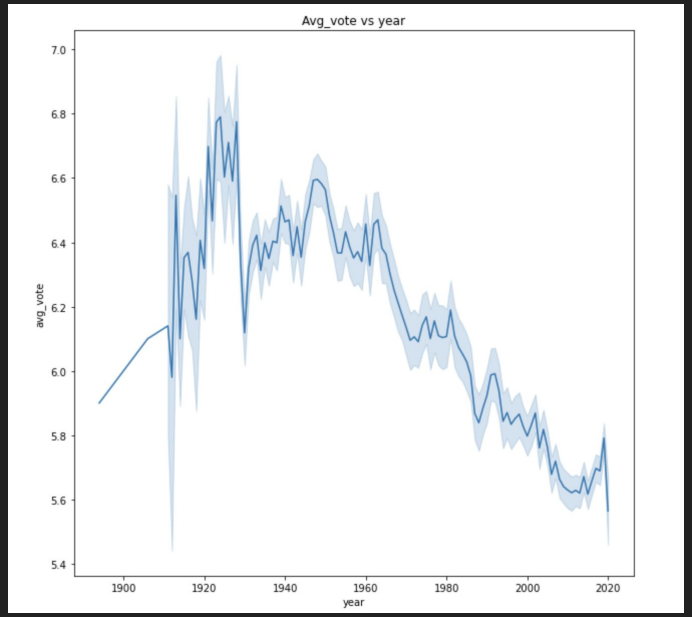


Exploratory Data Analysis

IMDB Dataset



Movie duration vs. year



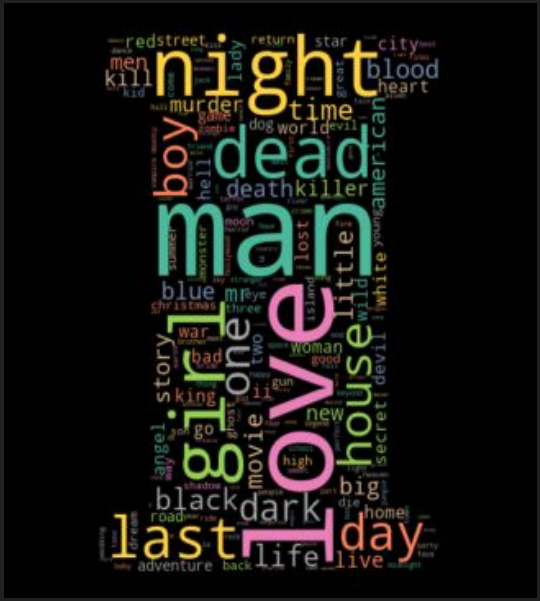
avg_vote vs. year

Exploratory Data Analysis

Netflix Prize Dataset

Top 5 key words in Netflix
Season, live, collection, love, material

IMDB Dataset

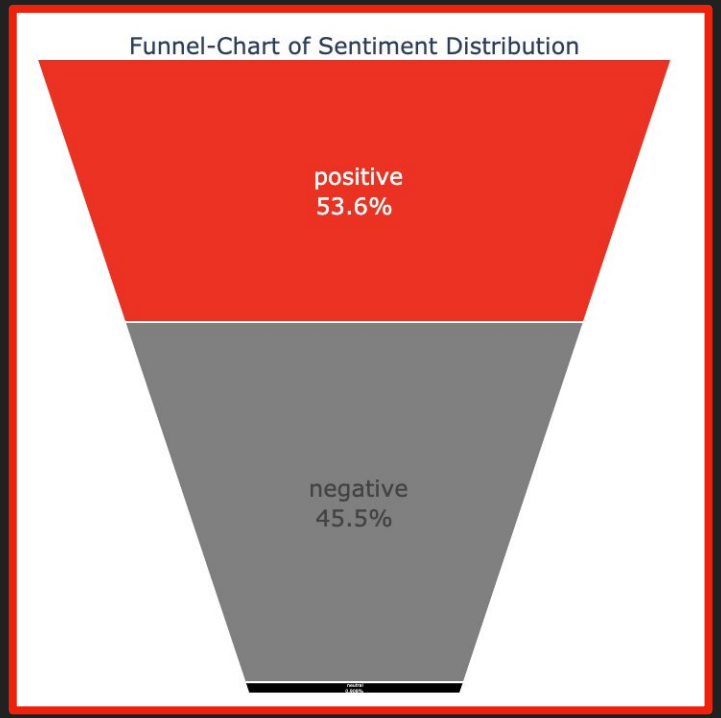


Top 5 key words in IMDB
Love, man, girl, night, dead

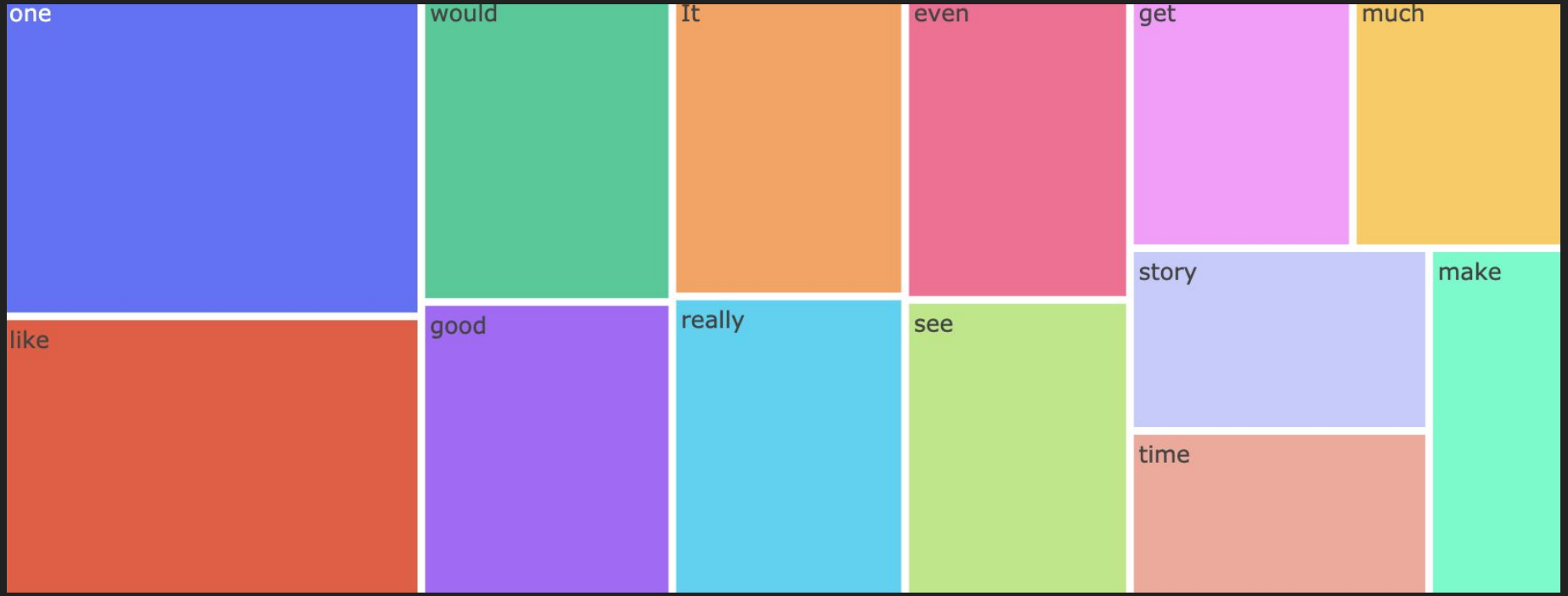
IMDB Sentiment Analysis

IMDB Review Sentiment Analysis

	sentiment	text
2	positive	26801
0	negative	22741
1	neutral	454

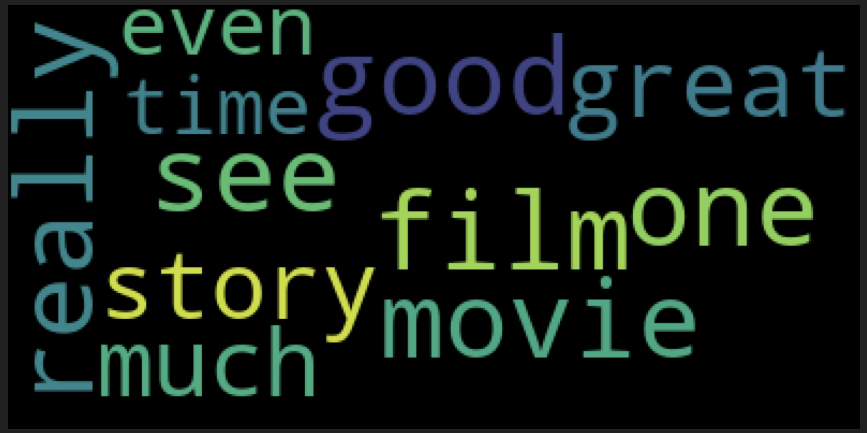


IMDB Review Sentiment Analysis

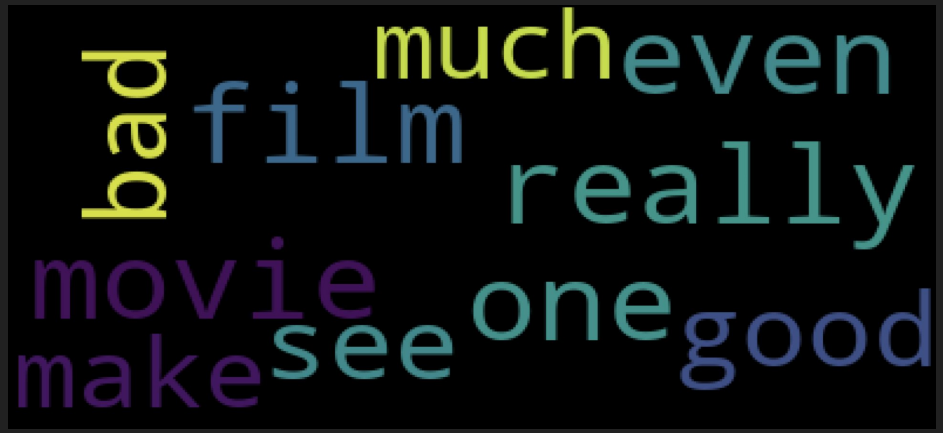


IMDB Review Sentiment Analysis - Top Words

Positive



Negative



Collaborative Filtering

Collaborative Filtering

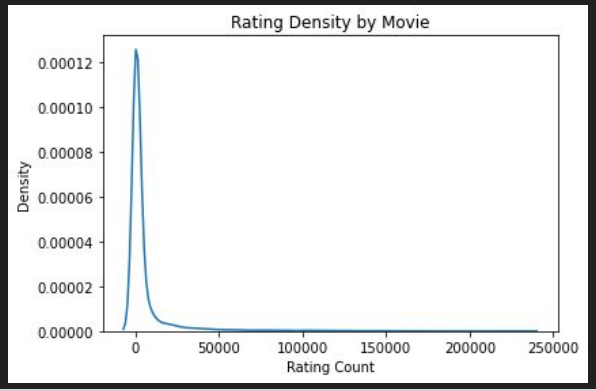
Data: Netflix Prize Dataset - user_id, movie_id, rating

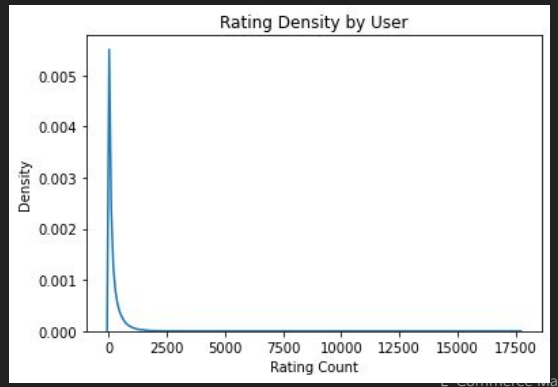
Target: rating

Percent of missing ratings: 98.86%

Outliers: outliers on the right represents influential users or popular movies

Density plot for number of ratings





Collaborative Filtering

Movielens 1M	RMSE	MAE	Time
SVD	0.873	0.686	0:02:13
SVD++	0.862	0.673	2:54:19
NMF	0.916	0.724	0:02:31
Slope One	0.907	0.715	0:02:31
k-NN	0.923	0.727	0:05:27
Centered k-NN	0.929	0.738	0:05:43
k-NN Baseline	0.895	0.706	0:05:55
Co-Clustering	0.915	0.717	0:00:31
Baseline	0.909	0.719	0:00:19
Random	1.504	1.206	0:00:19

Source: http://surpriselib.com/

Collaborative Filtering

- SVD
  - Training speed: ~1 hour 20 min
  - High memory cost: not always feasible to train on a local machine
- Ensembled SVD:
  - Partition data by user_id, then apply SVD over each group of users
  - Recommend using a weighted average of the SVD predictions
  - Training speed: ~1 hour 20 min
- Prediction:
  - The rating for each user for all the movies
  - Format: a list of dictionaries that can be stored as json
  - Speed (for non-ensembled SVD): 0.08s/user

```
{'user_id': 1,
 'recommend':         id     year         title          rating_pred
 3455    3456  2004.0       Lost: Season 1   4.648350
 8446    8447  2001.0         24: Season 1   4.586378
 13503  13504  2004.0                House   4.557277
 14549  14550  1994.0  The Shawshank Re...   4.556644
 17084  17085  2002.0         24: Season 2   4.542718
 ...      ...     ...                  ...          ...
 3574    3575  2005.0  The Worst Horror...   1.467725
 11767  11768  2004.0  Alone in a Haunt...   1.466566
 15572  15573  2005.0   Rise of the Undead   1.459531
 4201    4202  2004.0           Half-Caste   1.459440
 514      515  2005.0  Avia Vampire Hunter   1.330242

 [17770 rows x 4 columns]}
```

Collaborative Filtering

- Hyperparameter-tuning on 1M data
- RandomizedSearchCV
- Candidates:
  - 'n_factors': [50,100,200]
  - 'n_epochs': [20,40]
  - 'lr_all': [0.005,0.001]
  - 'reg_all': [0.05,0.02,0.01]}
- Best parameters:
  - ‘N_factors’ = 50
  - ‘N_epochs’ = 20
  - ‘Lr_all’ = 0.005
  - ‘Reg_all’ = 0.05

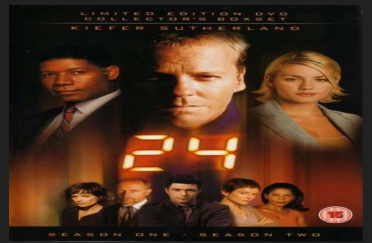
Collaborative Filtering - Evaluation

- SVD
  - 5 fold cross-validation (on full dataset)
  - RMSE: 0.8668
  - 5 fold cross-validation (on 1M data)
  - RMSE: 0.9716
- Ensembled SVD
  - Train test split
  - Test RMSE (on 1M data): 0.8770
- Benchmark: Naive model (randomly select a rating based on the population distribution)
  - RMSE: 1.5347

Recommended for you

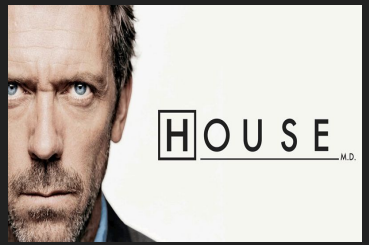
Lost: Season 1

The survivors of Oceanic Flight 815 were 1,000 miles off course when they crashed on a lush, mysterious island.



24: Season 1

Counterterrorism agent Jack Bauer fights the bad guys of the world, a day at a time. With each week's episode unfolding in real time, "24" covers a single day in the life of Bauer each season.



House

At fictional Princeton Plainsboro Teaching Hospital in New Jersey, prickly genius Dr. Gregory House tackles health mysteries as would a medical Sherlock Holmes.

The Shawshank Redemption

Andy Dufresne (Tim Robbins) is sentenced to two consecutive life terms in prison for the murders of his wife and her lover and is sentenced to a tough prison. However, only Andy knows he didn't commit the crimes.

Reasons of not Combining Two Datasets

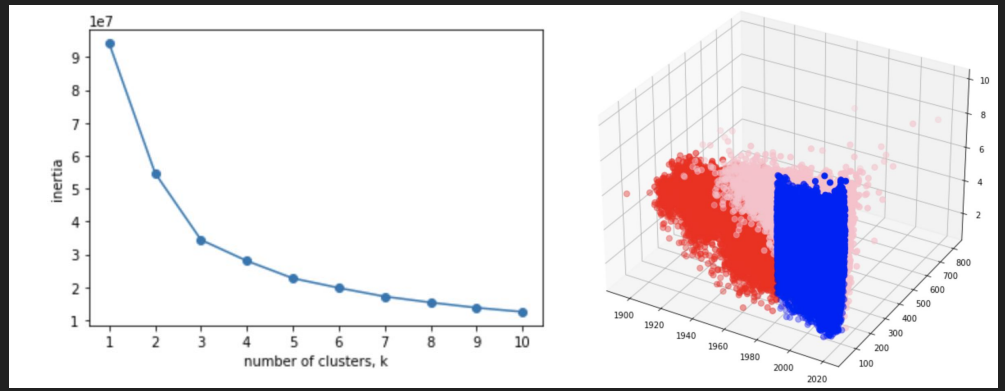
- Limitations
  - Netflix Dataset has only users' information and which movies users watched. (No movies' information)
  - IMDB Movies Dataset has only movies information. (No users' information)
  - The overlapping of two datasets only contains less than 4000 movies
  - Demo at the end the presentation to show how we can combine these two algorithms
- Our Current Approach
  - Use the IMDB_movie dataset to recommend Netflix which movies' copyrights they could include in its databases

	imdb_title_id	title	original_title	year	date_published	genre	duration	country	language	director	...	actors	description	avg_vote
0	tt0000009	Miss Jerry	Miss Jerry	1894	1894-10-09	Romance	45	USA	None	Alexander Black	...	Blanche Bayliss, William Courtenay, Chauncey D...	The adventures of a female reporter in the 1890s.	5.9
1	tt0000574	The Story of the Kelly Gang	The Story of the Kelly Gang	1906	1906-12-26	Biography, Crime, Drama	70	Australia	None	Charles Tait	...	Elizabeth Tait, John Tait, Norman Campbell, Be...	True story of notorious Australian outlaw Ned ...	6.1

Content-Based

Content Based Recommendation - Clustering

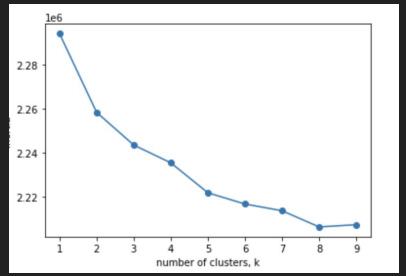
- Cluster based on numeric variables
  - Gather all the necessary numeric columns
  - Apply K-means (k = 3)
  - Assign the group for each movie by adding a new column called “description_group”



title	year	duration	avg_vote
Miss Jerry	1894	45	5.9
The Story of the Kelly Gang	1906	70	6.1
Den sorte drøm	1911	53	5.8
Cleopatra	1912	100	5.2
L'Inferno	1911	68	7.0
...	...	...	...
Le lion	2020	95	5.3
De Beentjes van Sint-Hildegard	2020	103	7.7
Padmavyuhathile Abhimanyu	2019	130	7.9
Sokagin Çocuklari	2019	98	6.4
La vida sense la Sara Amat	2019	74	6.7

Content Based Recommendation - Clustering

- Cluster based on “string” variables
  - Clean data
  - Turn all the cols to the “bag_of_word” in which we remove all the useless words
  - Apply K-means (k = 8)



  - Assign the group for each movie by adding a new column called “description_group”

```
EMovie.head()
```

title	title	genre	country	language	director	actors	Key_words
Miss Jerry	[miss, jerry]	[romance]	[usa]	[none]	alexanderblack	[blanchebayliss, williamcourtenay, chauncey depew]	[1890s, adventures, female, reporter]
The Story of the Kelly Gang	[the, story, of, the, kelly, gang]	[biography, crime, drama]	[australia]	[none]	charlestait	[elizabethtait, johntait, normancampbell]	[notorious, australian, outlaw, ned, kelly, 80...
Den sorte drøm	[den, sorte, drøm]	[drama]	[denmark, germany]	[nan]	urbangad	[astanielsen, valdemarpsilander, gunnarhelseng...	[home, famous, equestrian, acrobat, stella, je...
Cleopatra	[cleopatra]	[drama, history]	[usa]	[english]	charlesl.gaskill	[helengardner, pearlsindelar, missfielding]	[fabled, queen, egypt, ultimately, disastrous,...
L'Inferno	[l'inferno]	[adventure, drama, fantasy]	[italy]	[italian	escobertolini,adolfopadovan	[salvatorepapa, arturopirovano, giuseppedeligu...	[divine, comedy, illustrations, original, sile...

title	bag_of_words
Miss Jerry	miss jerry romance usa none alexanderblack bla...
The Story of the Kelly Gang	the story of the kelly gang biography crime ...
Den sorte drøm	den sorte drøm drama denmark germany nan urba...
Cleopatra	cleopatra drama history usa english charlesl....
L'Inferno	l'inferno adventure drama fantasy italy ital...

Content Based Recommendation - Clustering

- Cluster based on previous results
  - Manually assign the cluster by a function

title	description	numeric
Miss Jerry	1	0
The Story of the Kelly Gang	5	0
Den sorte drøm	5	0
Cleopatra	2	0
L'Inferno	1	
...	...	...
Le lion	7	1
De Beentjes van Sint-Hildegard	6	1
Padmavyuhathile Abhimanyu	5	2
Sokagin Çocuklari	5	1
La vida sense la Sara Amat	5	1

Assign new groups based on these two columns

title	description	numeric	cluster
Miss Jerry	1	0	1
The Story of the Kelly Gang	5	0	5
Den sorte drøm	5	0	5
Cleopatra	2	0	2
L'Inferno	1	0	1
...	...	...	...
Le lion	7	1	15
De Beentjes van Sint-Hildegard	6	1	14
Padmavyuhathile Abhimanyu	5	2	21
Sokagin Çocuklari	5	1	13
La vida sense la Sara Amat	5	1	1

Content Based Recommendation

- Use K-Means clustering to divide the whole dataset into 24 (3 * 8) clusters.
- Generate the “count vectorizer” and use cosine similarity to measure how similar the movies are under each cluster.
- Search recommendations based on movie names in clusters.
- Recommendations for ‘Miss Jerry’

title	bag_of_words	group
Miss Jerry	miss jerry romance usa none alexanderblack bla...	12
The Story of the Kelly Gang	the story of the kelly gang biography crime ...	12
Den sorte drøm	den sorte drøm drama denmark germany nan urba...	12
Cleopatra	cleopatra drama history usa english charlesl....	17
L'Inferno	l'inferno adventure drama fantasy italy ital...	12

```
['Il quarantunesimo',
 'Kesällä kello 5',
 'Pink Narcissus',
 'Stolen Moments',
 'Sangue e arena',
 'La commedia umana',
 'Barbed Wire',
 'Akitsu onsen',
 'Feng er ti ta cai',
 'Vase de noces']
```

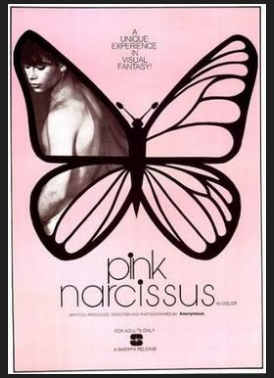
Miss Jerry

Il quarantunesimo

A 1930 American pre-Code crime film directed by Archie Mayo and starring Lew Ayres and James Cagney in his second film role.[2][3] The film was based on the story *A Handful of Clouds*, written by Rowland Brown. The film's title was typical of the sensationalistic titles of many Pre-Code films.[4] It was marketed with the tagline "The picture Gangland defied Hollywood to make!"[5]

'Kesällä kello 5'

A 1969 Australian comedy-drama film directed by Michael Powell. The film stars James Mason (co-producer with Powell), Helen Mirren in her first major film role, and Jack MacGowran, and features actress Neva Carr Glyn.



Pink Narcissus

Pink Narcissus is a 1942 Italian comedy film directed by Gianni Franciolini and starring Lilia Silvi, Amedeo Nazzari and Leonardo Cortese. It was based on a play by Claude-André Puget, which had been made into a French film Les jours heureux the previous year.

Stolen Moments

Stolen Moments is a 1936 American musical film starring Jeanette MacDonald, Nelson Eddy, and Reginald Owen that was directed by W. S. Van Dyke. It was the second of three movie adaptations from Metro-Goldwyn-Mayer of the 1924 Broadway musical of the same name.

Content Based Recommendation - Evaluation

Test by Overlapping from Netflix

1. Figure out recommendation lists for users based on one movie
2. Calculate Precision = # of Correctly Recommend / # of Recommendation Lists
3. Calculate Recall = # of Correctly Recommend / # of Users' Movie Lists

Limitation & Improvement:

1. Overlapping dataset has less than 4000 movies
2. Most users only rate several or dozens of movies
3. Provide recommendation lists based on multiple movies

```
0.07142857142857142,
0.04166666666666664,
0.21428571428571427,
0.023809523809523808,
0.07142857142857142,
0.078125,
0.0,
0.11428571428571428,
0.07142857142857142,
0.0,
0.031746031746031744,
```

Precision
0.056

Precision
0.082

Recall
0.013

Recall
0.042

```
0.026881720430107527,
0.015037593984962405,
0.01968503937007874,
0.01818181818181818,
0.016025641025641024,
0.013227513227513227,
0.0,
0.02185792349726776,
0.021367521367521368,
```

Challenges & Improvement

- Reasonably combine two datasets
  - Try to search overlapping movies between Netflix and IMDB dataset in order to create a small group of movies dataset
- Reasonably combine two algorithms
  - Try to get two lists of movies based on collaborative filtering and content based and compare the result to recommend proper ones
- No users' information
  - Not able to cluster users based on their profiles
- SVD memory cost
  - Solution: ensembled SVD
- Reasonably assign weight to movies' attributes in content-based algorithm
  - Need to do further research on users' preferences, and know which factors they consider the most important

Reference

- https://www.kaggle.com/shivamb/netflix-shows
- https://www.kaggle.com/netflix-inc/netflix-prize-data
- https://www.kaggle.com/stefanoleone992/imdb-extensive-dataset
- http://surpriselib.com/
- https://arxiv.org/pdf/1205.3193.pdf
- http://ai.stanford.edu/~amaas/data/sentiment/

NETFLIX

Q & A