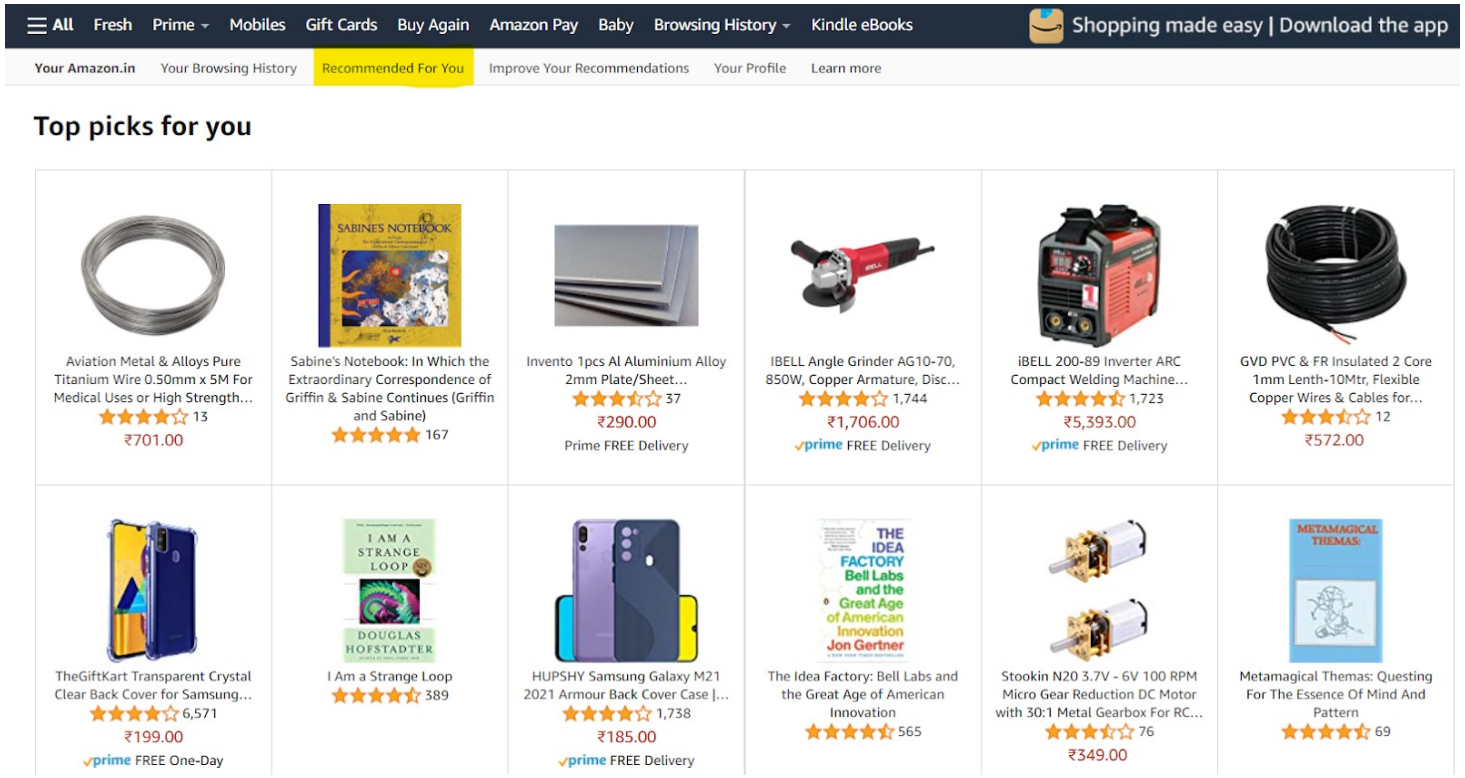
Hybrid recommendation model

Chi Nguyen

Table of content

Project introduction

- This project is to deeply explore recommendation algorithms.
- This project is based on real data from Amazon and its outcome will be a hybrid model to recommend products for Amazon' user.
- Objectives:
  - Earning knowledge about recommendation systems
  - Earning experience in building a hybrid recommender
  - Add an exciting project to portfolio



Data description

- First try with Amazon developer API.
- Finally, I found a data set containing product reviews and metadata from Amazon released by a group of data professionals at UCSD (University of California SanDiego)
- Data source: https://nijianmo.github.io/amazon/index.html

Data description

- Product reviews from amazon.
- The data used for the project is the subset of the full data set span a period of 22 years up to Oct 2018.
- Category narrow down: Movies & TV.
- Include 2 data sets:
  - 13.7 millions reviews
  - 748k metadata

SAMPLE OF REVIEW TABLE

	rating	title	text	parent_asin	user_id
0	5.0	Five Stars	Amazon, please buy the show! I'm hooked!	B013488XFS	AGGZ357AO26RQZVRLGU4D4N52DZQ
1	5.0	Five Stars	My Kiddos LOVE this show!!	B00CB6VTDS	AGKASBHYZPGTEPO6LWZPVJWB2BVA
2	3.0	Some decent moments...but...	Annabella Sciorra did her character justice wi...	B096Z8Z3R6	AG2L7H23R5LLKDKLBEF2Q3L2MVDA
4	5.0	What Love Is...	...isn't always how you expect it to be, but w...	B001H1SVZC	AG2L7H23R5LLKDKLBEF2Q3L2MVDA
5	5.0	QUIRKY TURNS TO HEARTSTRINGS	As you learn about the very unique characters ...	B06WVW16WY	AG2L7H23R5LLKDKLBEF2Q3L2MVDA

SAMPLE OF METADATA TABLE

	main_category	title	average_rating	rating_number	features	description	price	categories	details	parent_asin
1	Prime Video	One Perfect Wedding	3.0	6.0	[IMDb 6.1, 1 h 27 min, 2021, ALL]	[With her book tour in two weeks and his expan...	None	[Comedy, Drama, Romance]	{'Audio languages': ['English'], 'Subtitles': ...	B09WDLJ4HP
5	Movies & TV	Sacred Flesh	3.7	63.0	[]	[Product Description, Torn between sexual desi...	None	[Movies & TV, Genre for Featured Categories, H...	{'Format': 'Color, AC-3, Closed-captioned, Col...	B0001YJBJ6
6	Prime Video	Let Sleeping Corpses Lie	4.2	96.0	[IMDb 6.8, 1 h 32 min, 1975, R]	[A cop chases two hippies suspected of a serie...	None	[Horror, Atmospheric, Intense]	{'Subtitles': ['None available'], 'Directors':...	B018RDV40G
11	Movies & TV	Pink Cadillac [DVD]	4.6	972.0	[]	[A comic action-adventure about a modern-day b...	None	[Movies & TV, Studio Specials, Warner Home Vid...	{'Genre': 'Action & Adventure', 'Format': 'Ana...	B00009N83U
15	Movies & TV	Teddy	5.0	1.0	[]	[From creators Steve Goltz and Kevin Sommerfie...	None	[Movies & TV, Disc on Demand]	{'Format': 'NTSC', 'Contributor': 'Steve Goltz...	B0052ZO9TQ
...	...	...	...	...	...	...	...	...	...	...

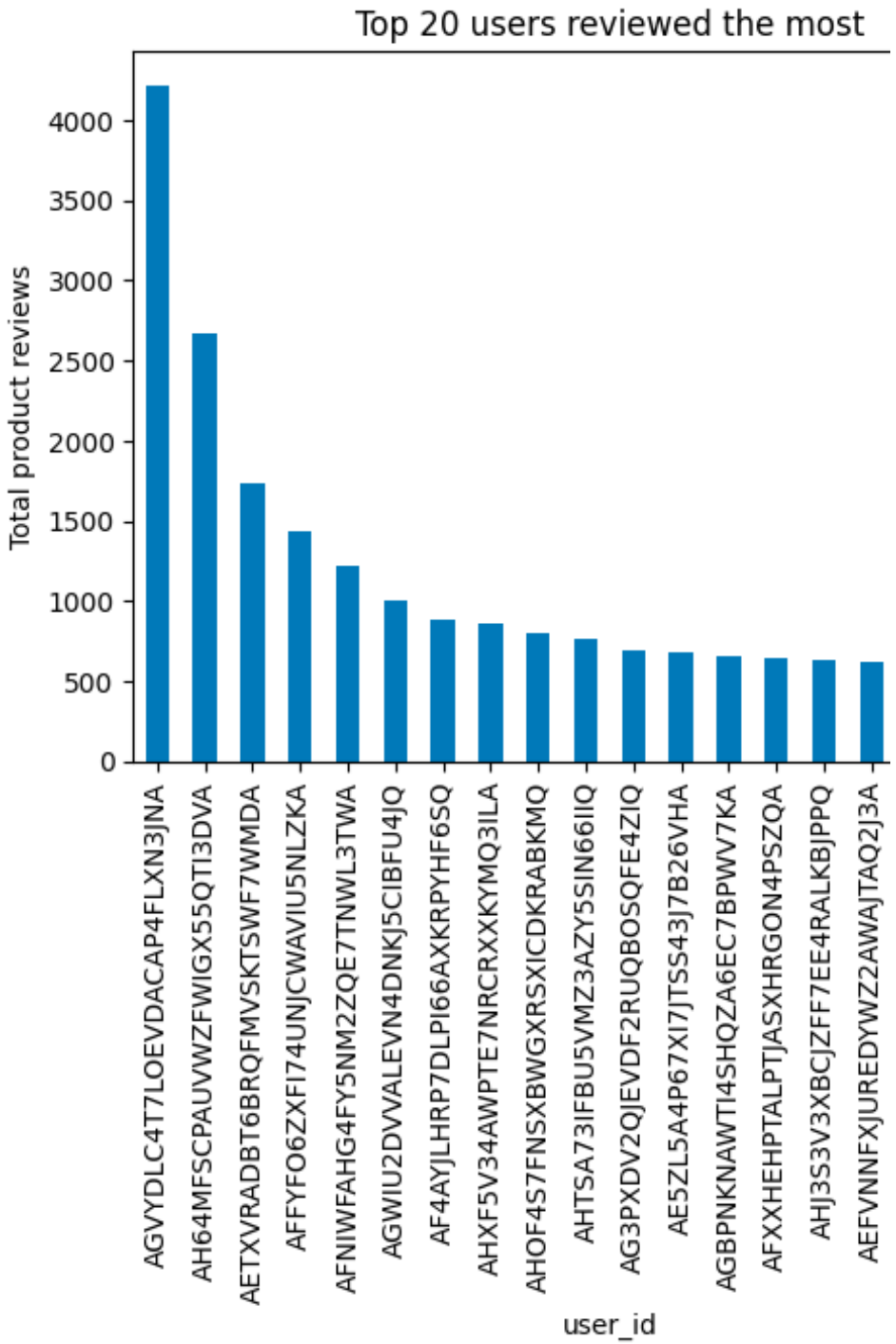
Clean & Inspect the data

Data Cleaning

- Drop missing values
- Preprocess text data in ‘review’, product ‘description’ columns: lowercase, remove punctuation, remove digit, remove stopwords.
- Drop unnecessary columns: features, details, image, video, store, etc.
- Drop unrelated categories:
- Drop duplicates and Merge

‘review’ table with ‘metadata’

table on ‘parent_asin’

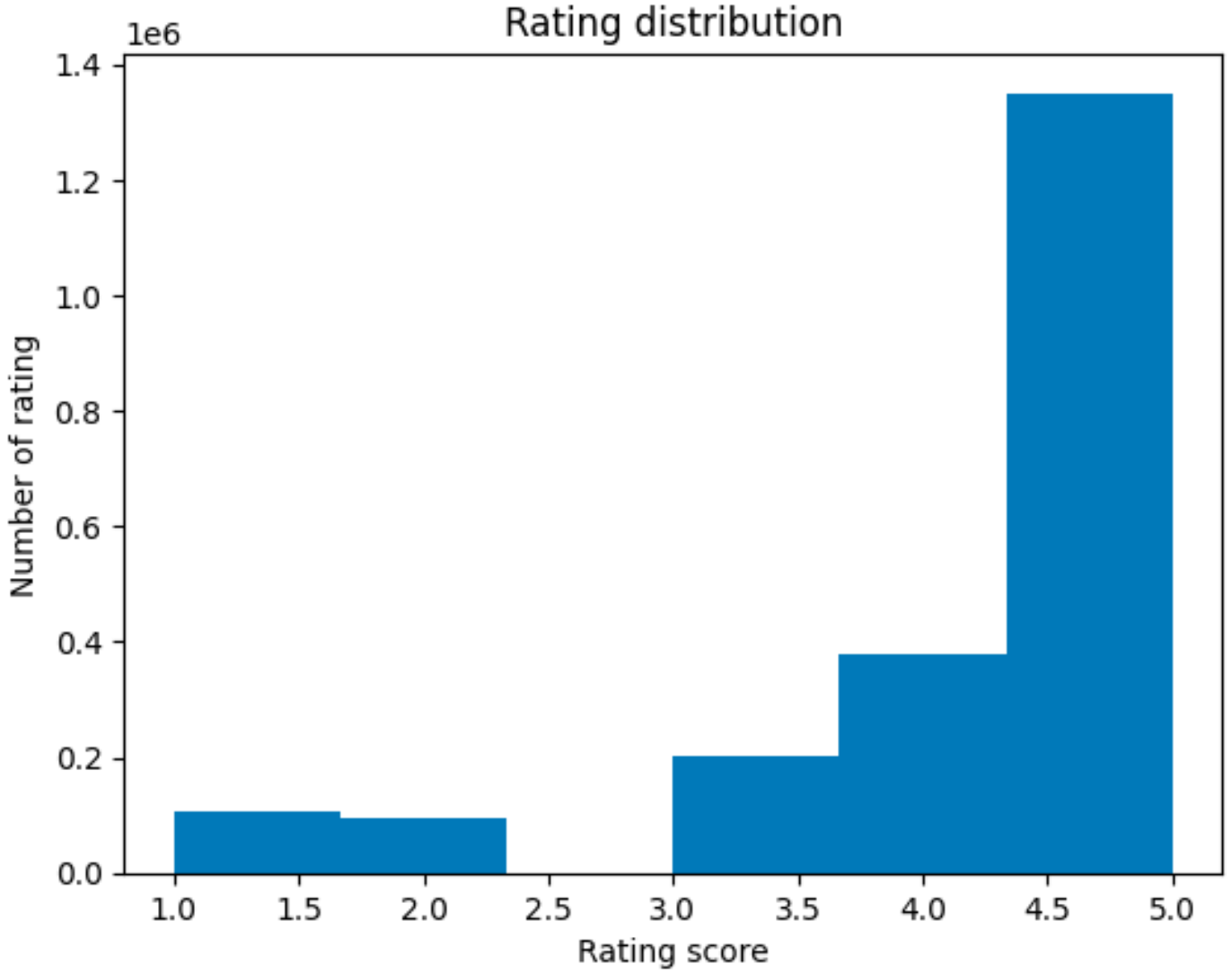


Data inspection: some user purchased huge amount of movies

- A user in the top 20 users gave more than 700 reviews.
- The most active user gave more than 4000 reviews.

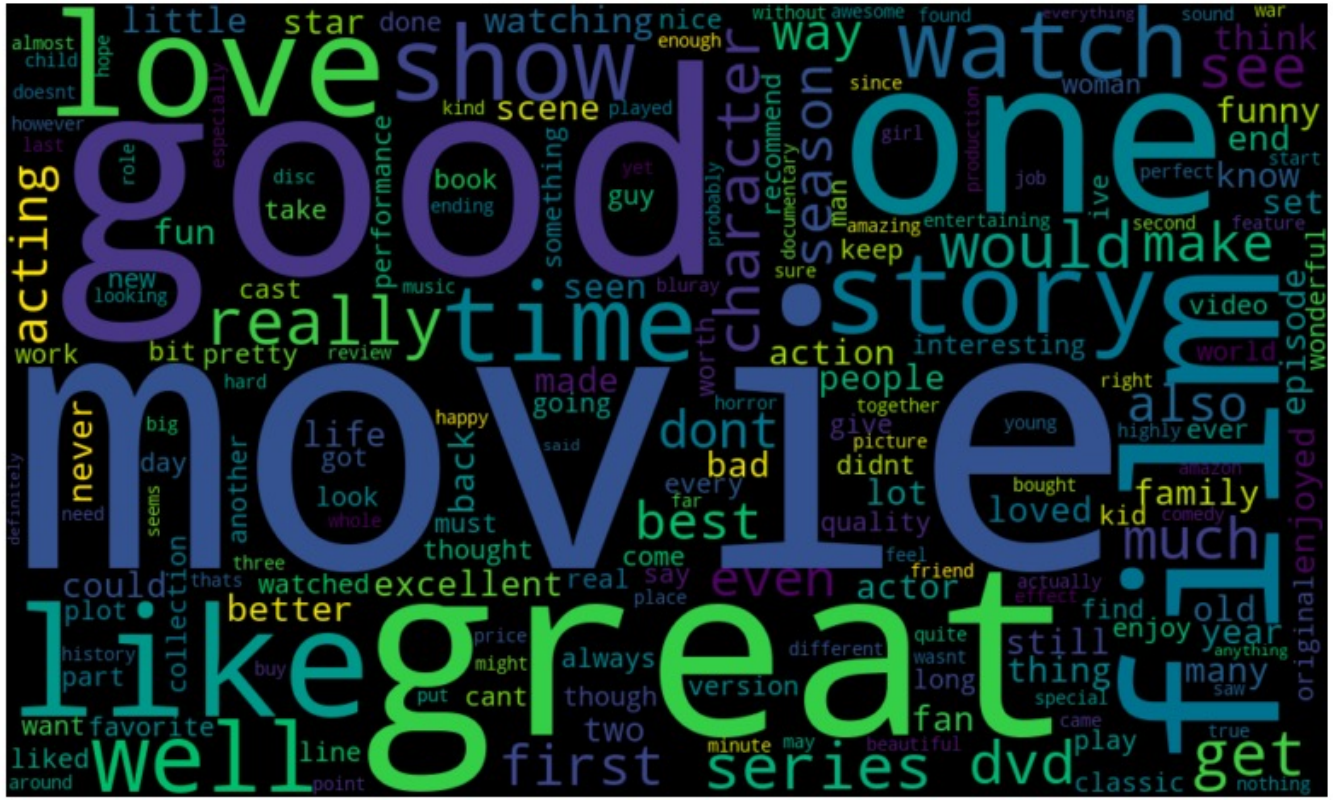
Most of user satisfied with the movies they viewed

- Most of the rating are above 4.0



Nice words given to rated movies

- This is the wordcloud visualization about the reviews user gave for movies they purchased.
- Consistent conclusion with the previous ‘rating distribution’ chart: movies received good feedback from users.



There’s a significant unbalance in distribution of average rating, number of rating, price.

Some movies are extremely more popular than others.

The price are also not balanced between movies.

Hybrid model for recommender

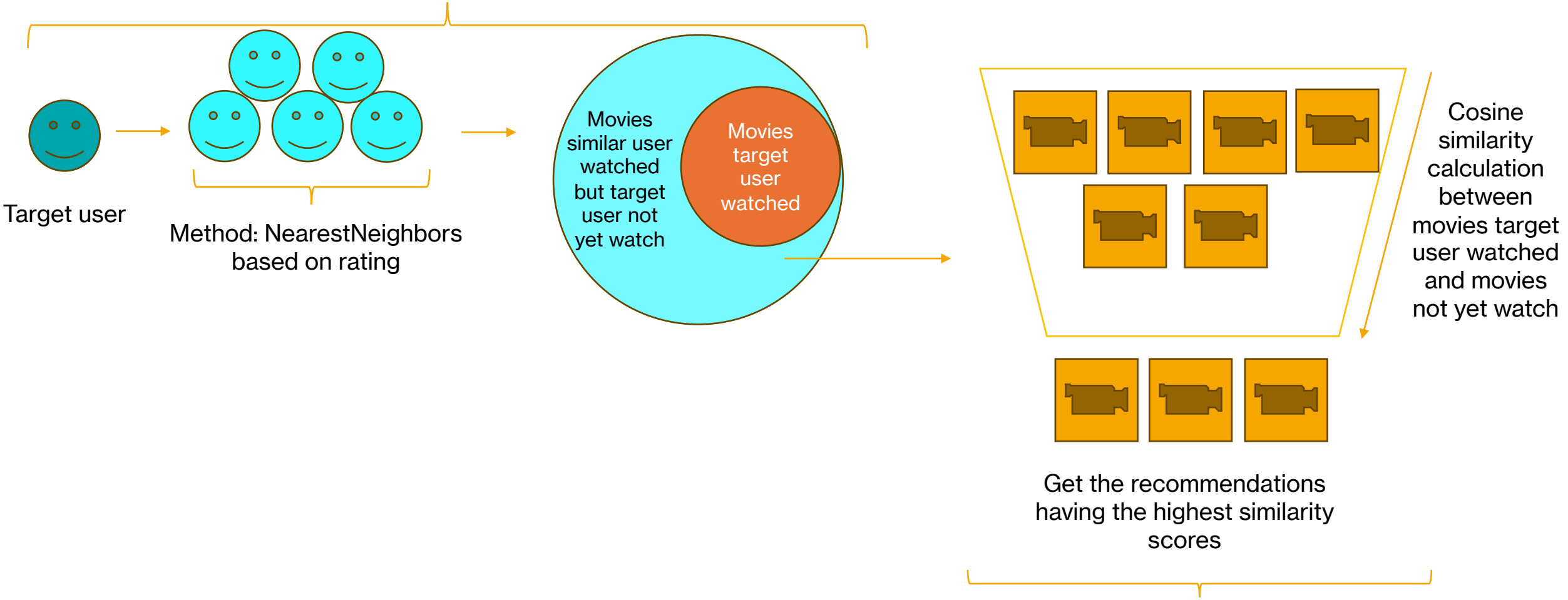
Recommender model

There are two popular methods for recommendation systems

- Collaborative filtering:
  - User-user collaborative filtering
  - Item-Item collaborative filtering
- Content-based filtering

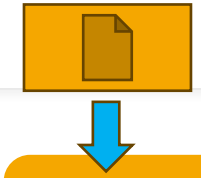
Hybrid model

User-based filtering algorithms



Content-based filtering algorithms

Steps to build the model



Data to feed the model is about users who made more than 30 reviews only.

User-based Collaborative Filtering:

- Input: user_id, dataframe
- Output: Suggested movies having 5star rating of top 5 similar users of ‘user_id’

Content-based filtering model:

- Input: dataframe contains data about content of each movie, ‘parent_asin’_product id
- Output: a dataframe of top similar movies with the ‘parent_asin’ input.

Hybrid model algorithms:

- With each user_id as input, the user-based model will run first to get a dataframe of suggested movies that similar users watched.

-Then, merge that result with a dataframe of watched movies of user_id.

-Next, user cosine-similarity method to filter the list of recommendations from User-based model by keeping the similar movies having the similar score more than 0.8 only.

Model performance:

With user_id = 'AHPRGNDWLTC4EIDDASPKFELSLZSQ' (randomly pick)

The User_based model suggested 98 movies:

```
test_1 = user_based_recommendations(df_narrow.iloc[6,3],df_narrow)
print('User_based model recommend: ',len(test_1),' movies')
test_1.head(3)
```

```
User_based model recommend:  98  movies
```

	rating	review_title	parent_asin	user_id	clean_review	main_
27322	5.0	Awesome movie!!	B0001I55SI	AENK4HJLBS5C4Y7FWLEU7JSS74BQ	saw movie tv child scene banshee death coach s...	
37599	5.0	three hours of beauty	B00AEFYSEA	AGSIXL4DPJMDIWMDTJYPT2MBB6TA	master director sergio leone delivers yet anot...	
39260	5.0	Classic Tarantino	B005LAIIJY	AHS5ZC5IVEBFQTFMDC44XW4QDIWQ	mr quentin one way worked christoph waltz osca...	

The Hybrid model help to narrow down the list to 85 movies by selecting movies with cosine_similarity score above 0.9.

```
#testing:
test_2=hybrid_model(df_narrow.iloc[6,3],df_narrow)
print('Hybrid model recommend: ',len(test_2),' movies')
test_2.head(3)
```

```
Hybrid model recommend:  85  movies
```

	parent_asin	user_id	product_name	rating	average_rating	rating_
1	B001CO42J8	AHCN6VJ6PAZFH2S3CIK554GOBYUQ	A Charlie Brown Christmas (Remastered Deluxe E...	5.0	4.8	
2	B00000G02H	AGSIXL4DPJMDIWMDTJYPT2MBB6TA	Punch-Drunk Love (Two-Disc Special Edition)	5.0	4.5	
6	B00AEFYSEA	AGSIXL4DPJMDIWMDTJYPT2MBB6TA	Once Upon A Time In The West	5.0	4.7	

Another example

```
test_3 = user_based_recommendations(df_narrow.iloc[100000,3],df_narrow)
print('User_based model recommend: ',len(test_3),' movies')
test_3.head(3)
```

```
User_based model recommend:  154  movies
```

	rating	review_title	parent_asin	user_id	clean_review	main_cate
748	5.0	Awesome	B002ZG981E	AGSVTH7RCAPXZAOSK23A4ZLBIZPA	visually entertaining story great	
3056	5.0	hunger games	B0189HKELU	AHZAJYSL7ZS65MB7XTXUWYT2MJ3Q	loved book loved movie	
13227	5.0	dr strange	B01M5EKXCA	AHZAJYSL7ZS65MB7XTXUWYT2MJ3Q	thing marvel	

```
#testing:
test_4=hybrid_model(df_narrow.iloc[100000,3],df_narrow)
print('Hybrid model recommend: ',len(test_4),' movies')
test_4.head(3)
```

```
Hybrid model recommend:  144  movies
```

	parent_asin	user_id	product_name	rating	average_r
7	B0189HKELU	AHZAJYSL7ZS65MB7XTXUWYT2MJ3Q	The Hunger Games: Complete 4 Film Collection	5.0	
8	B01M5EKXCA	AHZAJYSL7ZS65MB7XTXUWYT2MJ3Q	Doctor Strange	5.0	
9	B00OMC0W9G	AELGZ73C76HZ3TALZMNTHZJYE47Q	Fury [Blu-ray]	5.0	

Lesson learned

- Recommender system is really a challenging task.
- The heaviest part of a recommendation model is the algorithms to find recommendations.
- Model evaluation in this step is how effective the hybrid model can help to filter suggestions comparing to user_based model.


Thank you