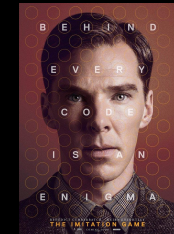


Recommender Systems

NETFLIX

Top Picks for You



Discover Weekly
Your weekly mixtape of fresh music. Enjoy new discoveries and deep cuts chosen just for you. Monday, so save your favourite!

Created by Spotify · 30 songs, 2 hr 5 min

TRACK	ARTIST	TIME	ALBUM	ADDED	USER
+ Bathed in Light	Gengahr	3:13	Powder / Bathed in Li	1 day ago	Spotify
+ I Love You All (Radio Mix) feat.	The Soronprfbs, Mi	3:52	I Love You All (From	1 day ago	Spotify
+ First Light	Django Django	4:49	First Light	1 day ago	Spotify
+ X Marks The Spot	Ghostpoet, Nadine	3:51	Shedding Skin	1 day ago	Spotify
+ All The Time	Bahamas	3:55	Bahamas Is Afie	1 day ago	Spotify
+ Occupy Your Mind	Villagers	2:53	Occupy Your Mind	1 day ago	Spotify
+ Shelter Song	Temples	3:11	Sun Structures	1 day ago	Spotify

Frequently bought together

Total price: **\$179.39**
[Add all three to Cart](#)
[Add all three to List](#)

- ⓘ These items are shipped from and sold by different sellers. Show details
- This Item:** Machine Learning: A Probabilistic Perspective (Adaptive Computation and Machine Learning) by Kevin P. Murphy Hardcover \$68.60
 - Pattern Recognition and Machine Learning (Information Science and Statistics)** by Christopher M. Bishop Hardcover \$60.79
 - Deep Learning (Adaptive Computation and Machine Learning)** by Ian Goodfellow Hardcover \$50.00



Customers who bought this item also bought

 Pattern Recognition and Machine Learning Information Science, ... Christopher M. Bishop ★★★★★ 132 Hardcover \$60.79 -prime	 The Elements of Statistical Learning: Data Mining, Inference, and... Trevor Hastie ★★★★★ 151 Hardcover \$24.67 -prime	 Deep Learning (Adaptive Computation and Machine Learning) Ian Goodfellow ★★★★★ 153 Hardcover \$50.00 -prime	 Probabilistic Graphical Models: Principles and Techniques (Adaptive... Daphne Koller ★★★★★ 45 Hardcover \$83.54	 Learning From Data Year S. Elu-Hosoda ★★★★★ 145 Hardcover \$28.00 -prime
---	---	---	---	--

SUNBEAM Fan Heater, 1 Touch Electric Thermostat,alci Plug
by Sunbeam

Rate this item
 ☆☆☆☆
 This was a gift
 Don't use for recommendations

Legend of the Guardians: The Owls of Ga'Hoole
~ Helen Mirren

Rate this item
 ☆☆☆☆
 This was a gift
 Don't use for recommendations

Mars Chocolate Favorites Halloween Candy Bars Variety Mix
96.2-Ounce 250-Piece Bag
by Mars

Rate this item
 ☆☆☆☆
 This was a gift
 Don't use for recommendations

amazon

How might my ratings change my recommendations?

Goals of Recommender Systems

- Show content that we're interested in
- Suggest new content that would interest us
- Suggest new content that is generally popular
- Adjust recommendations based on our feedback

Netfli Prize

Home Rules Leaderboard Update

COMPLETED

Congratulations!

The Netflix Prize sought to substantially improve the accuracy of predictions about how much someone is going to enjoy a movie based on their movie preferences.

On September 21, 2009 we awarded the \$1M Grand Prize to team "BellKor's Pragmatic Chaos". Read about their [algorithm](#), checkout team scores on the [Leaderboard](#), and join the discussions on the [Forum](#).

We applaud all the contributors to this quest, which improves our ability to connect people to the movies they love.

\$1M winning algorithm not actually used by Netflix

Researchers were able to de-anonymize data by comparing with IMDB ratings, resulting in a lawsuit

YouTube-8M Dataset

Dataset Explore Download Workshop About

YouTube-8M is a large-scale labeled video dataset that consists of millions of YouTube video IDs, with high-quality machine-generated annotations from a diverse vocabulary of 3,800+ visual entities. It comes with precomputed audio-visual features from billions of frames and audio segments, designed to fit on a single hard disk. This makes it possible to train a strong baseline model on this dataset in less than a day on a single GPU. At the same time, the dataset's scale and diversity can enable deep exploration of complex audio-visual models that can take weeks to train even in a distributed fashion.

Our goal is to accelerate research on large-scale video understanding, representation learning, noisy data modeling, transfer learning, and domain adaptation approaches for video. More details about the dataset and initial experiments can be found in our [technical report](#) and in last year's [workshop](#). Some statistics from the latest version of the dataset are included below.

6.1 Million Video IDs	350,000 Hours of Video	2.6 Billion Audio/Visual Features	3862 Classes	3.0 Avg. Labels / Video
--------------------------	---------------------------	--------------------------------------	-----------------	----------------------------

RecSys Challenge 2018

Welcome ACM RecSys Community! For this year's challenge, use the Spotify Million Playlist Dataset to help users create and extend their own playlists.

Read on for all the details. Good luck!

The RecSys Challenge 2018 is organized by Spotify, The University of Massachusetts, Amherst, and Johannes Kepler University, Linz.

Have a question, suggestion or concern? Let us know by emailing us at recsyschallenge@spotify.com

Yelp Dataset Challenge
Discover what insights lie hidden in our data.

What is the dataset challenge?

The challenge is a chance for students to conduct research or analysis on our data and share their discoveries with us. Whether you're trying to figure out how food trends start or identify the impact of different connections from the local graph, you'll have a chance to win cash prizes for your work! See some of the [past winners](#) and [hundreds of academic papers](#) written using the dataset.

Recommender Systems

Collaborative Filtering

- What makes two (Amazon) users similar?
 - Purchased the same set of items
 - Liked and disliked the same set of items
- What makes two items similar?
 - The same set of users purchased/liked them
 - Their titles, description, prices, other metadata

Content Based Recommendation

Collaborative Filtering

Create a *user-item* matrix

Sohie	5		2	
Brian				4
Christine		1	4	5
Orit	3	5	3	
Catherine	4	4		4

Similarity: Jaccard

- Measure similarity between a pair of user vectors (or a pair of item vectors)

$$U_A = [1, 0, 1, 0]$$

$$U_B = [0, 0, 1, 0]$$

$$Jaccard(U_A, U_B) = \frac{|U_A \cap U_B|}{|U_A \cup U_B|}$$

Problem:
does not work
for non-binary
vectors

When is
result 0?
When is it 1?

Similarity: Cosine

- Measure similarity between a pair of user vectors (or a pair of item vectors)

$$U_A = [0, 5, 2, 0]$$

$$U_B = [1, 0, 4, 2]$$

$$CosineSim(U_A, U_B) = \frac{U_A \cdot U_B}{\|U_A\| \|U_B\|}$$

When is
result 0?
When is it 1?

User-Based Collaborative Filtering

Task: predict rating on new user-item entry in matrix: U_A, I_p

- Among *users* that have rated I_p , select a set S_K of the K most similar *users* to U_A
- Predicted rating for U_A, I_p is average rating of I_p from *users* in S_K :

$$R(U_A, I_p) = \frac{\sum_{U_B \in S_K} R(U_B, I_p)}{K}$$

User-Based Collaborative Filtering

Task: predict rating on new user-item entry in matrix: U_A, I_p

- Some users have a tendency to be more or less generous
- Use **deviation from a user's average rating**, rather than a user's absolute rating

$$R(U_A, I_p) = \frac{\sum_{U_B \in S_K} R(U_B, I_p) - \text{mean}(U_B)}{K}$$

Item-Based Collaborative Filtering

Task: predict rating on new user-item entry in matrix: U_A, I_P

- Among *items* that have been rated by U_A , select a set S_K of the K most similar *items* to I_P
- Predicted rating for U_A, I_P is average rating of U_A from *items* in S_K :

$$R(U_A, I_P) = \frac{\sum_{I_Q \in S_K} R(U_A, I_Q)}{K}$$

Item-Based Collaborative Filtering

Task: predict rating on new user-item entry in matrix: U_A, I_P

- Among all *items* that have been rated by U_A , compute **weighted average** of U_A 's ratings (**weighted by similarity to I_P**)

$$R(U_A, I_P) = \frac{\sum_{I_Q \in S_{all}} \text{Sim}(I_P, I_Q) R(U_A, I_Q)}{\sum_{I_Q \in S_{all}} \text{Sim}(I_P, I_Q)}$$

Weighted Average

Mean is 73%

Compute final score in some class:

Weighted Mean is 80%

	<u>Score</u>	<u>Weight</u>
Class participation	60%	50 points
Homework	95%	200 points
Midterm Exam	50%	100 points
Final Exam	87%	150 points

$$\text{Weighted Mean} = \frac{\sum \text{Weight} \cdot \text{Score}}{\sum \text{Weight}} = \frac{50 \cdot 0.60 + 200 \cdot 0.95 + 100 \cdot 0.50 + 150 \cdot 0.87}{50 + 200 + 100 + 150}$$

Problems with Collaborative Filtering?

- If user-item matrix is too sparse, may not be useful
- “Cold-start problem”: how to handle new users and items?
- Won't encourage diverse results (echo chamber effect)

Content-Based Recommendations: Approach 1

- Define similarity between users (or similarity between items) in terms of **content features**, not **rating patterns**
 - Examples of item features: restaurant cuisine type, director or actors in movie, product details
 - Examples of user features: demographic information
- Apply same methods as for collaborative filtering

Content-Based Recommendations: Approach 2

- Featurize users and items under the **same** set of features
 - Features: words
 - user feature values = word counts in reviews
 - item feature values = word counts in descriptions
 - Features: demographics
 - user feature values = demographic info
 - item feature values = target demographics
- Compute similarity **between a given user and item**

Featurizing Text

- Bag of words: tokenizing, counting, tf-idf weighting

	bland	but	fast	food	good	no	parking	service
Fast service but bland food.	1	1	1	1	0	0	0	1
Good fast food.	0	0	1	1	1	0	0	0
No service, no parking, no good.	0	0	0	0	1	3	1	1

	bland	but	fast	food	good	no	parking	service
0.5	0.5	0.4	0.4	0	0	0	0	0.4
0	0	0.6	0.6	0.6	0	0	0	0
0	0	0	0	0.2	0.9	0.3	0.2	

- N-Grams

Fast service but bland food.	fast service	service no	no service	fast food	fast food	good fast
Good fast food.	fast service	service but	but bland	parking no	parking no	
No service, no parking, no good.	service but	no service	no good	no good	no parking	

Evaluation

- Task Type A:** Given test set of (user, item) pairs, predict ratings
 - Raw accuracy, e.g., percentage of ratings predicted exactly Too strict!
 - Root mean squared error (RMSE)

$$RMSE = \sqrt{\frac{1}{n} \sum_{i=1}^n (\hat{y}_i - y_i)^2}$$

- Task Type B:** Given test set of users, predict set of items to recommend

- Precision, Recall, F1 Score

TP: Recommended items user actually buys
FP: Recommended items user does not buy
TN: Items not recommended and user does not buy
FN: Items not recommended and user buys

Vectorization (Array Programming)

- Many scientific and numerical computing libraries, such as NumPy in Python, provide *vectorized* operations, i.e., operations that can be applied to an entire array (matrix)

```
np.random.randint(...)  
np.median(a)          a[a>10]  
a**2                 np.sum(a)   np.mean(a)   np.dot(a,b)  
np.ones(...)
```

- Whenever possible, it is usually a good idea to use *vectorization* rather than looping through an array and applying an operation to each element

Overview

