Build your first RecSys

Evgeniya Korneva
January 19, 2024
Barcelona

On the menu today

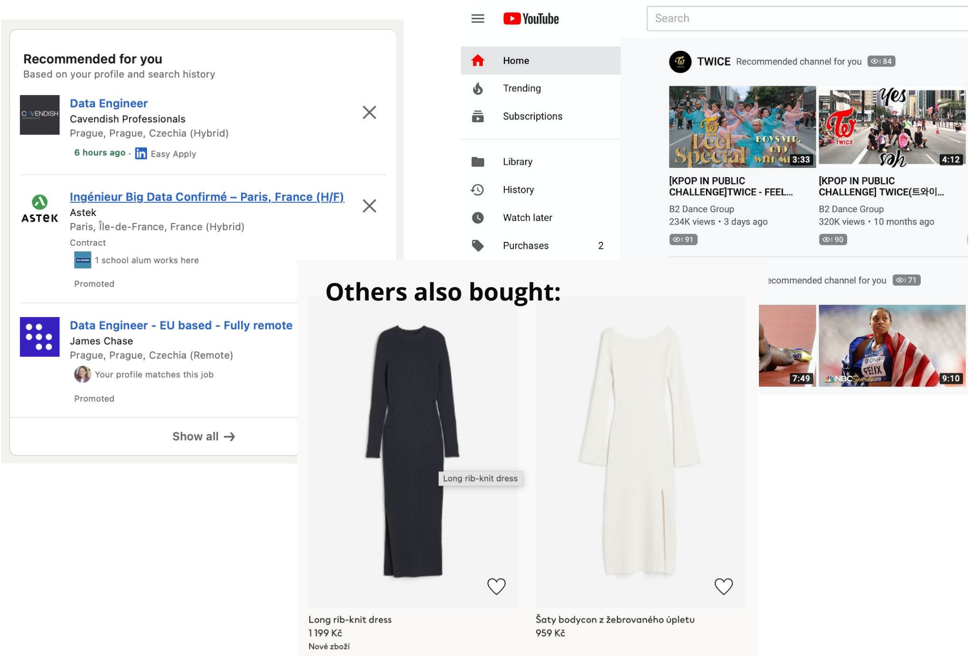
Overview of
recommendation systems

Collaborative filtering
in depth
(simple approach)

Practice time!

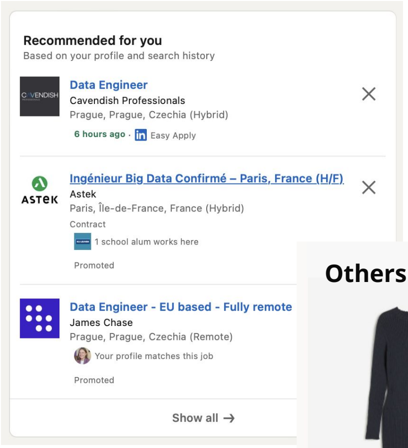
Recommendations are everywhere

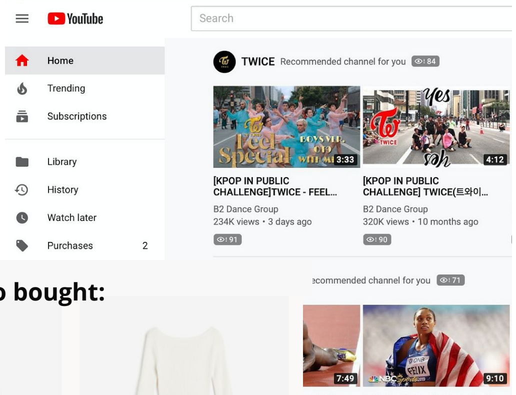
Hardly any part of our life isn't affected by personalized recommendations. While RecSys help users find compelling content in a large corpora, some consider personalization a serious concern

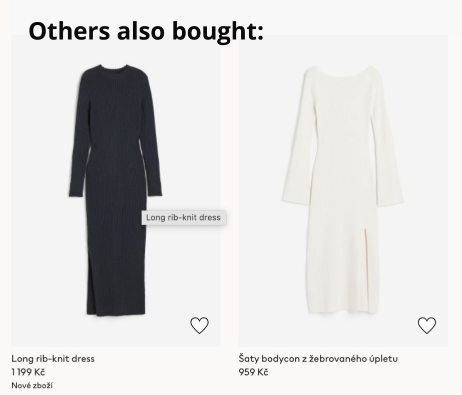


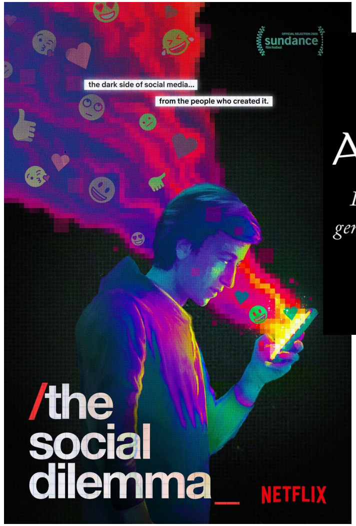
Recommendations are everywhere

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



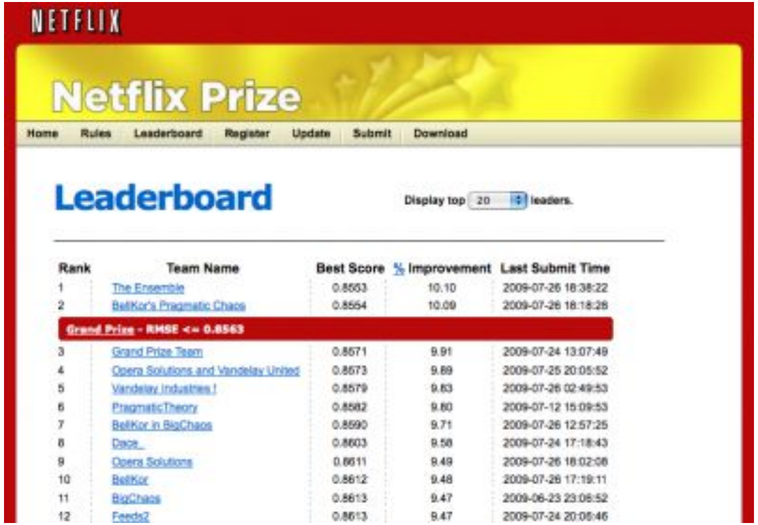






Netflix Prize boosted interest in recommenders

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



Full story: https://www.wired.com/2012/04/netflix-prize-costs/

There are different types of recommendations

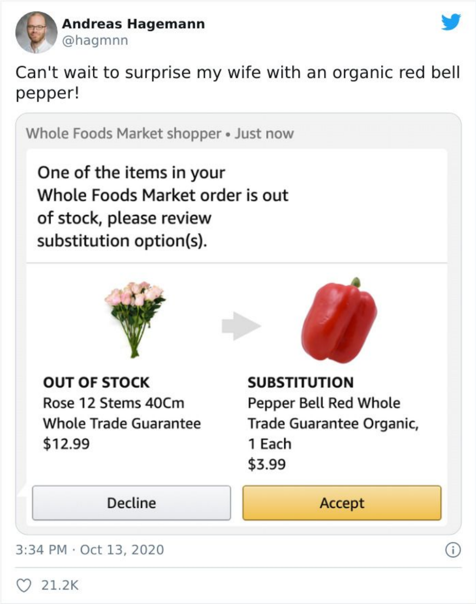
We need recommendations in different context

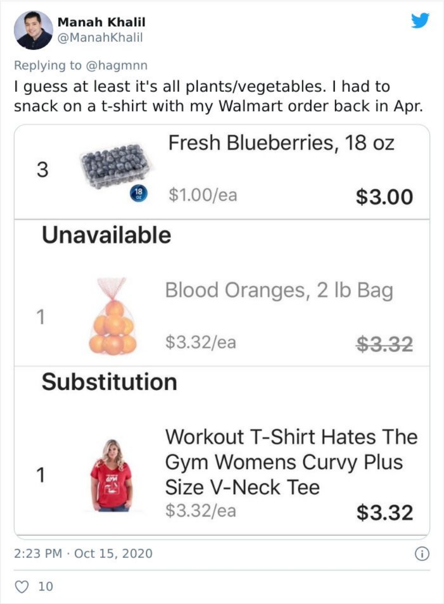
Different systems can be used to provide good recommendations in different situations. Not all of them are necessarily personalized or need a complex ML-based solution

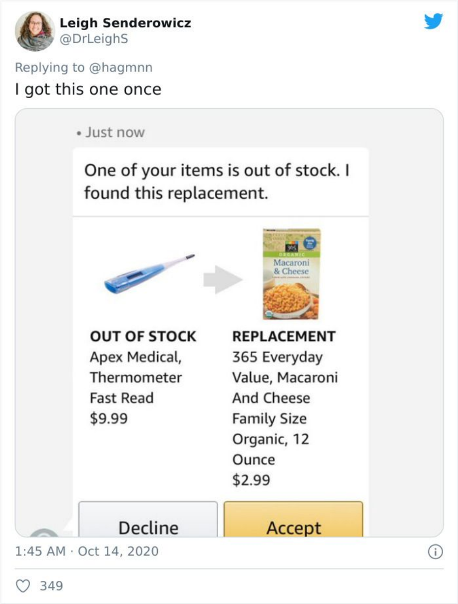
Non-personalised	• Substitutions for out-of-stock items • Cross-sales: items frequently bought together
Personalised	• *"You might also like …"* • *"Others also liked …"*

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…







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

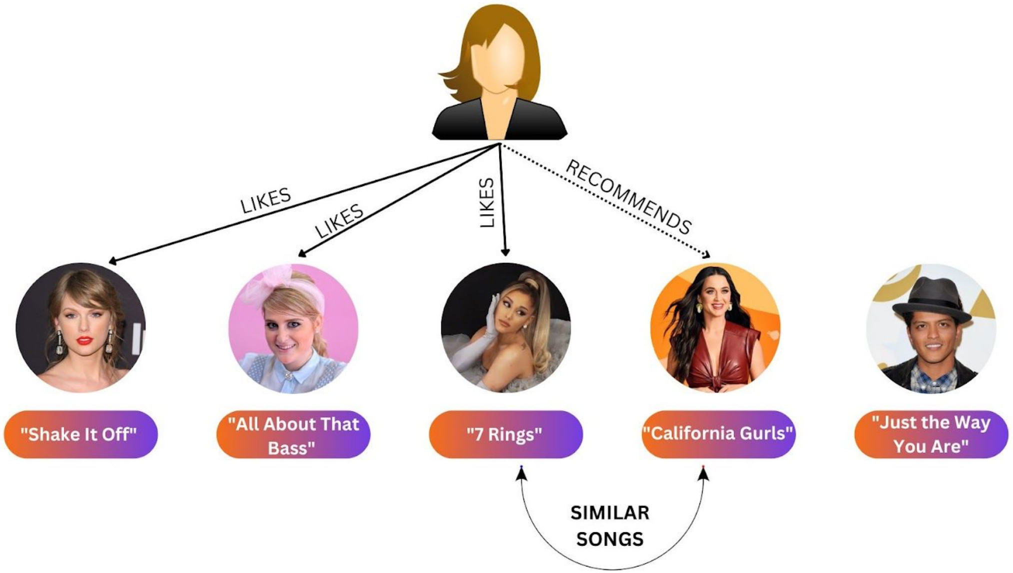


Image source: https://redfield.ai/content-based-recommendation/

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

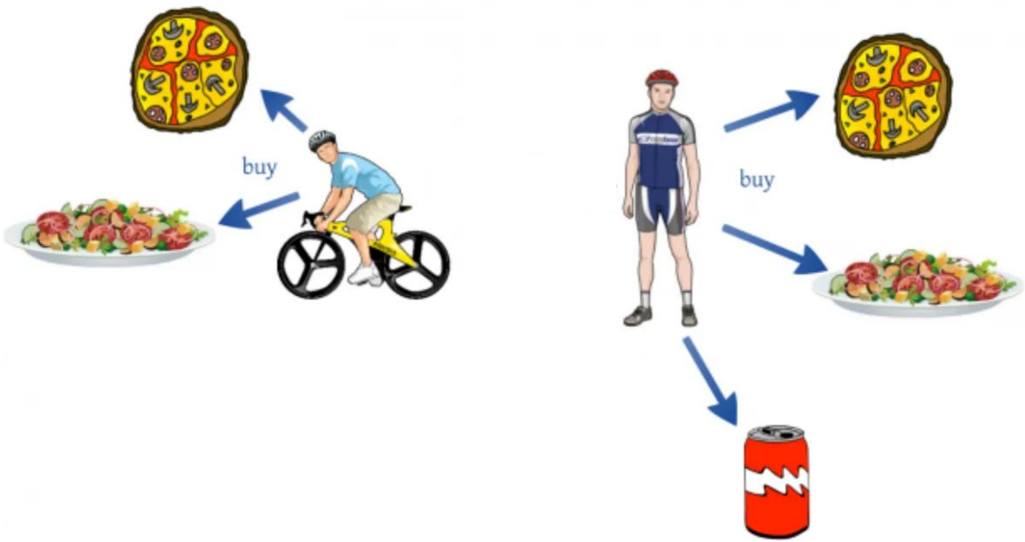


Image source: https://towardsdatascience.com/various-implementations-of-collaborative-filtering-100385c6dfe0

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

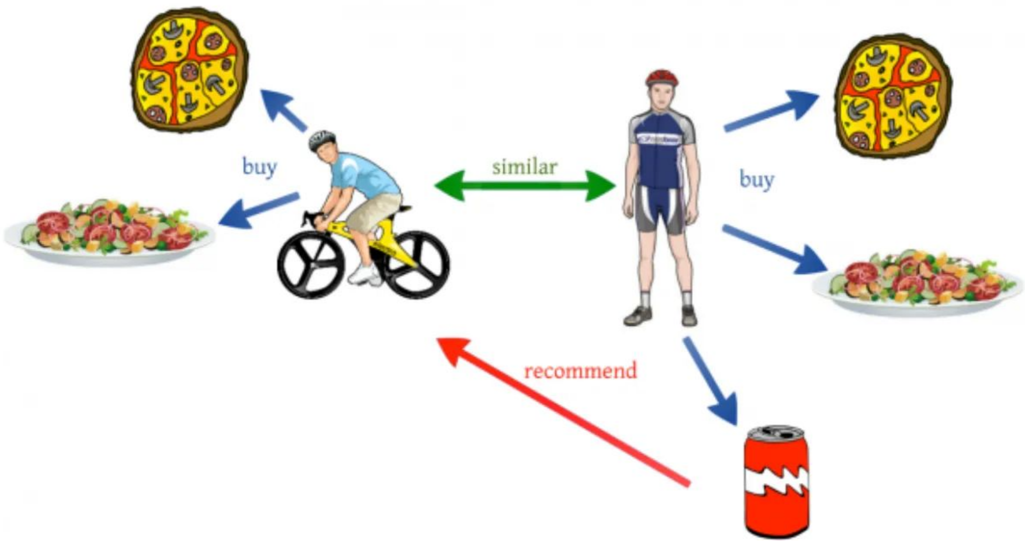


Image source: https://towardsdatascience.com/various-implementations-of-collaborative-filtering-100385c6dfe0

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 ar ethen 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 ar ethen 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

	ORANGE is the new BLACK	STRANGER THINGS	NARCOS	HOUSE of CARDS	DAREDEVIL
JOHN	1		1	2	2
LUCY	1	2	5	5	5
DIANE	4	5	3	3	
YOU	2	3	?	5	4

Image source: https://medium.com/the-graph/how-recommender-systems-make-their-suggestions-da6658029b76

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

	ORANGE is the new BLACK	STRANGER THINGS	NARCOS	HOUSE of CARDS	DAREDEVIL
JOHN	1		1	2	2
LUCY	1	2	5	5	5
DIANE	4	5	3	3	
YOU	2	3	?	5	4

Image source: https://medium.com/the-graph/how-recommender-systems-make-their-suggestions-da6658029b76

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.

	ORANGE is the new BLACK	STRANGER THINGS	NARCOS	HOUSE of CARDS	DAREDEVIL
JOHN	1		1	2	2
YOU	2	3	?	5	4

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

Image source: https://medium.com/the-graph/how-recommender-systems-make-their-suggestions-da6658029b76

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

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.

	ORANGE is the new BLACK	STRANGER THINGS	NARCOS	HOUSE of CARDS	DAREDEVIL
JOHN (*Avg.: 1.5*)	1		1	2	2
YOU (*Avg.: 3.5*)	2	3	?	5	4

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

Image source: https://medium.com/the-graph/how-recommender-systems-make-their-suggestions-da6658029b76

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

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.

	ORANGE is the new BLACK	STRANGER THINGS	NARCOS	HOUSE of CARDS	DAREDEVIL
JOHN	1		1	2	2
LUCY	1	2	5	5	5
DIANE	4	5	3	3	
YOU	2	3	?	5	4

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

Image source: https://medium.com/the-graph/how-recommender-systems-make-their-suggestions-da6658029b76

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

	ORANGE is the new BLACK	STRANGER THINGS	NARCOS	HOUSE of CARDS	DAREDEVIL
JOHN	1		1	2	2
LUCY	1	2	5	5	5
DIANE	4	5	3	3	
YOU	2	3	?	5	4

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

Image source: https://medium.com/the-graph/how-recommender-systems-make-their-suggestions-da6658029b76

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

	ORANGE is the new BLACK	STRANGER THINGS	NARCOS	HOUSE of CARDS	DAREDEVIL
JOHN	1		1	2	2
LUCY	1	2	5	5	5
DIANE	4	5	3	3	
YOU	2	3	?	5	4

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

💡 Need to account for different personal scales

Image source: https://medium.com/the-graph/how-recommender-systems-make-their-suggestions-da6658029b76

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

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

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

	ORANGE is the new BLACK	STRANGER THINGS	NARCOS	HOUSE of CARDS	DAREDEVIL
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	GITHUB
https://shorturl.at/dkwA5	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