

Build your first RecSys

Evgeniya Korneva
January 19, 2024
Barcelona



On the menu today



Overview of
recommendation systems



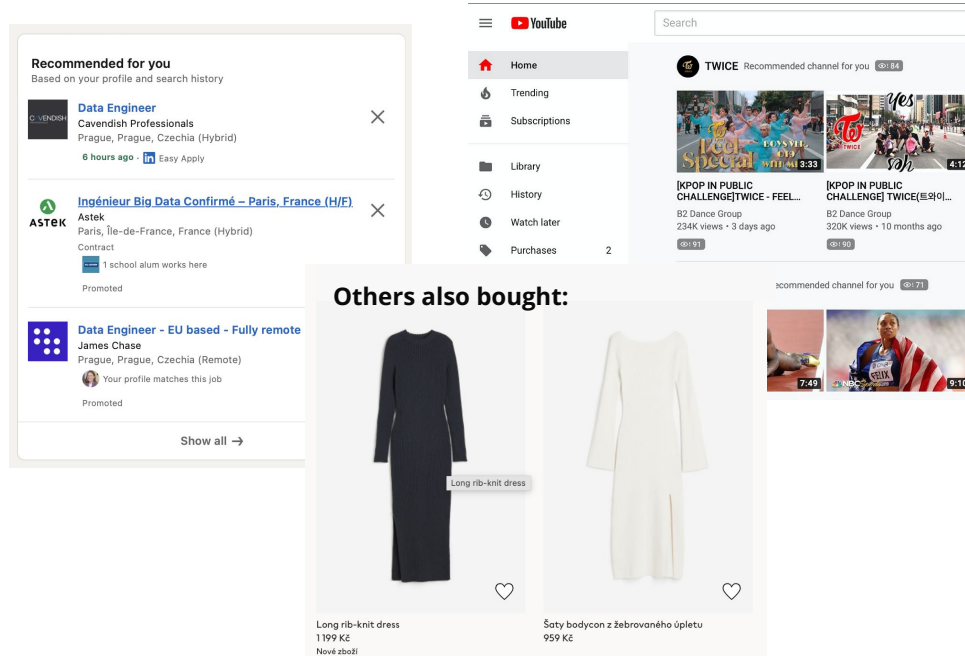
Collaborative filtering
in depth
(simple approach)



Practice time!

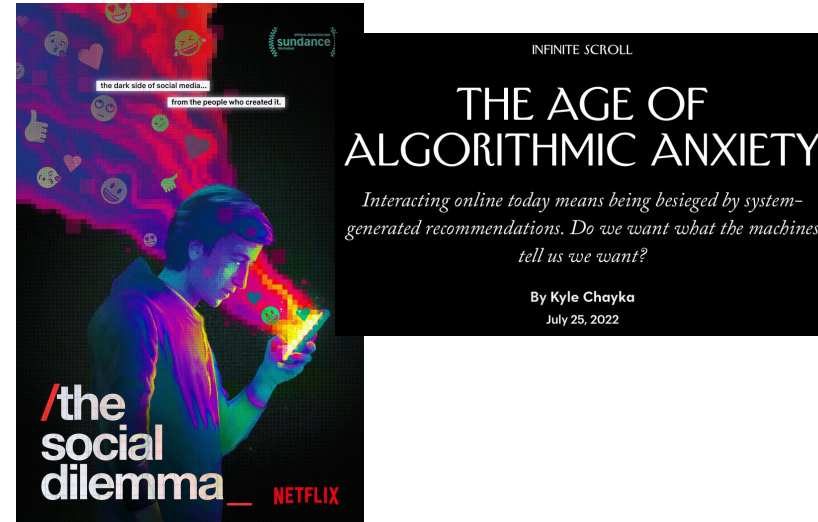
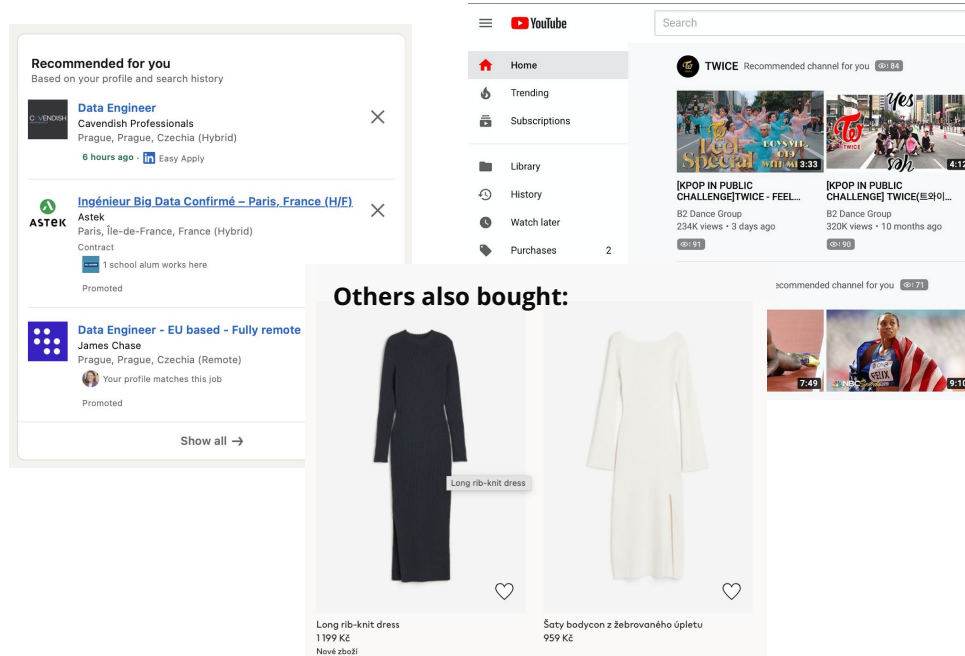
Recommendations are everywhere

Hardly any part of our life isn't affected by personalized recommendations. While RecSys help users find compelling content in a large corpora, some consider personalization a serious concern



Recommendations are everywhere

Hardly any part of our life isn't affected by personalized recommendations. While RecSys help users find compelling content in a large corpora, some consider personalization a serious concern



Netflix Prize boosted interest in recommenders

The competition ran for 2 years, and the winning team won a \$1M prize. However, Netflix never used the winning solution as it was too difficult to implement in production

A screenshot of the Netflix Prize Leaderboard. The header features the Netflix logo and the title "Netflix Prize". Below the header is a navigation bar with links: Home, Rules, Leaderboard, Register, Update, Submit, and Download. The main content area is titled "Leaderboard" and includes a "Display top 20 leaders" dropdown menu. Below the title is a table listing the top teams and their scores. The table has five columns: Rank, Team Name, Best Score, % Improvement, and Last Submit Time. The first two teams are "The Ensemble" and "BellKor's Pragmatic Chaos". A red banner below the table indicates the "Grand Prize - RMSE <= 0.8563". The table lists 12 teams in total.

Rank	Team Name	Best Score	% Improvement	Last Submit Time
1	The Ensemble	0.8553	10.10	2009-07-26 18:38:22
2	BellKor's Pragmatic Chaos	0.8554	10.09	2009-07-26 18:18:28
Grand Prize - RMSE <= 0.8563				
3	Grand Prize Team	0.8571	9.91	2009-07-24 13:07:49
4	Opera Solutions and Vandelav United	0.8573	9.89	2009-07-25 20:05:52
5	Vandelav Industries I	0.8579	9.83	2009-07-26 02:49:53
6	Pragmatic Theory	0.8582	9.80	2009-07-12 15:09:53
7	BellKor in BigChaos	0.8590	9.71	2009-07-26 12:57:25
8	Dance	0.8603	9.58	2009-07-24 17:18:43
9	Opera Solutions	0.8611	9.49	2009-07-26 18:02:08
10	BellKor	0.8612	9.48	2009-07-26 17:19:11
11	BigChaos	0.8613	9.47	2009-06-23 23:06:52
12	Feeds2	0.8613	9.47	2009-07-24 20:06:46

**There are different types
of recommendations**

We need recommendations in different context

Different systems can be used to provide good recommendations in different situations. Not all of them are necessarily personalized or need a complex ML-based solution

Non-personalised

- Substitutions for out-of-stock items
- Cross-sales: items frequently bought together

Personalised

- *"You might also like ..."*
- *"Others also liked ..."*

Look for **substitution** when an item is unavailable



Substitution means finding the most similar item to the one out of stock. Given some vector representations of items, this is a straightforward task, but bad representations can result in unexpected suggestions...

Andreas Hagemann @hagmnn

Can't wait to surprise my wife with an organic red bell pepper!

Whole Foods Market shopper • Just now

One of the items in your Whole Foods Market order is out of stock, please review substitution option(s).




OUT OF STOCK	SUBSTITUTION
 Rose 12 Stems 40Cm Whole Trade Guarantee \$12.99	 Pepper Bell Red Whole Trade Guarantee Organic, 1 Each \$3.99
<input type="button" value="Decline"/>	<input type="button" value="Accept"/>

3:34 PM · Oct 13, 2020

21.2K

Manah Khalil @ManahKhalil

Replying to @hagmnn
I guess at least it's all plants/vegetables. I had to snack on a t-shirt with my Walmart order back in Apr.

3	 Fresh Blueberries, 18 oz \$1.00/ea \$3.00
Unavailable	
1	 Blood Oranges, 2 lb Bag \$3.32/ea \$3.32
Substitution	
1	 Workout T-Shirt Hates The Gym Womens Curvy Plus Size V-Neck Tee \$3.32/ea \$3.32

2:23 PM · Oct 15, 2020



10

Leigh Senderowicz @DrLeighS

Replying to @hagmnn
I got this one once

Just now

One of your items is out of stock. I found this replacement.

OUT OF STOCK	REPLACEMENT
 Apex Medical, Thermometer Fast Read \$9.99	 365 Everyday Value, Macaroni And Cheese Family Size Organic, 12 Ounce \$2.99
<input type="button" value="Decline"/>	<input type="button" value="Accept"/>

1:45 AM · Oct 14, 2020

349

Full story: <https://www.boredpanda.com/hilarious-amazon-out-of-stock-substitution-options/>

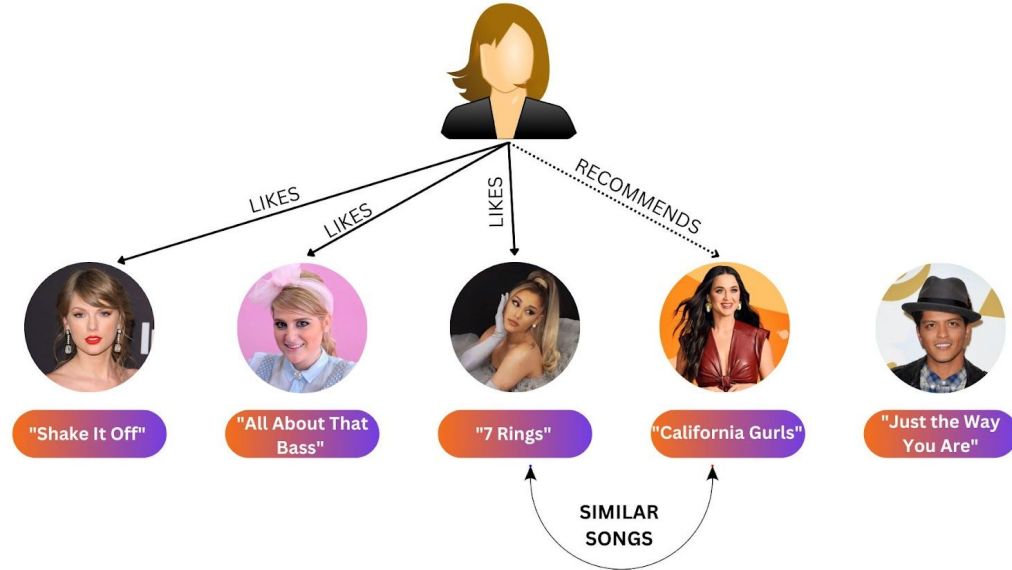
Cross-selling helps maximize basket size

We can suggest items that are frequently bought together with those that the user already added to the basket. Association rule mining techniques help identify such patterns from raw transactional data



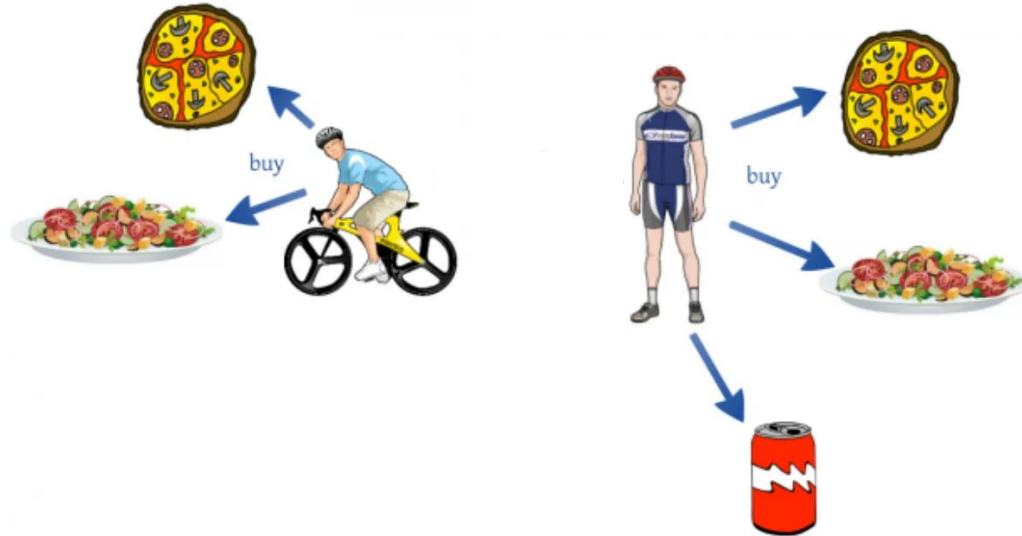
Content filtering brings more similar content

We can suggest items that are similar to those the user has already liked. The only information we need for that is some vector representation of the items



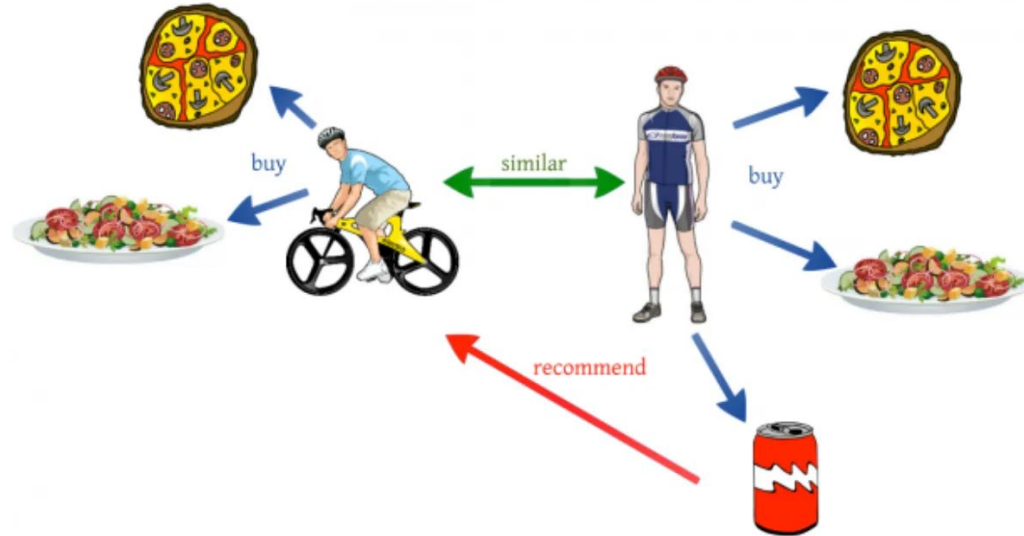
Collaborative filtering looks at the others, too

We can assume that users who liked similar items in the past will continue to have similar tastes in the future. Thus, when generating recommendations for a given user, we can use similar users as inspiration



Collaborative filtering looks at the others, too

We can assume that users who liked similar items in the past will continue to have similar tastes in the future. Thus, when generating recommendations for a given user, we can use similar users as inspiration



Both approaches have their (dis)advantages

In practice, hybrid approaches are often used to combine the two worlds. Often, candidates are generated by different recommendation engines and are then re-ranked according to additional criteria

Content-based filtering

Collaborative filtering

Both approaches have their (dis)advantages

In practice, hybrid approaches are often used to combine the two worlds. Often, candidates are generated by different recommendation engines and are then re-ranked according to additional criteria

Content-based filtering

- ✓ Captures specific interests, can recommend niche items only few people are interested in.
- ✓ Recommendations are user-specific, no info about the other users needed.
- ✗ Only as good as the features representing the items
- ✗ Can't handle new users
- ✗ Puts the user in their own bubble

Collaborative filtering

- ✓ No item- or user features needed, only user-item interactions
- ✓ Can help users discover new interests
- ✗ Can't handle neither fresh items nor new users
- ✗ Hard to include additional information about the users and/or items

Diving deeper into collaborative filtering

A very basic implementation




Users with similar taste will agree in the future

So, when predicting how much a given user will like a particular movie, we need to look at how much other users liked it, and how similar taste they have to the user in question

					
JOHN	1		1	2	2
LUCY	1	2	5	5	5
DIANE	4	5	3	3	
YOU	2	3	?	5	4

We need to measure similarity between users

This can be done by comparing how the two users rated a subset of movies they both saw. The more similar those vectors are, the more similar are the users' preferences

					
JOHN	1		1	2	2
LUCY	1	2	5	5	5
DIANE	4	5	3	3	
YOU	2	3	?	5	4

Correlation is one possible similarity measure

Correlation between the ratings, computed a subset of items rated by both users, shows how aligned the users are in their tastes. It ranges from +1 (very similar) to -1 (opposite) tastes.

					
JOHN	1		1	2	2
YOU	2	3	?	5	4

$$W_{ij} = \frac{\sum_k (R_{ik} - \bar{R}_i) (R_{jk} - \bar{R}_j)}{[\sum_k (R_{ik} - \bar{R}_i)^2 \sum_k (R_{jk} - \bar{R}_j)^2]^{0.5}}$$

$$W_{ij} = \frac{\sum_k (R_{ik} - \bar{R}_i) (R_{jk} - \bar{R}_j)}{[\sum_k (R_{ik} - \bar{R}_i)^2 \sum_k (R_{jk} - \bar{R}_j)^2]^{0.5}}$$

Correlation is one possible similarity measure

Correlation between the ratings, computed a subset of items rated by both users, shows how aligned the users are in their tastes. It ranges from +1 (very similar) to -1 (opposite) tastes.

					
Avg.: 1.5					
JOHN	1		1	2	2
YOU	2	3	?	5	4
Avg.: 3.5					

$$W_{John, You} = \frac{(1-1.5) \cdot (2-3.5) + (2-1.5) \cdot (5-3.5) + (2-1.5) \cdot (4-3.5)}{\sqrt{[(1-1.5)^2 + (2-1.5)^2 + (2-1.5)^2] \cdot [(2-3.5)^2 + (5-3.5)^2 + (4-3.5)^2]}} \approx 0.95$$

$$W_{ij} = \frac{\sum_k (R_{ik} - \bar{R}_i) (R_{jk} - \bar{R}_j)}{[\sum_k (R_{ik} - \bar{R}_i)^2 \sum_k (R_{jk} - \bar{R}_j)^2]^{0.5}}$$

Correlation is one possible similarity measure

Correlation between the ratings, computed a subset of items rated by both users, shows how aligned the users are in their tastes. It ranges from +1 (very similar) to -1 (opposite) tastes.

					
JOHN	1		1	2	2
LUCY	1	2	5	5	5
DIANE	4	5	3	3	
YOU	2	3	?	5	4

$$w_{John, You} \approx 0.95, \quad w_{Lucy, You} \approx 0.94, \quad w_{Diane, You} \approx -0.65$$

Predicted rating is a “weighted average”

Given user's rating for a specific movie can be predicted as a weighted average of the ratings given by all the other users who watch it, with weights being the similarity between those users and the one in question

					
JOHN	1		1	2	2
LUCY	1	2	5	5	5
DIANE	4	5	3	3	
YOU	2	3	?	5	4

$$\hat{R}_{ik} = \bar{R}_i + \sum_{X_j \in N_i} W_{ij} (R_{jk} - \bar{R}_j)$$

Predicted rating is a weighted average

Given user's rating for a specific movie can be predicted as a weighted average of the ratings given by all the other users who watch it, with weights being the similarity between those users and the one in question

					
JOHN	1		1	2	2
LUCY	1	2	5	5	5
DIANE	4	5	3	3	
YOU	2	3	?	5	4

$$\hat{R}_{ik} = \bar{R}_i + \sum_{X_j \in N_i} W_{ij} (R_{jk} - \bar{R}_j)$$

💡 Need to account for different personal scales

$$\hat{R}_{ik} = \bar{R}_i + \sum_{X_j \in \mathbf{N}_i} W_{ij} (R_{jk} - \bar{R}_j)$$

$$w_{John, You} \approx 0.95, \quad w_{Lucy, You} \approx 0.94, \quad w_{Diane, You} \approx -0.65$$

Predicted rating is a weighted average

Given user's rating for a specific movie can be predicted as a weighted average of the ratings given by all the other users who watch it, with weights being the similarity between those users and the one in question

					
JOHN	1		1	2	2
LUCY	1	2	5	5	5
DIANE	4	5	3	3	
YOU	2	3	?	5	4

$$\hat{R}_{You, Narcos} = 3.5 + 0.95 \cdot (1 - 1.5) + 0.94 \cdot (5 - 3.6) - 0.65 \cdot (3 - 3.75) \approx 4.8$$

Try to implement this!

COLAB NOTEBOOK

<https://shorturl.at/dkwA5>



GITHUB

<https://shorturl.at/JLOX8>



Thank you!

Today's materials will be here



<https://github.com/evgeniyako-edu/build-your-first-recsys-workshop>

Let's stay connected!

Evgeniya Korneva

Senior Data Scientist, Monster



evgeniakorneva@gmail.com



<https://www.linkedin.com/in/evgeniyako>

