MOVIE RECOMMENDER SYSTEMS

FROM FIRST VISIT, EXPLORING DIFFERENT SYSTEMS OF RECOMMENDATIONS TO SUIT DIFFERENT BUSINESS NEEDS

EVE BEN EZRA

WHAT IS A RECOMMENDER SYSTEM?

- A Recommender System refers to any customization of items that 'recommends' other items. They can be personalized or universal, and their applications are endless.

- Recommender systems are widely applicable, from Netflix recommending what movie you should watch next to the New York Times recommending popular articles, to voting systems on sites like Reddit and Hacker News

- Generally recommender systems are content-based or collaborative (based on preferences) and lie on a spectrum of personalization. Recommenders can also be popularity-based, or a hybrid of these features.

- A variety of data can be used. Examples: movie metadata, how many minutes of a YouTube video you watch, a 'thumbs-up' or 'thumbs-down' on Pandora, a 'heart' on Spotify, a click on a NYT article, or engagement with an Instagram Ad.

THE IDEA

- A new app works similarly to the iTunes store, recommending movies and outsourcing to buy or rent the movie, or indicating where the movie can be streamed.

- When you sign up, there will be 0 data on you, the user, so the recommender will ask you to put in one of your favorite movies and return suggestions based on that using NLP off movie metadata (plot, rating, writer, director, actors, release year, language, and country). There will also be Universally-available popularity charts.

- Internal data (clicking on movies) and external data (the user clicking) is gathered, and a secondary recommender system works to provide more and more personalized results through collaborative recommending and hybrid systems

1ST RECOMMENDER: NON-PERSONALIZED POPULARITY

- A simple filter that considers movies whose total votes are in the 95th percentile and whose scores are above average.

- Completely non-personalized, every user sees the same list. This is a 'top 250 movies' list

- Based on the idea that 'Hey, if 1 million people like it, you probably will too!'

- Popularity recommenders are used on websites like the New York Time's 'What's Popular', IMDB's top 250 movies list, Reddit's r/all, and more.

- This simple recommender uses all of the data in my scraped dataset, and has an initial shape of (265417, 19)

- Notebook and code: http://bit.ly/eb-poprec

1ST RECOMMENDER: NON-PERSONALIZED POPULARITY

- Make a list of 'qualified' movies, using the total number of IMDB ratings and the average movie ratings.

- The fewer total ratings a movie has, the more an individual rating can influence the overall rating. To prevent massive swinging in rankings, a weighted average is used.

- Process: if a movie has beyond a certain number of votes x, its rating won't be impacted by the transformation. For a movie below that threshold, we will add in a number of votes to bring the total rating up to x, where each of those votes is the mean rating for all movies. This helps the stability of the chart. In this case:

$$(WR) = (v \div (v+m)) \times R + (m \div (v+m)) \times C$$

Where:
R = average rating R for the movie
v = the number of votes cast for the movie
m = the minimum vote threshold required
C = the mean rating of **all** movies
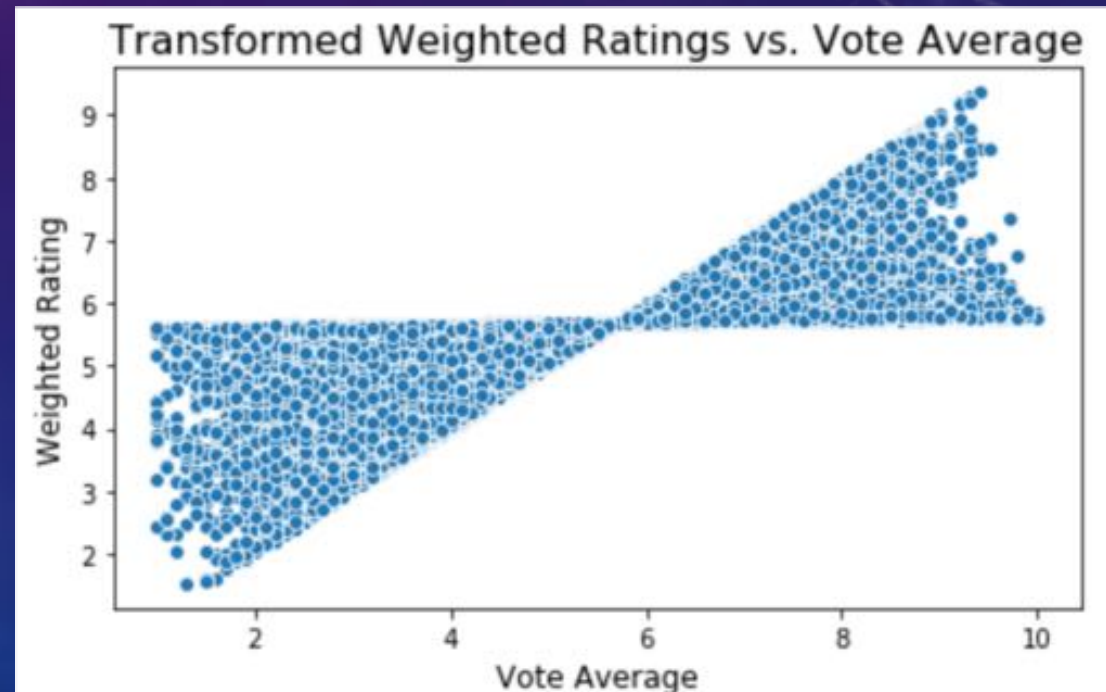
1ST RECOMMENDER: NON-PERSONALIZED POPULARITY

Movies with most votes

| title	imdb_votes	imdb_rating	weighted_rating
The Shawshank Redemption	2084570.0	9.3	9.299422
The Dark Knight	2056482.0	9.0	8.999463
Inception	1826846.0	8.8	8.799432
Fight Club	1665349.0	8.8	8.799377
Pulp Fiction	1632190.0	8.9	8.899344
Forrest Gump	1598520.0	8.8	8.799351
The Lord of the Rings: The Fellowship of the Ring	1503756.0	8.8	8.799310
The Matrix	1500156.0	8.7	8.699330
The Lord of the Rings: The Return of the King	1477575.0	8.9	8.899275
The Godfather	1425400.0	9.2	9.199178

Movies with least votes

title	imdb_votes	imdb_rating	weighted_rating
Cradlesong	5.0	2.8	5.640999
The Most Gigantic Lying Mouth of All Time	5.0	7.6	5.725705
Good Morning	5.0	6.4	5.704529
Burning Calcutta	5.0	8.4	5.739823
A Miracle from Heaven	5.0	5.8	5.693940
Fu shi	5.0	5.6	5.690411
Fondali notturni	5.0	7.2	5.718646
Fine pena mai	5.0	5.2	5.683352
Dou ji	5.0	5.0	5.679823
Boishommo	5.0	6.8	5.711587

Overall, this lowered the global mean from ~7 to around 5.5-6

Transformed Weighted Ratings vs. Vote Average

1ST RECOMMENDER: NON-PERSONALIZED POPULARITY

- The distribution of votes follows a 'long tail' distribution.

- Typically in these distributions, the majority of occurrences are accounted for by the first 20% of items in the distribution.

- Total movies a year vs. big blockbuster hits

- These distributions are important in many recommenders as we consider the entire dataset, and what sort of 'novelty' and 'diversity' we want in our recommendations. A thorough recommender will consider the entire set, not just the Avatars of the movie world.
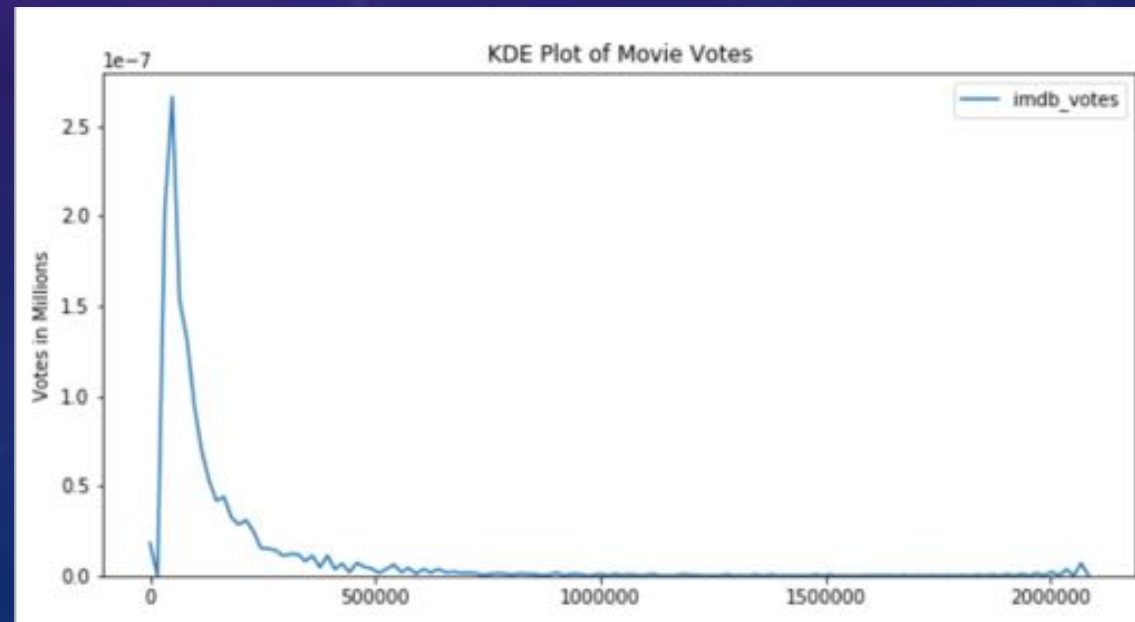
1ST RECOMMENDER: NON-PERSONALIZED POPULARITY

- Our initial list considers all 265,417 movies in our dataset, released from 1891 to 2019.

- Our 'qualified list' contains movies with more than 7691 votes and with an unweighted rating above 6.14

title	year	imdb_votes	imdb_rating	weighted_rating	genre
The Chaos Class	1975	33306	9	9.363186	Comedy, Drama
The Shawshank Redemption	1994	2084570	9	9.299422	Drama
Aynabaji	2016	18209	9	9.235016	Crime, Mystery, Thriller
Wheels	2014	17281	9	9.231593	Drama
The Godfather	1972	1425400	9	9.199178	Crime, Drama
CM101MMXI Fundamentals	2013	41371	9	9.171907	Documentary, Comedy
The Dark Knight	2008	2056482	9	8.999463	Action, Crime, Drama, Thriller
The Godfather: Part II	1974	996807	9	8.998892	Crime, Drama
The Chaos Class Failed the Class	1976	19329	9	8.943813	Comedy
Tosun Pasa	1976	18896	9	8.942548	Comedy, History
The Foster Brothers	1976	15530	9	8.930359	Comedy, Family, Mystery
The Marathon Family	1982	12667	9	8.915024	Comedy, Drama
Pulp Fiction	1994	1632190	8	8.899344	Crime, Drama
The Lord of the Rings: The Return of the King	2003	1477575	8	8.899275	Adventure, Drama, Fantasy
Schindler's List	1993	1077467	8	8.899006	Biography, Drama, History
The Good, the Bad and the Ugly	1966	618548	8	8.898269	Western
12 Angry Men	1957	588758	8	8.898181	Crime, Drama
Avengers: Endgame	2019	360133	8	8.897028	Action, Adventure, Fantasy, Sci-Fi
The Chaos Class Is Waking Up	1977	16016	8	8.834471	Comedy
Saban, Son of Saban	1977	12823	8	8.818570	Comedy, Crime, Mystery

1ST RECOMMENDER: NON-PERSONALIZED POPULARITY

- There are many different ways to make popular charts.

- The previous chart showed all movies available. Next we look at movies worldwide from 1990 or later.

- This time, the 95th percentile for ratings is 11,707 votes, and the average score is 6.13. Only movies that meet both these requirements will be considered.

- Our results still contain foreign films, though we see many more US blockbusters

title	year	imdb_votes	imdb_rating	weighted_rating
The Shawshank Redemption	1994	2084570.0	9.3	9.299422
Aynabaji	2016	18209.0	9.3	9.235016
Wheels	2014	17281.0	9.3	9.231593
CM101MMXI Fundamentals	2013	41371.0	9.2	9.171907
The Dark Knight	2008	2056482.0	9.0	8.999463
Pulp Fiction	1994	1632190.0	8.9	8.899344
The Lord of the Rings: The Return of the King	2003	1477575.0	8.9	8.899275
Schindler's List	1993	1077467.0	8.9	8.899006
Avengers: Endgame	2019	360133.0	8.9	8.897028
Inception	2010	1826846.0	8.8	8.799432
Fight Club	1999	1665349.0	8.8	8.799377
Forrest Gump	1994	1598520.0	8.8	8.799351
The Lord of the Rings: The Fellowship of the Ring	2001	1503756.0	8.8	8.799310
Drishyam	2013	23810.0	8.8	8.757007
Anbe Sivam	2003	12670.0	8.8	8.720180
The Matrix	1999	1500156.0	8.7	8.699330
The Lord of the Rings: The Two Towers	2002	1343746.0	8.7	8.699253
Goodfellas	1990	897033.0	8.7	8.698880
Earthlings	2005	16063.0	8.7	8.638733
Pretty Village, Pretty Flame	1996	13185.0	8.7	8.625691

1ST RECOMMENDER: NON-PERSONALIZED POPULARITY

- Lastly we make a theoretical 'US-Centric' popularity recommendation. We filter out foreign films like Bollywood hits, leaving only Hollywood's greatest movies, as decided by users.

- This list was generated using movies in The Movie Dataset, which is Hollywood-centric and has a base set of 45,466 movies.

- Our cutoffs for this set were 434 votes and a rating of 5.62

- Depending on our cutoffs, movies that appear on some lists would be filtered out and not qualify for others

title	vote_average	vote_count	weighted_rating
The Shawshank Redemption	8.5	8358.0	8.486814
The Godfather	8.5	6024.0	8.481734
Your Name.	8.5	1030.0	8.396081
The Dark Knight	8.3	12269.0	8.291558
Fight Club	8.3	9678.0	8.289306
Pulp Fiction	8.3	8670.0	8.288067
Schindler's List	8.3	4436.0	8.276767
Whiplash	8.3	4376.0	8.276451
Spirited Away	8.3	3968.0	8.274051
Life Is Beautiful	8.3	3643.0	8.271758
The Godfather: Part II	8.3	3418.0	8.269918
One Flew Over the Cuckoo's Nest	8.3	3001.0	8.265786
Psycho	8.3	2405.0	8.257429
Once Upon a Time in America	8.3	1104.0	8.208803
Forrest Gump	8.2	8147.0	8.187720
The Empire Strikes Back	8.2	5998.0	8.183346
The Intouchables	8.2	5410.0	8.181548
Leon: The Professional	8.2	4293.0	8.176785
The Green Mile	8.2	4166.0	8.176083
GoodFellas	8.2	3211.0	8.169047

1ST RECOMMENDER: NON-PERSONALIZED POPULARITY

- We can also use this method to make top-n charts of different movie genres, languages, newness, or other features

- Slightly more personalized: filtering out movies the user has rated or watched

Top 10 for Japanese Language

title	year	imdb_votes	imdb_rating	weighted_rating
Seven Samurai	1954	282175	8	8.696444
Harakiri	1962	24235	8	8.659110
Spirited Away	2001	550352	8	8.598236
The Human Condition III: A Soldier's Prayer	1961	3845	8	8.551660
Grave of the Fireflies	1988	193412	8	8.495159
Woman in the Dunes	1964	14753	8	8.437841
The Human Condition I: No Greater Love	1959	5016	8	8.418490
Princess Mononoke	1997	292566	8	8.396912
Your Name.	2016	130781	8	8.393102
The Human Condition II: Road to Eternity	1959	3918	8	8.371629

Top 10 for Comedy

title	vote_count	vote_average	weighted_rating
Dilwale Dulhania Le Jayenge	661	9	8.463024
Forrest Gump	8147	8	7.963363
Back to the Future	6239	8	7.952358
The Intouchables	5410	8	7.945207
The Grand Budapest Hotel	4644	8	7.936384
Life Is Beautiful	3643	8	7.919430
Dr. Strangelove or: How I Learned to Stop Worr...	1472	8	7.809073
Modern Times	881	8	7.695554
Some Like It Hot	835	8	7.680781
The Great Dictator	756	8	7.651762

1ST RECOMMENDER: NON-PERSONALIZED POPULARITY

Implementation

- Popularity charts are a good starting point, but they lack personalization.

- A good starting point, but due to the long-tail effect, but due to the long-tail effect, our hardcore movie buffs will be disappointed by the lack of novelty and diversity.

- People who don't like high fantasy will not like Lord of the Rings, no matter how popular it is.

Evaluation

- A/B testing, taking initial data and tuning by looking at increased activity with charts, increased movies rented/sold, or other metrics.

- Potential quantitative evaluation by looking at other features, such as total revenue, and comparing with movies selected as 'popular'. The most popular movies will be the ones that make the most money.

2ND RECOMMENDER: CONTENT-BASED

- Recommends based on metadata (plot, genres, actors, director, etc)
- Non-personalized: each user gets back the same results based on search queries and filtering.
- Based on mathematical vector similarity
- Notebook: http://bit.ly/bow-cossim

2ND RECOMMENDER: CONTENT-BASED

- Recommends based on metadata (plot, genres, actors, director, etc)
- Non-personalized: each user gets back the same results based on search queries and filtering.
- Based on mathematical vector similarity
- Notebook: http://bit.ly/bow-cossim
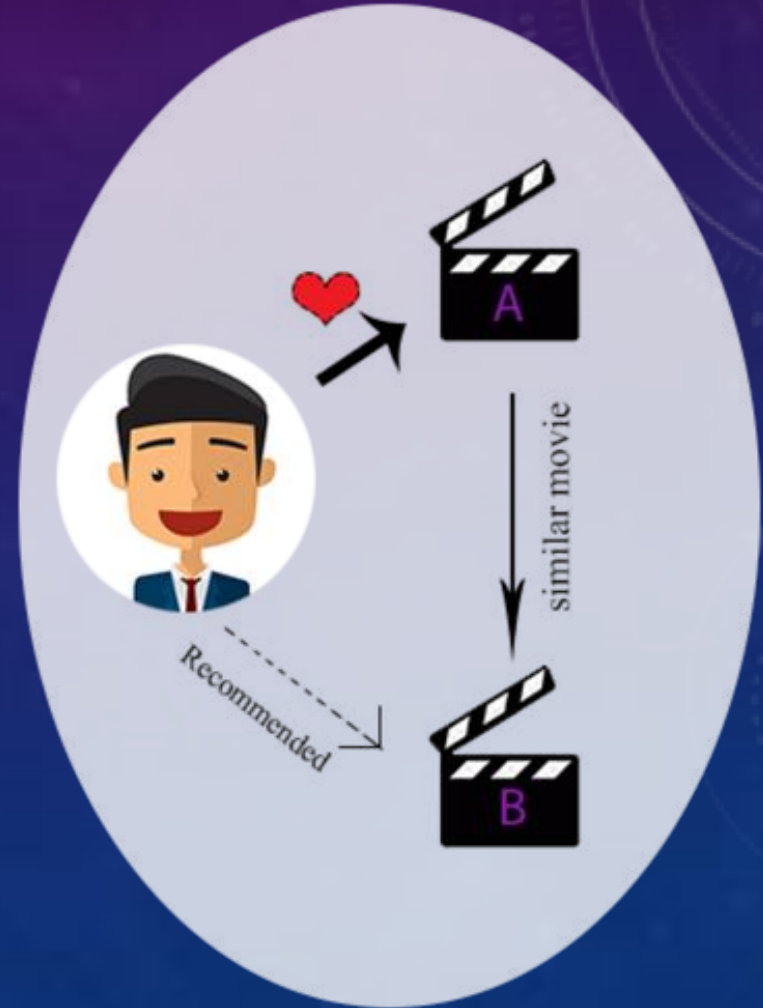
2ND RECOMMENDER: CONTENT-BASED
PREPROCESSING THE TEXT

- Manually tokenizing the text, and using nltk-rake as a keyword extractor for plots

- For the names of actors, directors, and writers, put the first and last names together (John Lasseter becomes johnlasseter)

- Features included: MPAA rating, writer, genres, director, actors, language, country, year, and key words extracted from plot

- Scale: filtered movies down to ~55,000 to make computation manageable: a 55,000x55,000 matrix has just over 3 billion entries. A 100,000-movie matrix has 10 billion.

- Filtered movies by date (included after 1900 only), popularity (41th percentile and above in terms of total votes), and rating (unweighted rating in the 31st percentile or above)

title_year	rated	writer	genre	director	actors	language	country	year	Key_words
Toy Story (1995)	G	[johnlasseter]	[animation, adventure, comedy, family, fan...	[johnlasseter]	[tomhanks, timallen, donrickles, jimvarney]	[english]	[usa]	[1995]	[profoundly, threatened, top, toy, cowboy, dol...
Jumanji (1995)	PG	[jonathanhensleigh]	[adventure, comedy, family, fantasy]	[joejohnston]	[robinwilliams, jonathanhyde, kirstendunst, br...	[english, french]	[usa]	[1995]	[host, magical, board, game, man, trapped, dan...

title_year	bag_of_words
Toy Story (1995)	G johnlasseter animation adventure comedy f...
Jumanji (1995)	PG jonathanhensleigh adventure comedy family...

2ND RECOMMENDER: CONTENT-BASED
COSINE SIMILARITY

- Using CountVectorizer(), we tokenize our bag of words and create count vectors for each movie

- This creates a huge sparse matrix

- We compute cosine similarity for all the count vectors, and are returned an n x n matrix of cosine similarities

- We return the movies with the highest cosine similarities, sorted by their weighted IMDB rating

Similar scores
Score Vectors in same direction
Angle between then is near 0 deg.
Cosine of angle is near 1 i.e. 100%

Unrelated scores
Score Vectors are nearly orthogonal
Angle between then is near 90 deg.
Cosine of angle is near 0 i.e. 0%

Opposite scores
Score Vectors in opposite direction
Angle between then is near 180 deg.
Cosine of angle is near -1 i.e. -100%
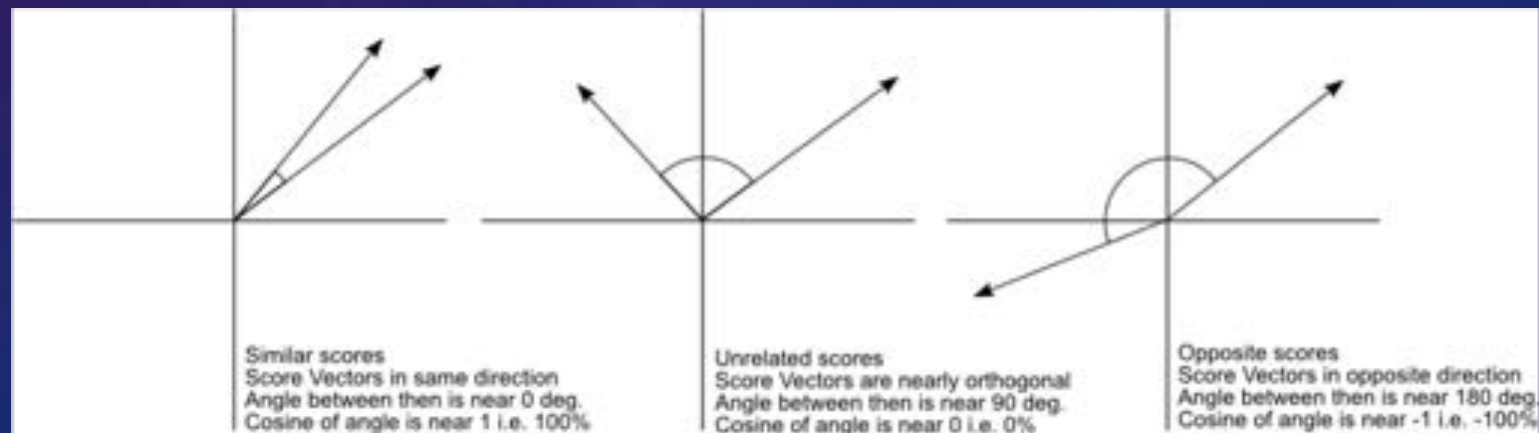
2ND RECOMMENDER: CONTENT-BASED
COSINE SIMILARITY

- The recommendations for some movies make more sense than for others, but throughout all of them, the algorithm managed to capture the themes that run through movies

- Movies returned for LoTR all have common thematic elements such as: power attracting the corruptible, inevitable loss, emotional struggle in the face of sacrifice, and how its worth it anyway

2ᴺᴰ RECOMMENDER: CONTENT-BASED
COSINE SIMILARITY

- The conundrum of similarity between cultures, languages, and movies

- Thematic elements: strong female leads, moving to a new place, encountering incredible creatures and monsters, coming of age, becoming less selfish

Here are the movie recommendations for Spirited Away (2001):

title_year	rated	plot	year	genre	director	actors	language	weighted_rating
Howl's Moving Castle (2004)	PG	When an unconfident young woman is cursed with...	2004	Animation, Adventure, Family, Fantasy	Hayao Miyazaki	Chieko Baishô, Takuya Kimura, Akihiro Miwa, Ta...	Japanese	8.196526
My Neighbor Totoro (1988)	G	When two girls move to the country to be near ...	1988	Animation, Family, Fantasy	Hayao Miyazaki	Noriko Hidaka, Chika Sakamoto, Shigesato Itoi,...	Japanese	8.196011
Castle in the Sky (1986)	PG	A young boy and a girl with a magic crystal mu...	1986	Animation, Adventure, Drama, Fantasy, Sci-Fi	Hayao Miyazaki	Anna Paquin, James Van Der Beek, Cloris Leachm...	Japanese	8.092852
Porco Rosso (1992)	PG	In 1930s Italy, a veteran World War I pilot is...	1992	Animation, Adventure, Comedy, Fantasy, Romance	Hayao Miyazaki	Shûichirô Moriyama, Tokiko Katô, Bunshi Katsur...	Japanese	7.786986
Ponyo (2008)	G	A five-year-old boy develops a relationship wi...	2008	Animation, Adventure, Comedy, Family, Fantasy	Hayao Miyazaki	Tomoko Yamaguchi, Kazushige Nagashima, Yûki Am...	Japanese	7.692918
Harry Potter and the Sorcerer's Stone (2001)	PG	An orphaned boy enrolls in a school of wizardr...	2001	Adventure, Family, Fantasy	Chris Columbus	Richard Harris, Maggie Smith, Robbie Coltrane,...	English	7.598725
Mei and the Kitten Bus (2002)	NaN	This is the story of Mei and her friend the ki...	2002	Animation, Short, Comedy	Hayao Miyazaki	Hayao Miyazaki, Chika Sakamoto	Japanese	7.492856
Pom Poko (1994)	PG	A community of magical shape-shifting raccoon ...	1994	Animation, Comedy, Drama, Family, Fantasy	Isao Takahata	Shinchô Kokontei, Makoto Nonomura, Yuriko Ishi...	Japanese, English	7.367225
The Cat Returns (2002)	G	After helping a cat, a 17-year-old girl finds ...	2002	Animation, Adventure, Comedy, Family, Fantasy,...	Hiroyuki Morita	Chizuru Ikewaki, Yoshihiko Hakamada, Aki Maeda...	Japanese, English	7.284617
My Neighbors the Yamadas (1999)	PG	The life and misadventures of a family in cont...	1999	Animation, Comedy, Family	Isao Takahata	Yukiji Asaoka, Tôru Masuoka, Masako Araki, Hay...	Japanese	7.233928

2ND RECOMMENDER: CONTENT-BASED
COSINE SIMILARITY

- TfidfVectorizer() and pairwise_kernel for different calculations

- Recommendations similar, but different

- Issues: rating similarity, smaller sets (~24,000 movies) get seemingly random recommendations

title_year	
Tangled Ever After (2012)	G
The Secret Life of Pets (2016.0)	PG
Wreck-It Ralph (2012)	PG
Bolt (2008)	PG
Tangled (2010)	PG
Moana (2016)	PG
Who Framed Roger Rabbit (1988)	PG
Life in a Day (2011)	PG-13
The Family Fang (2015)	R
Tintin and the Lake of Sharks (1972)	NaN

Results for
Zootopia
(2016),
Rated PG

Unweighted   Weighted

title_year	
The Secret Life of Pets (2016.0)	PG
Wreck-It Ralph (2012)	PG
Tangled Ever After (2012)	G
Bolt (2008)	PG
Tangled (2010)	PG
Moana (2016)	PG
Who Framed Roger Rabbit (1988)	PG
The Pink Panther (2006)	PG
Fantastic Mr. Fox (2009)	PG
Porco Rosso (1992)	PG

Film Ratings

R 9.2%
NOT RATED 8.6%
PG-13 2.8%
PG 2.7%
APPROVED 2.4%
UNRATED 1.6%
G 0.8%
NC-17 0.8%
NO INFO 71.2%
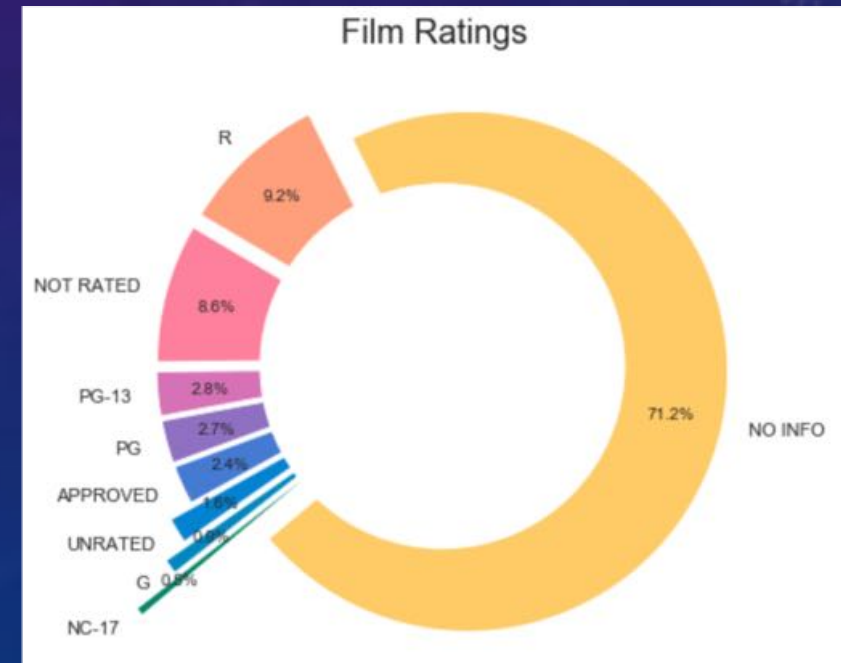
2ᴺᴰ RECOMMENDER: CONTENT-BASED
## COSINE SIMILARITY

Evaluation

- Cosine similarity doesn't scale well, making it difficult to add more movies as it continues to take up computational power and memory (alternative for scale is tf-idf vectorization and pairwise_kernel, which parallelizes the process. Since tf-idf vectors are weighted, the dot product is equivalent to the cosine similarity.

- Evaluation is qualitative, looking at the results, trying different methods and weights, including or taking out features (ex: adding a director's name 3 times)

- User-generated feedback, where users have the ability to add 'genres' and vote on them, giving us more keywords like 'High fantasy' vs fantasy, tuning our results.

Further Tuning

- Possibilities of further filtering and tuning along the lines of the popularity recommender: filtering out certain genres or using them in tandem, returning the top n list of popular movies along with similar movies using cosine similarity.

- Could also introduce other filters for increased variety/diversity: no films with the same director, or only films from a certain year on.

- Allowing users to thumbs-up or thumbs-down on recommendations and considering those weights in the recalculations
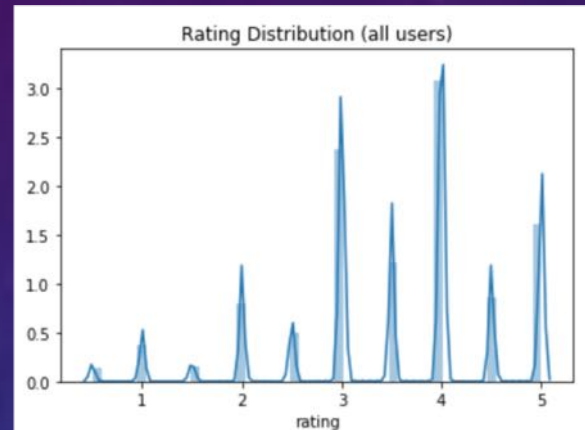
3RD RECOMMENDER: COLLABORATIVE
MATRIX FACTORIZATION

- Factors out a sparse matrix of movies and user ratings into two distinct matrices (item and user)

- Predicts what a user would rate a movie based on hidden features

- Predicted rating is the dot product of these features, or embeddings

- The algorithm uses a method called stochastic gradient descent to minimize the error
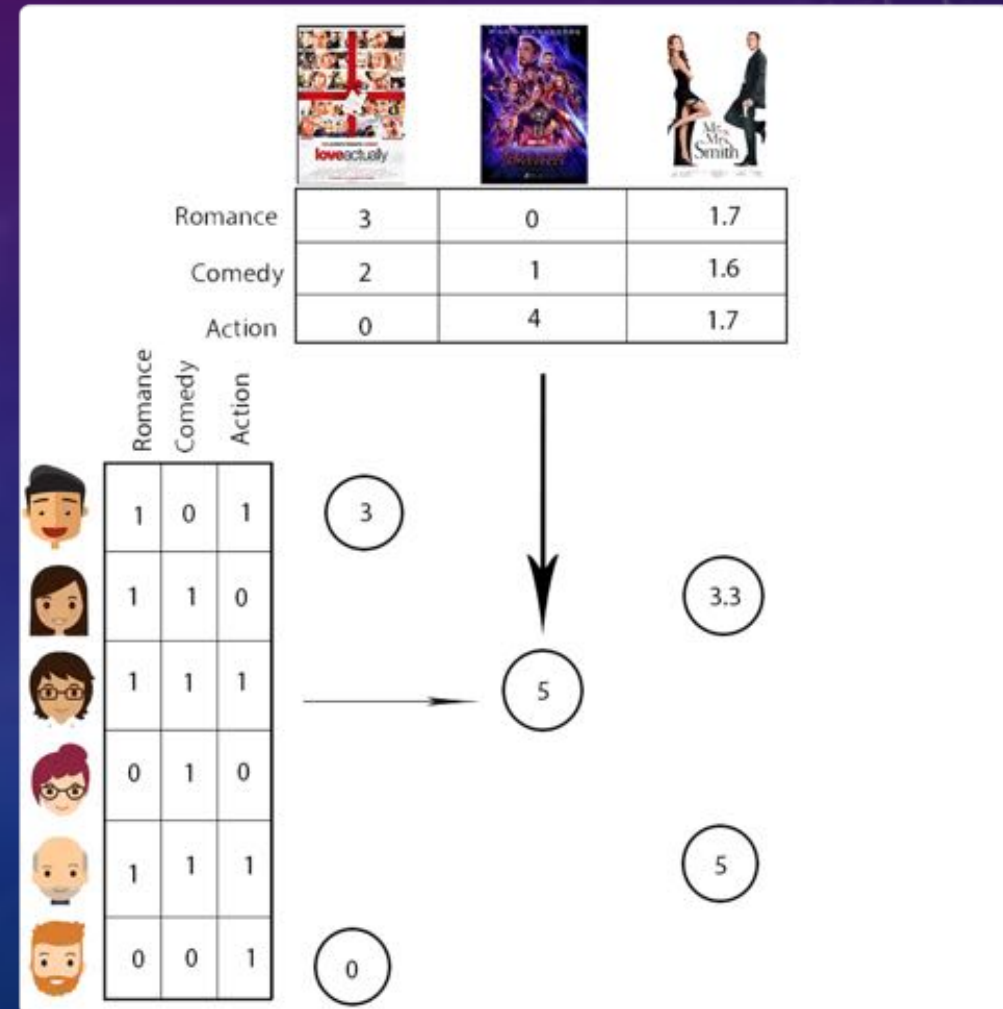
3RD RECOMMENDER: COLLABORATIVE
MATRIX FACTORIZATION

- Datasets used: **The Movie Dataset** (26,024,289 ratings from 120184 users on 4574 movies) and the MovieLens 20M dataset (20,000,263 ratings from 138,493 users on 26,744 movies)

- Algorithms: SVD and SVDpp

- Training and tuning: GridSearchCV, RMSE (stochastic gradient descent)

- Best performing algorithm: SVD, 5-fold CV, 20 epochs, learning rate 0.005, regularization term 0.02

- RMSE of test set: [0.7862, 0.7856, 0.7856, 0.7864, 0.7858]

- Training time: 1.47 hours (88 minutes)

- Final model: http://bit.ly/eb-svd

All ratings

Ratings randomly subsampled
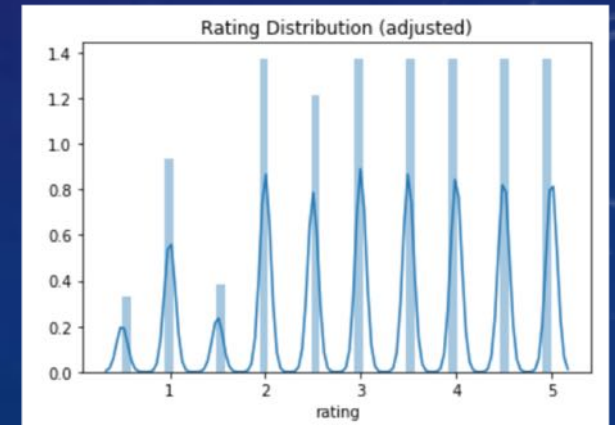
3RD RECOMMENDER: COLLABORATIVE
MATRIX FACTORIZATION

- Class representation subsampling

4.0	5561926
3.0	4291193
5.0	2898660
3.5	2200156
4.5	1534824
2.0	1430997
2.5	883398
1.0	680732
1.5	279252
0.5	239125

5.0	1000000
4.5	1000000
4.0	1000000
3.5	1000000
3.0	1000000
2.0	1000000
2.5	883398
1.0	680732
1.5	279252
0.5	239125

The mean RMSE of the full rating set is: 0.78592

The mean RMSE of the adjusted rating set is: 0.94348

3RD RECOMMENDER: COLLABORATIVE EVALUATION

- Predict against movies a user has already rated
- The number of ratings doesn't seem to impact average error for the user

User 118205 (9254 ratings)

title	rating	est	error	avg_error
Immediate Family (1989)	3.0	3.13	0.13	-0.001189
Executive Decision (1996)	4.0	3.30	-0.70	-0.001189
Body of Evidence (1993)	2.5	2.38	-0.12	-0.001189
Happy Go Lovely (1951)	3.0	2.87	-0.13	-0.001189
Somebody Up There Likes Me (1956)	3.5	3.55	0.05	-0.001189
Ready to Wear (Pret-A-Porter) (1994)	3.5	2.63	-0.87	-0.001189
Scarlet Street (1945)	4.0	3.73	-0.27	-0.001189
Bullet for a Badman (Renegade Posse) (1964)	3.0	3.30	0.30	-0.001189
Son of the Bride (Hijo de la novia, El) (2001)	4.0	3.85	-0.15	-0.001189
I'm Going Home (Je rentre à la maison) (2001)	3.0	3.34	0.34	-0.001189

User 1250 (144 ratings)

title	rating	est	error	avg_error
What's Love Got to Do with It? (1993)	4.0	3.64	-0.36	0.001111
Interview with the Vampire: The Vampire Chroni...	3.0	3.45	0.45	0.001111
Last of the Mohicans, The (1992)	5.0	3.72	-1.28	0.001111
Titanic (1997)	4.0	3.66	-0.34	0.001111
Thousand Acres, A (1997)	2.0	3.13	1.13	0.001111
Braveheart (1995)	5.0	4.38	-0.62	0.001111
Rudy (1993)	3.0	3.63	0.63	0.001111
Emma (1996)	4.0	4.00	0.00	0.001111
Batman Forever (1995)	3.0	3.12	0.12	0.001111
Jumanji (1995)	4.0	3.53	-0.47	0.001111

User 9 (35 ratings)

title	rating	est	error	avg_error
Psycho (1960)	3.0	3.63	0.63	0.011667
Unbreakable (2000)	3.0	3.44	0.44	0.011667
Hannibal (2001)	5.0	3.48	-1.52	0.011667
Dracula 2000 (2000)	2.0	2.70	0.70	0.011667
Forrest Gump (1994)	4.0	3.95	-0.05	0.011667
Little Nicky (2000)	3.0	2.31	-0.69	0.011667
American Pie (1999)	4.0	3.17	-0.83	0.011667
Stigmata (1999)	3.0	3.20	0.20	0.011667
Traffic (2000)	1.0	3.52	2.52	0.011667
What Lies Beneath (2000)	5.0	3.60	-1.40	0.011667

3RD RECOMMENDER: COLLABORATIVE EVALUATION

- Recommend new movies by filtering out movies already watched by the user, calculating the estimated rating on some (or all) of the remaining movies, and return the results sorted by the estimated rating.

- Some movies are highly rated by many users and this might have to do with the average rating of the movie, (the more popular the movie and the more highly rated overall, the more likely to be returned for all users)

User 118205 (9254 Ratings)

title	est
Frozen Planet (2011)	4.35
Civil War, The (1990)	4.28
North & South (2004)	4.27
Godfather, The (1972)	4.23
Death on the Staircase (Soupçons) (2004)	4.19
Alone in the Wilderness (2004)	4.18
Connections (1978)	4.18
Prime Suspect (1991)	4.17
Bill Hicks: Revelations (1993)	4.16
Casablanca (1942)	4.16

User 1250 (144 ratings)

title	est
Moulin Rouge (2001)	4.65
Amelie (Fabuleux destin d'Amélie Poulain, Le) ...	4.64
Life Is Beautiful (La Vita è bella) (1997)	4.63
Boondock Saints, The (2000)	4.62
Frozen Planet (2011)	4.57
Interstellar (2014)	4.55
Inception (2010)	4.53
Bill Hicks: Revelations (1993)	4.53
Black Mirror (2011)	4.52
City of God (Cidade de Deus) (2002)	4.52

User 9 (35 ratings)

| title	est
Frozen Planet (2011)	4.43
Dark Knight, The (2008)	4.34
Inception (2010)	4.29
Dark Knight Rises, The (2012)	4.28
Fight Club (1999)	4.26
Matrix, The (1999)	4.25
Snatch (2000)	4.24
Prime Suspect (1991)	4.24
Louis C.K.: Hilarious (2010)	4.22
Louis C.K.: Live at the Beacon Theater (2011)	4.18

3RD RECOMMENDER: COLLABORATIVE
EVALUATION

- RMSE could probably be further reduced by careful filtering of the initial ratings dataframe

- Taking out users with too many or too few reviews (bots or critics / noise)

- Taking out movies with too many or too few ratings (noise)

- Weighted or dynamic (users encouraged to 'thumbs up' or 'thumbs down' movies based on relevance)

4TH RECOMMENDER: HYBRID
SVD + COSINE SIMILARITY

- The final recommender combines cosine similarity scores and predicted user ratings

- A user enters a movie they liked and are returned a personalized list of movies they haven't yet rated/watched, ordered by their predicted rating

- Notebook: http://bit.ly/eb-hybrid (github) or http://bit.ly/nb-hybrid (nbviewer)
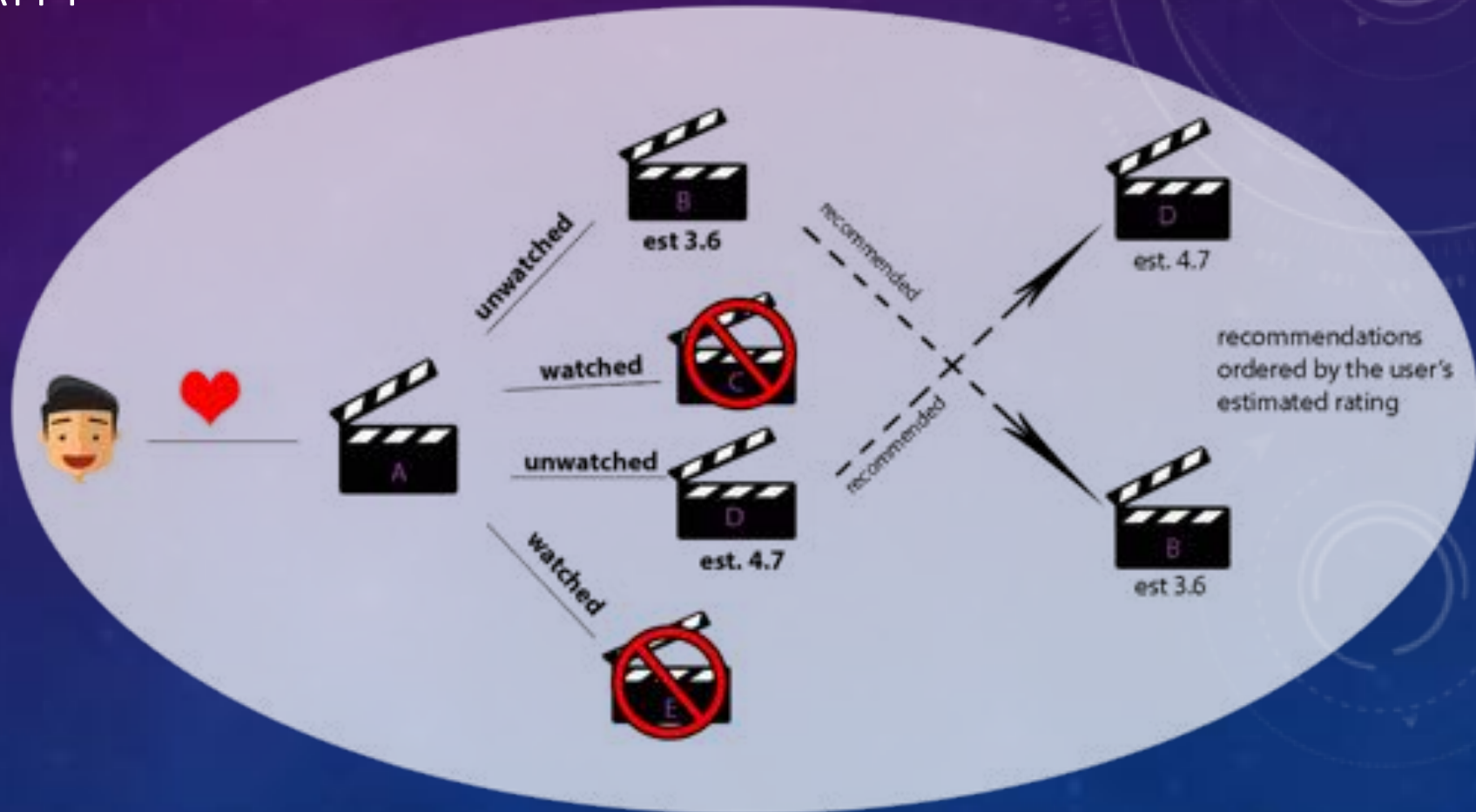
4TH RECOMMENDER: HYBRID

title_year	rated	year	genre	director	plot	language	country	est
My Neighbor Totoro (1988)	G	1988	Animation, Family, Fantasy	Hayao Miyazaki	When two girls move to the country to be near ...	Japanese	Japan	3.95
Song of the Sea (2014)	PG	2014	Animation, Adventure, Drama, Family, Fantasy	Tomm Moore	Ben, a young Irish boy, and his little sister ...	English, Irish	Ireland, Denmark, Belgium, Luxembourg, France	3.84
Howl's Moving Castle (2004)	PG	2004	Animation, Adventure, Family, Fantasy	Hayao Miyazaki	When an unconfident young woman is cursed with...	Japanese	Japan	3.83
Kiki's Delivery Service (1989)	G	1989	Animation, Adventure, Drama, Family, Fantasy	Hayao Miyazaki	A young witch, on her mandatory year of indepe...	Japanese, English	Japan	3.80
Shrek (2001)	PG	2001	Animation, Adventure, Comedy, Family, Fantasy	Andrew Adamson, Vicky Jenson	A mean lord exiles fairytale creatures to the ...	English	USA	3.80
Nausicaä of the Valley of the Wind (1984)	PG	1984	Animation, Adventure, Fantasy, Sci-Fi	Hayao Miyazaki	Warrior and pacifist Princess Nausicaä despera...	Japanese, English	Japan, USA	3.75
The Incredibles (2004)	PG	2004	Animation, Action, Adventure, Family	Brad Bird	A family of undercover superheroes, while tryi...	English, French	USA	3.75
Castle in the Sky (1986)	PG	1986	Animation, Adventure, Drama, Fantasy, Sci-Fi	Hayao Miyazaki	A young boy and a girl with a magic crystal mu...	Japanese	Japan	3.74
The Tale of The Princess Kaguya (2013)	PG	2013	Animation, Adventure, Drama, Family, Fantasy	Isao Takahata	Found inside a shining stalk of bamboo by an o...	Japanese	Japan	3.70
Porco Rosso (1992)	PG	1992	Animation, Adventure, Comedy, Fantasy, Romance	Hayao Miyazaki	In 1930s Italy, a veteran World War I pilot is...	Japanese	Japan	3.67

Previously-rated movies returned

Recommendations for user 118205 for 'Spirited Away'

Only unwatched movies returned

title_year	rated	year
Song of the Sea (2014)	PG	2014
Nausicaä of the Valley of the Wind (1984)	PG	1984
The Tale of The Princess Kaguya (2013)	PG	2013
Porco Rosso (1992)	PG	1992
Millennium Actress (2001)	PG	2001
Ponyo (2008)	G	2008
The Cat Returns (2002)	G	2002
Rise of the Guardians (2012)	PG	2012
My Neighbors the Yamadas (1999)	PG	1999
Pom Poko (1994)	PG	1994

4TH RECOMMENDER: HYBRID

"You haven't seen yet…"

User 118205

title_year		
Song of the Sea (2014)	PG	2014
Nausicaä of the Valley of the Wind (1984)	PG	1984
The Tale of The Princess Kaguya (2013)	PG	2013
Porco Rosso (1992)	PG	1992
Millennium Actress (2001)	PG	2001
Ponyo (2008)	G	2008
The Cat Returns (2002)	G	2002
Rise of the Guardians (2012)	PG	2012
My Neighbors the Yamadas (1999)	PG	1999
Pom Poko (1994)	PG	1994

User 1250

title_year		
Harry Potter and the Sorcerer's Stone (2001)	PG	2001
Shrek (2001)	PG	2001
Harry Potter and the Prisoner of Azkaban (2004)	PG	2004
Harry Potter and the Half-Blood Prince (2009)	PG	2009
Shrek 2 (2004)	PG	2004
Song of the Sea (2014)	PG	2014
Howl's Moving Castle (2004)	PG	2004
The Incredibles (2004)	PG	2004
Tangled (2010)	PG	2010
Kiki's Delivery Service (1989)	G	1989

User 9

title_year		
Howl's Moving Castle (2004)	PG	2004
Harry Potter and the Half-Blood Prince (2009)	PG	2009
Song of the Sea (2014)	PG	2014
Kiki's Delivery Service (1989)	G	1989
Harry Potter and the Prisoner of Azkaban (2004)	PG	2004
My Neighbor Totoro (1988)	G	1988
Castle in the Sky (1986)	PG	1986
Nausicaä of the Valley of the Wind (1984)	PG	1984
Shrek (2001)	PG	2001
Millennium Actress (2001)	PG	2001

RECAP

Users can now use the app in a variety of ways, with various levels of engagement:

- Users who don't want to sign up for an account can view popularity charts and sort/filter them by categories (language, genre, actor, etc)

- Users who want a bit more depth and variety (outside of top charts) can use content filters to get similar, less-known movies. These results can be filtered by their weighted rating (bringing a popularity aspect to it), or simply returned from highest to lowest similarity

- Users who want to sign up and set preferences can get personalized recommendations or use a 'recommend a random movie' feature. The user's tastes are considered against the entire database and returned in order of the user's estimated rating.

- Users who want personalized recommendations of a movie they've previously liked can filter their results by movies they haven't watched, or include all of them (leading to the opportunity for re-watching movies they've forgotten about)

OTHER CONSIDERATIONS

- Though systems like this are seen on many movie-streaming sites, this site takes into account an entire database of movies and outsources the streaming/buying/downloading, acting as a go-between for users who want movie recommendations but don't want to be limited by a certain site's catalogue.

- Movies can be automatically recommended for users based on movies they've already rated highly.

- Tastes change! The user's past ratings would have to eventually expire or be weighted to be less impactful than recent ratings.

- Exact user scores don't have to be given back to users: instead the layout could be more of a "We think you'd like this!" instead of "We think you'd rate this 4 stars!" while still being ordered by estimated rating internally