


Bălălău Andrei

Lican Ștefăniță

Sistem de recomandare pentru cărți: implementare, analiză și interpretare rezultate

Folosind algoritmul - descompunerea valorilor singulare

- 
1. **Introducere**
 2. **Arhitectura sistemului de recomandare**
 3. **Prezentare implementare (algoritmi, formule matematice)**
 4. **Interpretarea rezultatelor obținute**
 5. **Beneficii și îmbunătățiri**

Cuprins

1 | Introducere

Ce este un sistem de recomandare?

Un sistem care sugerează conținut personalizat pentru utilizatori, bazându-se pe datele lor.

Exemple populare:

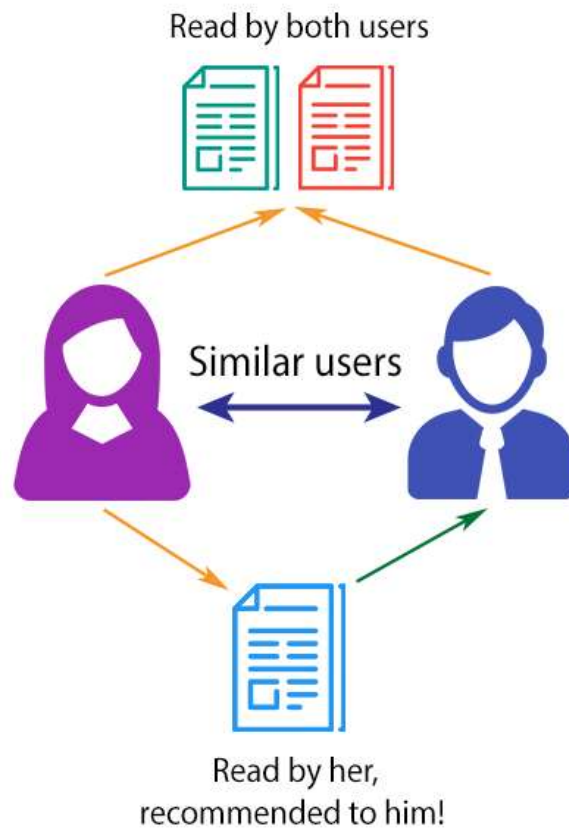
Amazon (e-commerce), Netflix (streaming video),
Goodreads (cărți).

Obiectivul nostru (în cadrul proiectului):

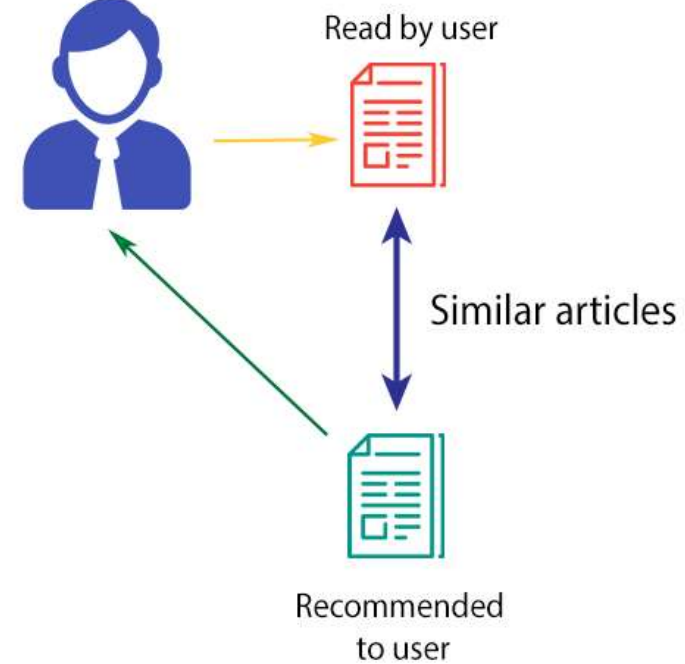
Dezvoltarea unui sistem care prezice rating-urile pentru un set de cărți și generează, astfel, recomandări personalizate.

1 | Introducere

COLLABORATIVE FILTERING



CONTENT-BASED FILTERING



2 | Arhitectura aplicației

Fluxul procesului:

1. Preprocesare a datelor (filtrare și pregătire matrice).
2. Aplicare de factorizare matricială (SVD).
3. Calculul rating-urilor prezise.
4. Generarea recomandărilor personalizate.

Biblioteci:

1. Pandas (pentru gestionarea datelor din CSV)
2. NumPy (calculare numerice)
3. SciPy (gestionarea calculelor cu matricea rară)
4. Seaborn (vizualizare grafică)
5. Matplotlib (din Matlab, pentru vizualizarea datelor)

2 | Seturi de date și procesări

Books.csv:

Conține informații despre cărți (ISBN, titlu, autor).

Ratings.csv:

Conține rating-urile atribuite de utilizatori pentru cărțile evaluate.

Procesare:

- Eliminarea duplicatelor (prima etapă).
- Filtrarea utilizatorilor și cărților fără suficiente evaluări (a doua etapă).

Statistici:

Dimensiunea finală a datelor (filtrate): utilizatori: **17,744**; cărți: **15,055**;

Densitatea matricei: **0.15%** - proporția de elemente nenule din matrice, raportată la numărul total de elemente.

3 | Implementare - formule matematice

➤ Matricea ratingurilor

$R \in \mathbb{R}^{m \times n}$, unde: m - numărul de utilizatori și n - numărul de cărți din baza de date. Elementele R_{ij} : ratingul dat de utilizatorul i cărții j .

➤ Factorizarea matricei cu SVD:

$R \approx U\Sigma V^T$, unde: U - matricea preferințelor utilizatorilor, Σ - matricea diagonală, cu valori singulare și V^T - matricea caracteristicilor cărților.

➤ Prezicerea rating-urilor:

Dedusă de pe urma DVS: $R_{ij} = \mu + b_i + b_j + \sum_{f=1}^k U_{if}\Sigma_f V_{fj}$, cu μ : rating mediu global; b_i, b_j : bias-uri – aprecieri stricte (utilizator/carte); $U_{if}\Sigma_f V_{fj}$: factori latent.

3 | Implementarea practică - Etape

1. Preprocesare:

Încărcarea datelor
Filtrarea cărților și utilizatorilor

2. Centrare matrice:

Eliminarea “bias-urilor” ((ne)aprecierilor
generale / individuale și filtrarea datelor rămase)

3. Factorizare:

Aplicare *SVD custom* și compararea cu *SVD standard* (`scipy.sparse.linalg.svds`)

4. Predicție:

Calculul rating-urilor prezise, după formula
descrișă anterior

5. Generare recomandări:

Lista de cărți sugerate pentru utilizatori – top 10
recomandări, pe baza predicțiilor / estimărilor de rigoare

1. Dimensiunea matricei: $17,744 \times 15,055$.
2. Densitatea (proporția de elemente nenule din matrice, raportată la numărul total de elemente): 0.15% (foarte rară).
3. Interpretare:
 - Matricea conține multe celule goale (ratinguri lipsă)
 - Metodele de completare precum SVD sunt esențiale pentru acest tip de date.

4 | Interpretarea rezultatelor

Matricea rară și densitatea

Comparăm valorile singulare:

Valorile singulare din SVD personalizat:

```
[226.64102 156.50238 115.535385 111.24676 108.36207 107.26156
103.48345 100.219574 99.799126 97.020035 93.62046 93.38408
92.16195 89.75323 89.05042 88.37679 87.844574 85.81656
85.24523 83.87518 83.37187 82.74826 82.05515 81.80576
81.26917 80.65301 80.319405 79.79019 79.36329 78.99025
78.699104 78.06072 77.72968 77.574715 77.39171 77.06373
76.562614 75.95742 75.62087 75.44317 74.99136 74.313965
73.9203 73.82392 73.52397 73.34763 72.92846 72.628044
72.40747 72.28821 ]
```

Valorile singulare din SVD standard:

```
[226.64116 156.50243 115.53542 111.24667 108.36202 107.26153
103.483536 100.21953 99.79913 97.02004 93.620476 93.38407
92.16193 89.75318 89.05044 88.37673 87.84455 85.81661
85.24525 83.87522 83.37194 82.7483 82.0551 81.8058
81.26922 80.65303 80.31943 79.79021 79.3633 78.99025
78.699135 78.06073 77.72968 77.574715 77.39172 77.063736
76.56261 75.95741 75.62084 75.44316 74.99135 74.31397
73.92029 73.82394 73.52394 73.34761 72.92843 72.62806
72.4075 72.2882 ]
```

Diferențele valorilor singulare:

```
[1.3732910e-04 4.5776367e-05 3.8146973e-05 8.3923340e-05 4.5776367e-05
3.0517578e-05 8.3923340e-05 4.5776367e-05 7.6293945e-06 7.6293945e-06
1.5258789e-05 7.6293945e-06 2.2888184e-05 4.5776367e-05 1.5258789e-05
6.1035156e-05 2.2888184e-05 5.3405762e-05 1.5258789e-05 3.8146973e-05
6.8664551e-05 3.8146973e-05 5.3405762e-05 3.8146973e-05 4.5776367e-05
2.2888184e-05 2.2888184e-05 1.5258789e-05 7.6293945e-06 0.0000000e+00
3.0517578e-05 7.6293945e-06 0.0000000e+00 0.0000000e+00 1.5258789e-05
7.6293945e-06 7.6293945e-06 7.6293945e-06 3.0517578e-05 7.6293945e-06
1.5258789e-05 7.6293945e-06 1.5258789e-05 1.5258789e-05 3.0517578e-05
2.2888184e-05 3.0517578e-05 1.5258789e-05 3.0517578e-05 7.6293945e-06]
```

Comparăm vectorii singulari stânga (U):

Norma Frobenius a diferenței pentru U: 10.583005

Comparăm vectorii singulari dreapta (V^T):

Norma Frobenius a diferenței pentru V^T: 10.582986

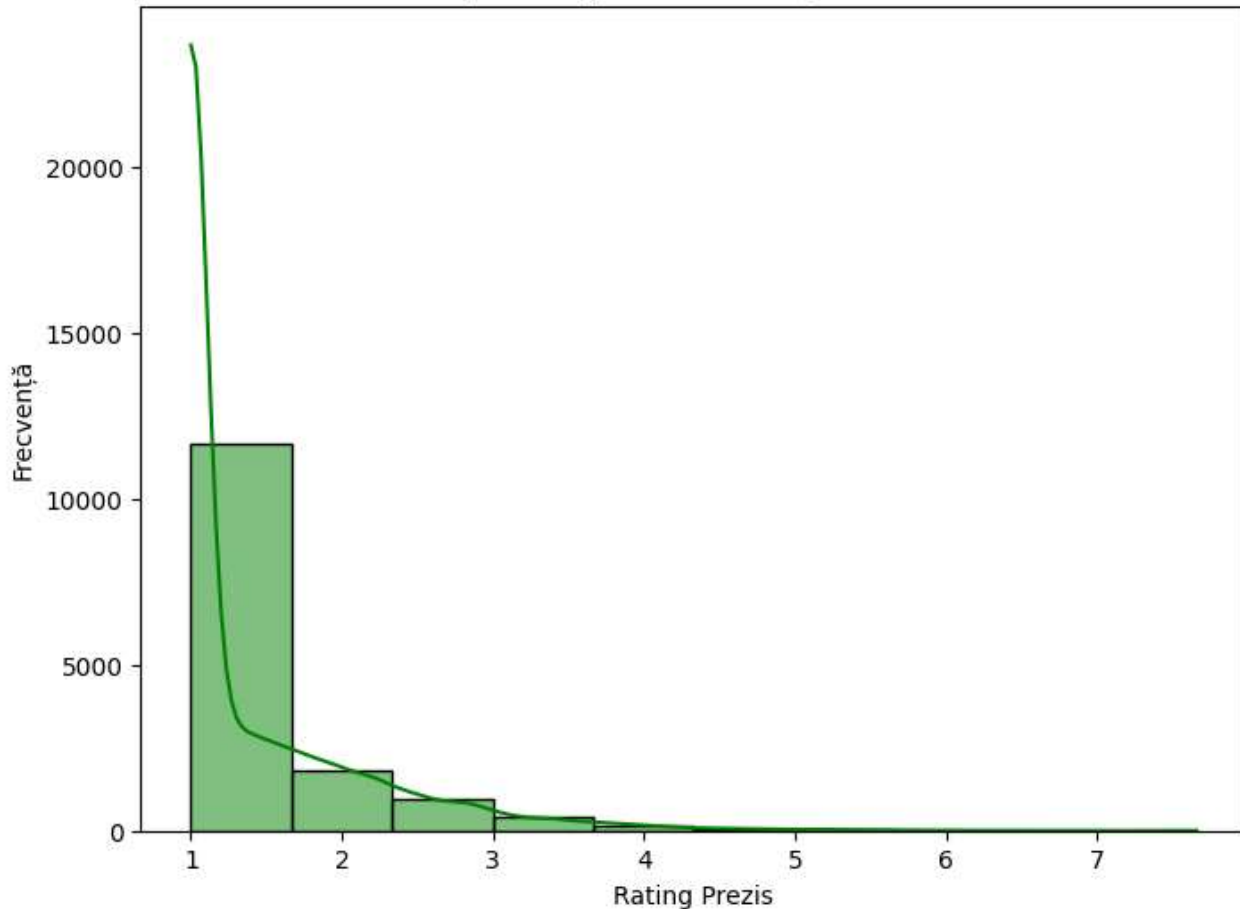
4 | Interpretarea rezultatelor

1. Valorile singulare: Diferența medie: $\approx 10^{-4}$ (aproape insignifiantă)
2. Norma Frobenius a diferențelor: $U: \approx 10$; $V^T: \approx 10$ (reducă, poate oscila, în funcție de cazul tratat – adică de utilizatorul selectat).
3. Interpretare: Rezultatele sunt aproape identice între implementarea noastră de SVD și cea standard, din python.

Comparare DVS

4 | Interpretarea rezultatelor

Distribuția Ratingurilor Prezise pentru Utilizator



1. Graficul distribuției: (Frecvența / modul de prezicere al rating-ului pentru restul cărților neevaluate de către utilizatorul selectat din listă)

Majoritatea ratingurilor prezise sunt între 3.5 și 7.5 (cel mai mare număr de rating-uri acordate la nivel general, valabil pentru toți userii – se observă că în graficul din stânga, pentru acest user, rating-ul predominant e extrem de mic – de pe urma unor evaluări mai critice a cărților).

Evitarea extremelor (rating-uri de 1 sau 10).

2. Interpretare:

Distribuția este echilibrată, indicând o bună funcționare a algoritmului. Pe baza acesteia, se determină primele 10 recomandări după predicția descrescătoare.

Distribuția rating-urilor

Recomandări de top pentru utilizatorul 276746:

- ```
=====
```
1. La Fata Carabina  
Autor: Daniel Pennac  
ISBN: 8807812576  
Rating Prezis: 7.66
  2. The Blue Day Book: A Lesson in Cheering Yourself Up  
Autor: Bradley Trevor Greive  
ISBN: 0091842050  
Rating Prezis: 6.99
  3. Das Zweite Gedächtnis.  
Autor: Ken Follett  
ISBN: 3404149114  
Rating Prezis: 6.66
  4. Love You Forever  
Autor: Robert Munsch  
ISBN: 0920668364  
Rating Prezis: 6.55
  5. Harry Potter Und Der Feuerkelch  
Autor: Joanne K. Rowling  
ISBN: 3551551936  
Rating Prezis: 6.36
  6. Jesus Freaks: DC Talk and The Voice of the Martyrs - Stories of Those Who Stood For Jesus, the Ultimate Jesus Freaks  
Autor: DC Talk  
ISBN: 1577780728  
Rating Prezis: 6.36
  7. The Princess Bride: S. Morgenstern's Classic Tale of True Love and High Adventure (The 'Good Parts' Version)  
Autor: William Goldman  
ISBN: 034543014X  
Rating Prezis: 6.16
  8. The Outlandish Companion  
Autor: DIANA GABALDON  
ISBN: 0385324138  
Rating Prezis: 6.06
  9. The Baby Book: Everything You Need to Know About Your Baby from Birth to Age Two  
Autor: Martha Sears  
ISBN: 0316779059  
Rating Prezis: 5.93
  10. When Did Ignorance Become A Point Of View  
Autor: Scott Adams  
ISBN: 0740718398  
Rating Prezis: 5.91

## 4 | Interpretarea rezultatelor

Recomandările sunt relevante și populare. Se poate observa că, în funcție de datele graficului, se generează lista primelor 10 filme recomandate, bazate pe predicția de apreciere a acestora.

## Legătura predicție - recomandări



### **Rezultate obținute:**

Sistemul poate prezice rating-uri cu acuratețe și poate genera recomandări personalizate.  
SVD este un algoritm eficient pentru factorizarea matricilor *sparse*.

### **Limitări:**

Densitatea scăzută a datelor poate afecta acuratețea.  
Scalabilitatea poate fi o problemă pentru seturi de date foarte mari (timp de execuție crescut).

# 5 | Concluzii



Bălălău Andrei

Lican Ștefăniță

Vă mulțumim!

321AC