Building a Recommender System with Python

•••

David Lim Cheng (X017427)
COMPSCI X433.3 - 005

Recommender Systems Overview

- Content-Based Filtering
  - Uses information about the description and attributes of previously-consumed items
  - Candidate items are compared to previously-consumed items, and similarly-matched items are recommended
- Collaborative Filtering
  - Makes predictions about user interests based on the interests of many users (collaborating)
  - If Person A has the same opinion as Person B on Item 1, then Person A is more likely to share Person B's opinion for any given item

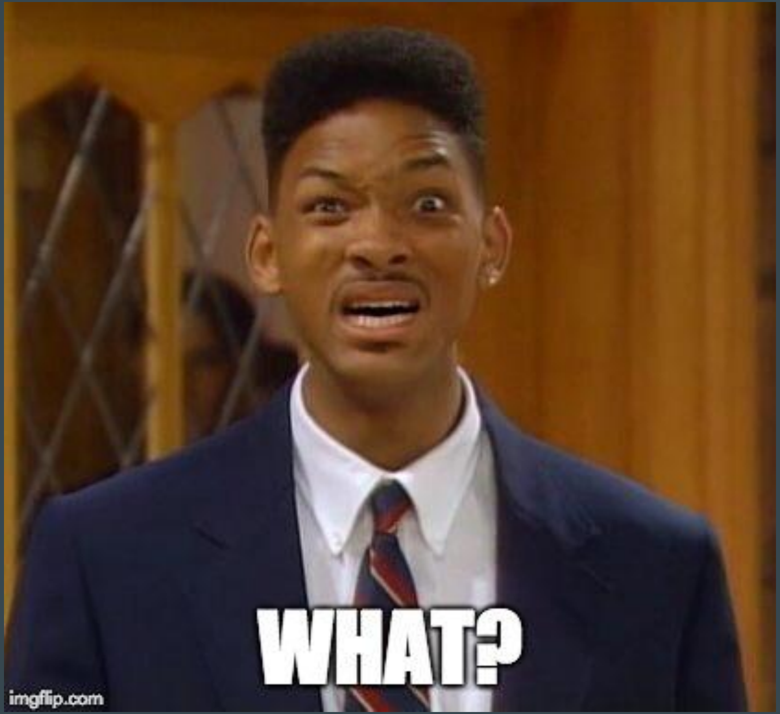
Singular Value Decomposition (SVD)

- Algorithm that leverages a latent factor model to capture similarities between users and items
- Decreases dimensionality of the utility matrix by extracting latent factors
- Given by the following equation:

$$R = M\Sigma U^T$$

- Where:
  - M is orthogonal to the column space of R,
  - U is orthonormal to the row space of R

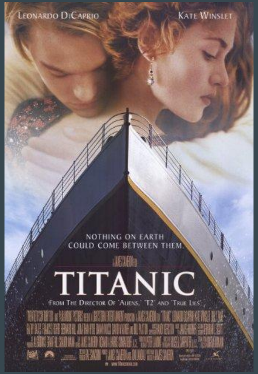
Uhhhhhhh...............



Latent Factors

*"Algorithm that leverages a **latent factor model** to capture similarities between users and items"*

- Important part of **Principal Component Analysis (PCA)**, a precursor to SVD
- Another way to say "latent factors" is "typical traits"
- **Oversimplified** example:
  - Our dataset is full of **movies** and user ratings
  - "Latent factors" or "typical traits" of movies: **GENRE!**

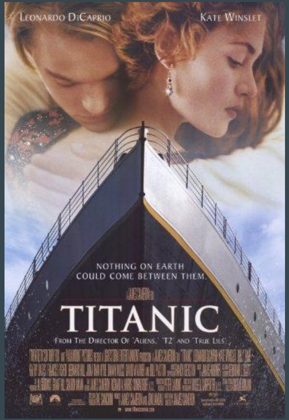

LEONARDO DiCAPRIO
KATE WINSLET
NOTHING ON EARTH COULD COME BETWEEN THEM.
TITANIC
FROM THE DIRECTOR OF 'ALIENS,' 'T2' AND 'TRUE LIES'

"45% Drama, 25% Action, 10% Comedy, 5% Thriller…"

Latent Factors (Continued)

- **Oversimplified** example:
  - Our dataset is full of movies and **user ratings**
  - “Latent factors” or “typical traits” of user ratings: **AFFINITY TO GENRES**

“50% Affinity for Drama, 25% Affinity for Comedy, 15% Affinity for True Crime…”


Leonardo DiCaprio
Kate Winslet
Nothing on Earth could come between them.
TITANIC
From the Director of 'Aliens,' 'T2' and 'True Lies'

"45% Drama
25% Action
10% Comedy
5% Thriller…"

"50% Affinity for Drama,
25% Affinity for Comedy,
15% Affinity for True Crime…"

SVD Says: "Bet that old lady likes Titanic!"

Matrix Factorization

"Decreases dimensionality of the utility matrix"

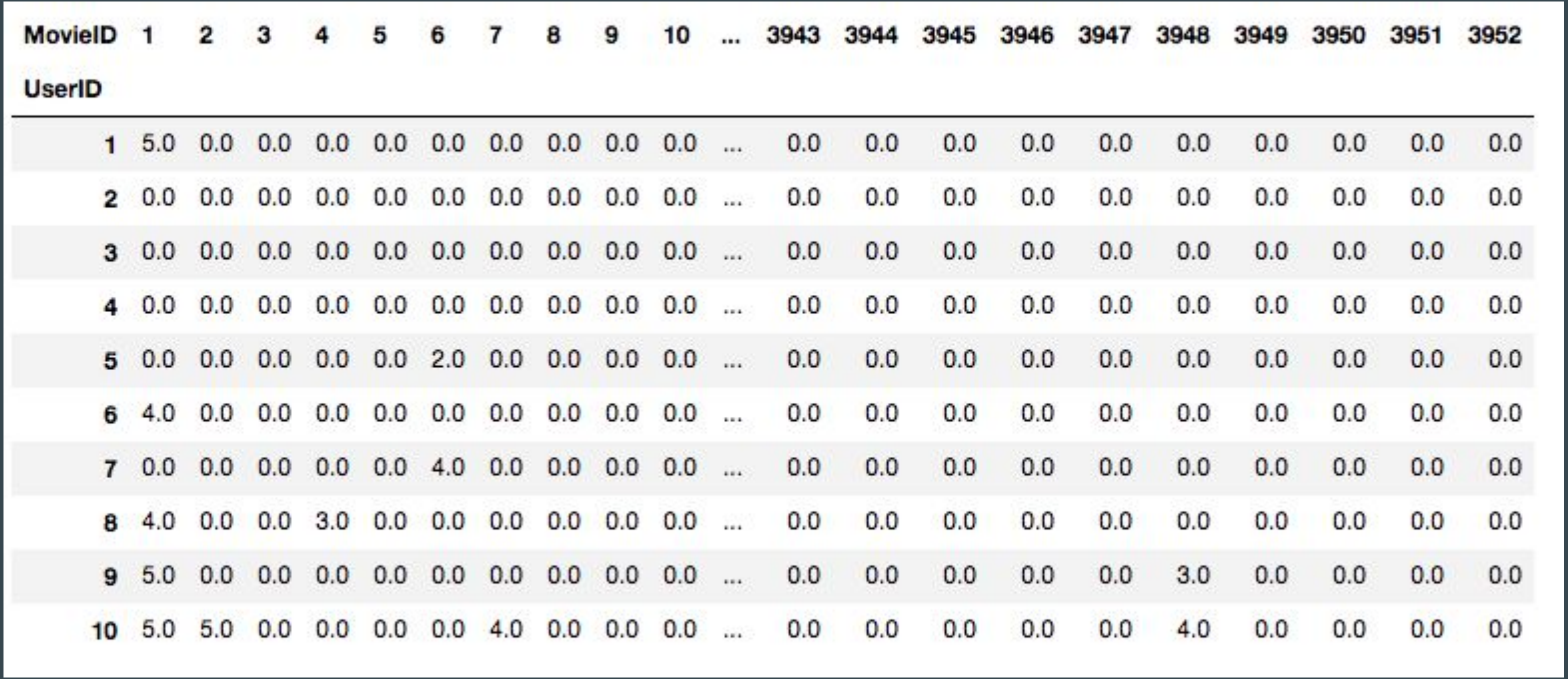
- Analogy: Polynomial Factorization
  - $X^2 - 4 \rightarrow (x - 2)(x + 2)$
  - Took **one** item, and simplified it by turning it into **two** items
  - We can *rebuild* the original term using the simplified terms
- Matrix Factorization is the same thing
  - Matrix → (Vector)(Vector)(Vector)
  - Took **one** item, and simplified it by turning it into **three** items
  - We can *rebuild* the original matrix using the simplified vectors

Matrix = (Vector)(Vector)(Vector)

$$R = M\Sigma U^T$$

Matrix Factorization (Continued)

*"Decreases dimensionality of the **utility matrix**"*



MovieID	1	2	3	4	5	6	7	8	9	10	...	3943	3944	3945	3946	3947	3948	3949	3950	3951	3952
UserID																					
1	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	0.0	0.0	0.0	0.0	0.0	2.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	4.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8	4.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	5.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	3.0	0.0	0.0	0.0	0.0
10	5.0	5.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	...	0.0	0.0	0.0	0.0	0.0	4.0	0.0	0.0	0.0	0.0

Recap

- Reveal the latent factors of both the movies and the users using SVD
- Rebuild the utility matrix using these latent factors
- Use the rebuilt matrix to make predictions of what movies would receive a high rating for a particular user

“SVD? Yeahhhh, baby!”