Recommendation Systems

Agendas

01

Introduction

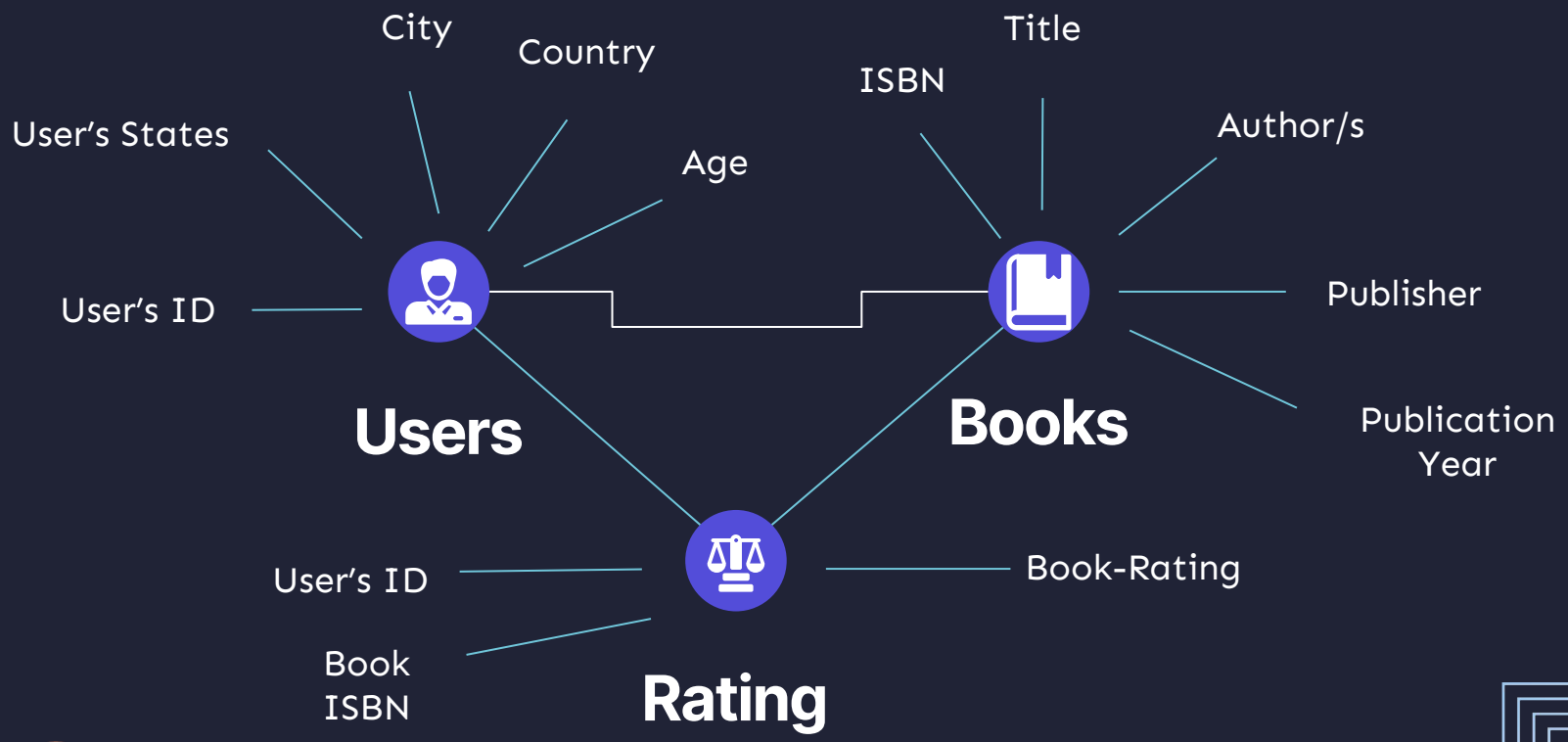
Recommendation Systems

- How does a system know what content you like?

Our aims

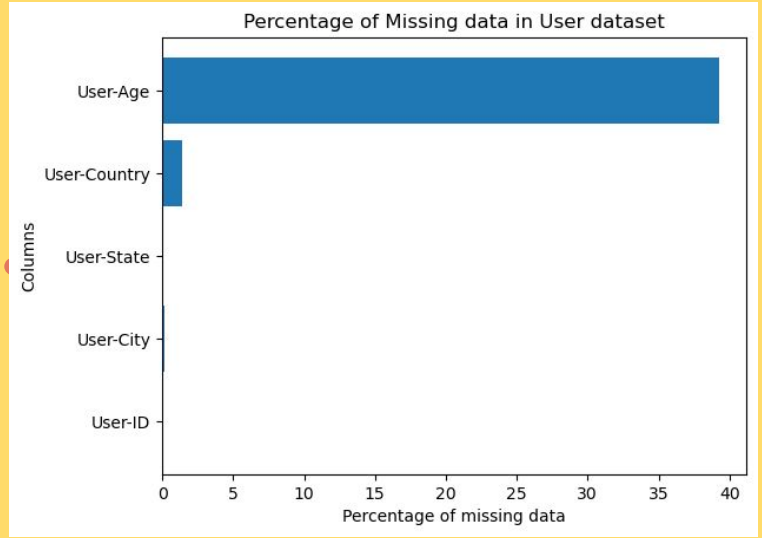
- Can the datasets be used to implement a recommendation system?
- How content-based and collaborative filtering methods are used in recommendation systems?

About the dataset



02

Data Preprocessing



Handle Missing Data

User-Age: Drop

User-Country: be filled by “other”

Text Processing

```
0              canada"
1                 usa"
2                 usa"
3                 usa"
4                  NaN
             ...
48294          canada"
48295          canada"
48296             usa"
48297       australia"
48298             usa"
```

```
0                Canada
1         United States
2         United States
3         United States
4                others
             ...
48294            Canada
48295            Canada
48296     United States
48297         Australia
48298     United States
```

Text Processing

Book-Title	preprocessed_title
Clara Callan	clara callan
Flu: The Story of the Great Influenza Pandemic...	flu story great influenza pandemic search viru...
The Kitchen God's Wife	kitchen god wife
The Testament	testament
Beloved (Plume Contemporary Fiction)	beloved plume contemporary fiction
...	...
Love, Etc.	love etc
The Wit And Whimsy Of Mary Engelbreit	wit whimsy mary engelbreit
Los Detectives Salvajes	los detective salvajes
The Ice House (TV Tie-In Edition)	ice house tv tie edition
Trouble Is My Business (Vintage Crime/Black Li...	trouble business vintage crime black lizard

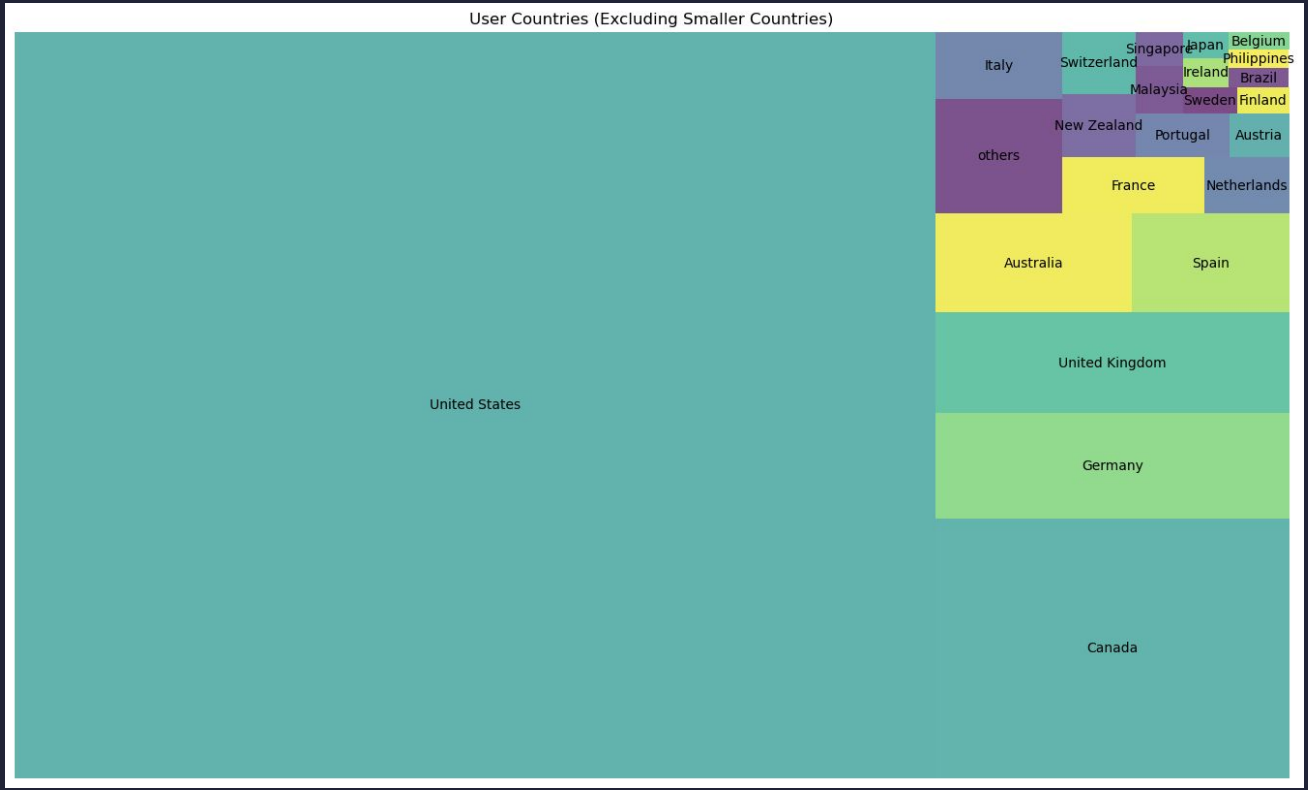
Discretization

Year Of Publication Category	
1991-2000	9517
After 2000	5330
1981-1990	2465
1971-1980	382
Before 1950	330
1961-1970	103
1950-1960	55

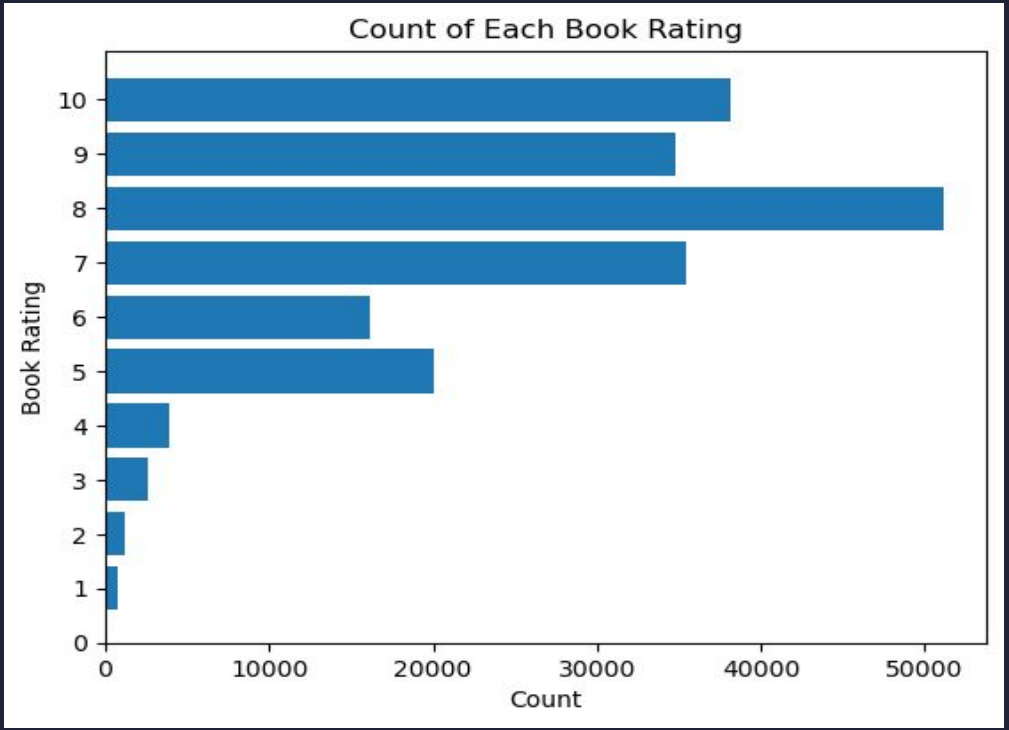
03

Exploratory data analysis

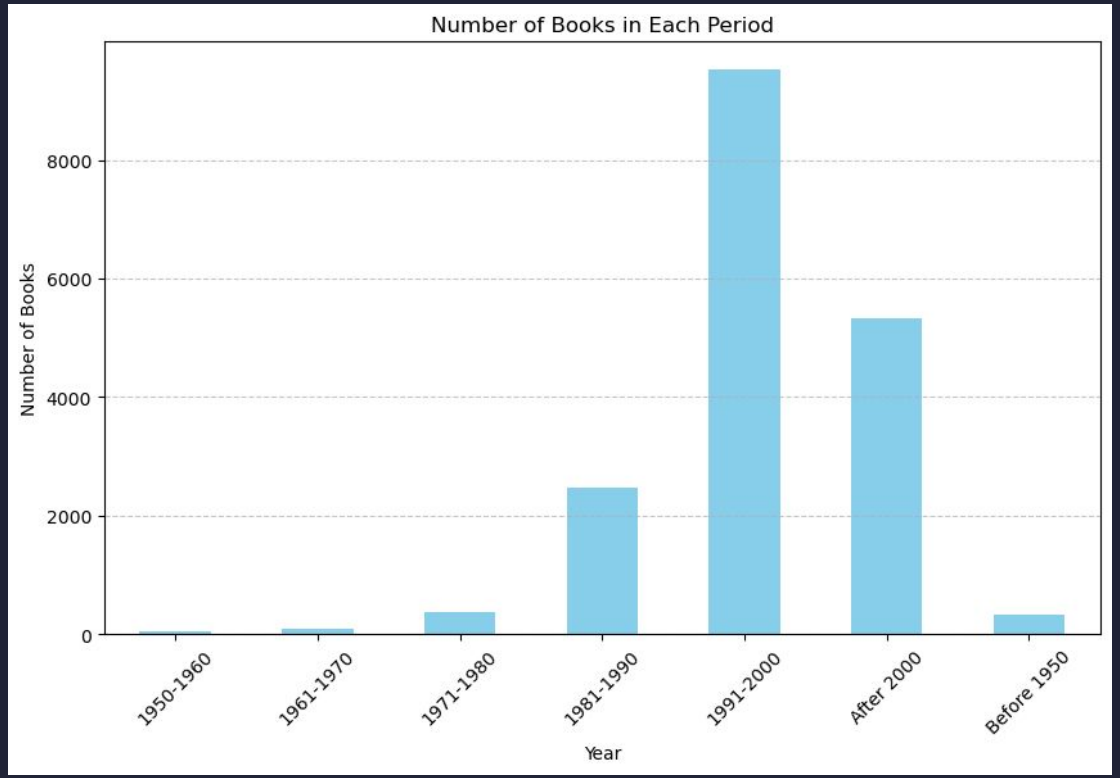
Countries distribution among users



Distribution of Book Rating



Distribution of Book publication year



04

Recommendation Systems

Recommendation system models

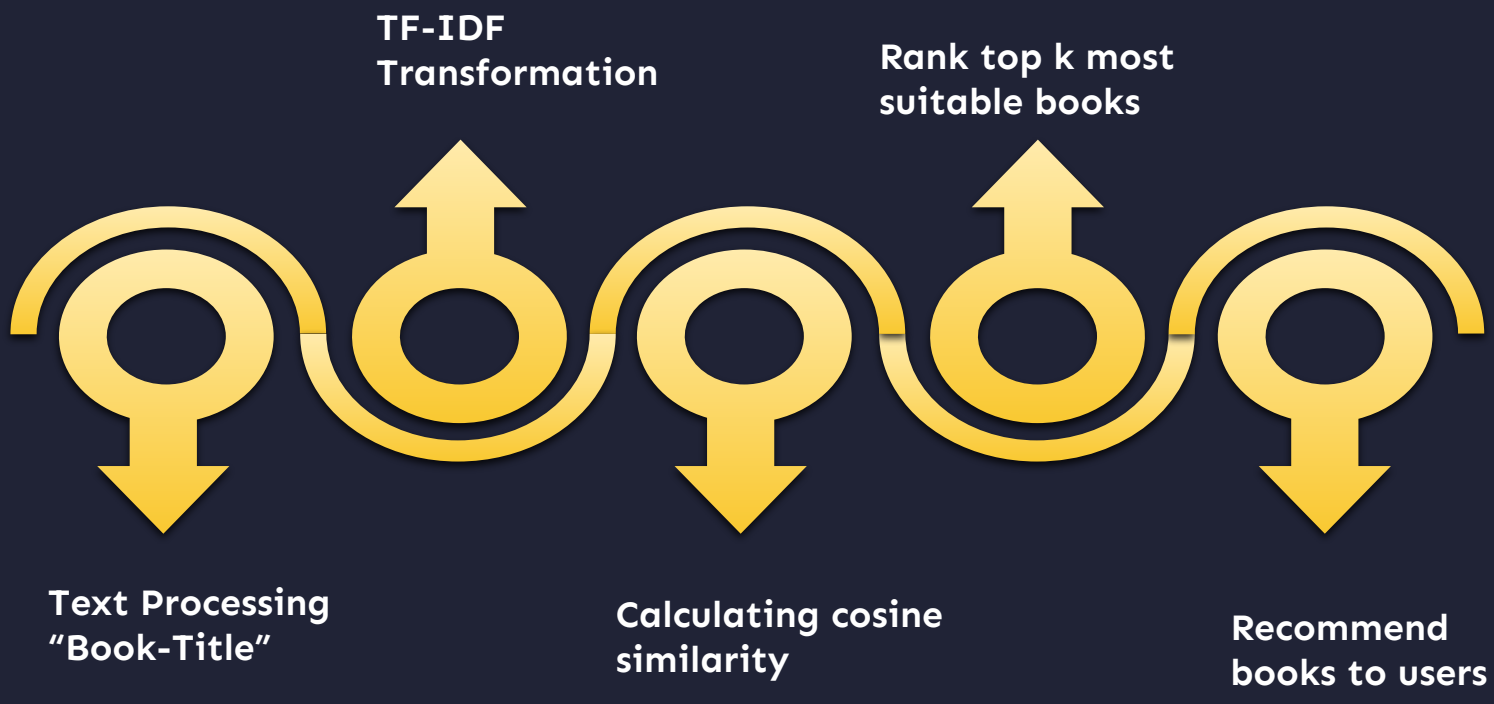
Collaborative filtering method

- Matrix Factorization using Singular Value Decomposition (SVD)

Content-based

- Term frequency-Inverse Document Frequency (TF-IDF)
- Cosine similarity

Content-Based Model



Creating TF-IDF vectors

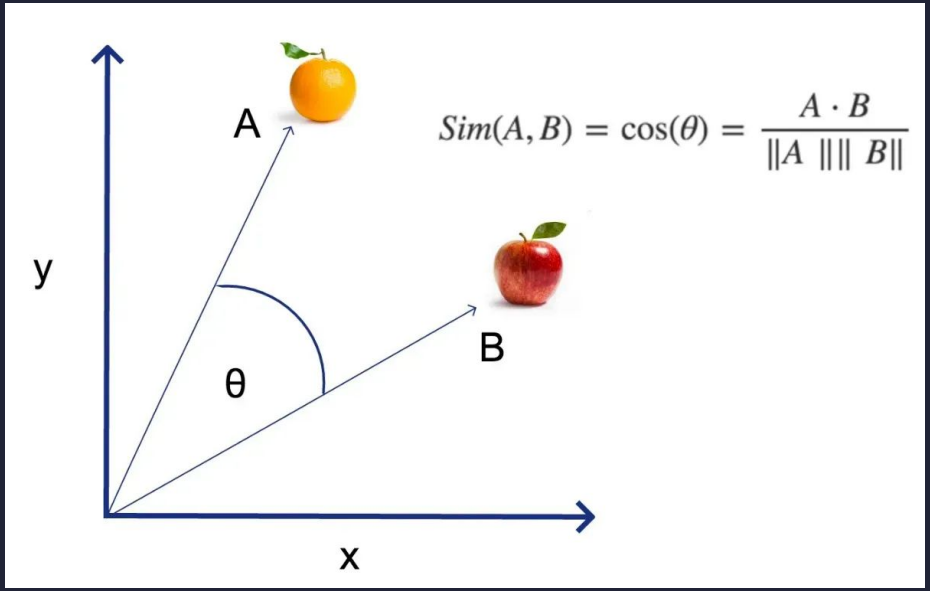
Term-Frequency:

- Is the number of a word or term appear in the document compared to the whole document

Inverse Document Frequency:

- The proportion of documents that contain that word/term in the corpus
  => It gives us the importance of the word

Calculating Cosine Similarity



$$Sim(A, B) = \cos(\theta) = \frac{A \cdot B}{\|A\| \|B\|}$$

Process of recommending books to users

Clicking on desired book

- User clicking on the book that they're interested in
- User search their desired book
- User choosing their favorite books when they first created an account

Calculating similarity

From the title, the model calculates the similarity of the book to other in the dataset

Recommend k books

After calculation, it sort the similarity score, and get the first K books

06

Limitations

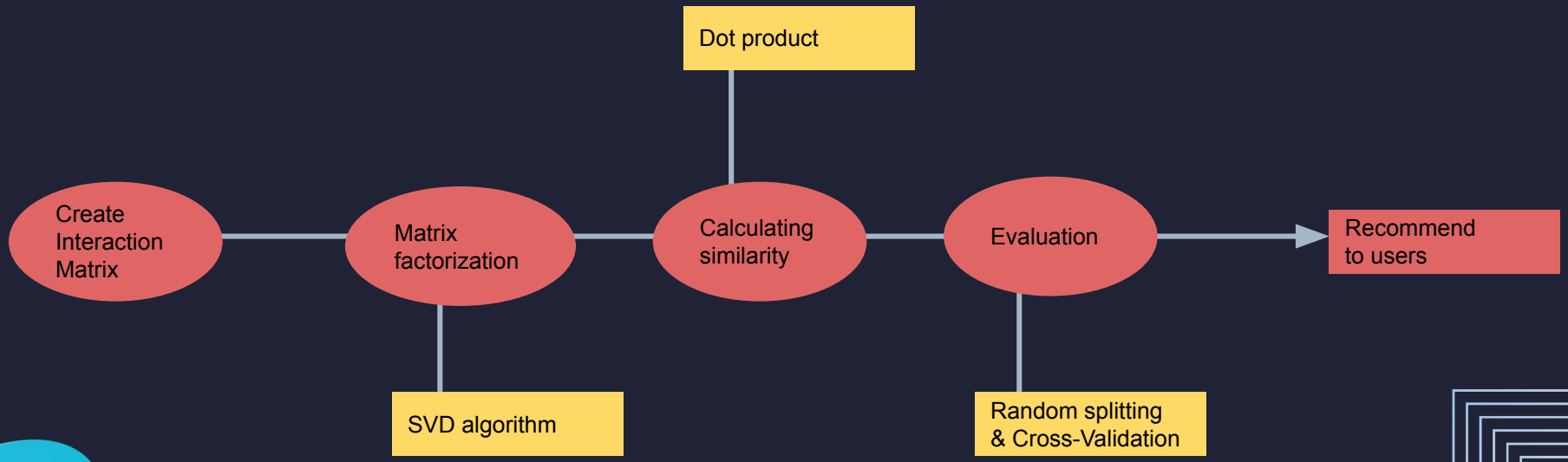
Limitations of the Content-based Model:

1. Ineffectiveness of TF-IDF: Limited common words reduce recommendation accuracy.

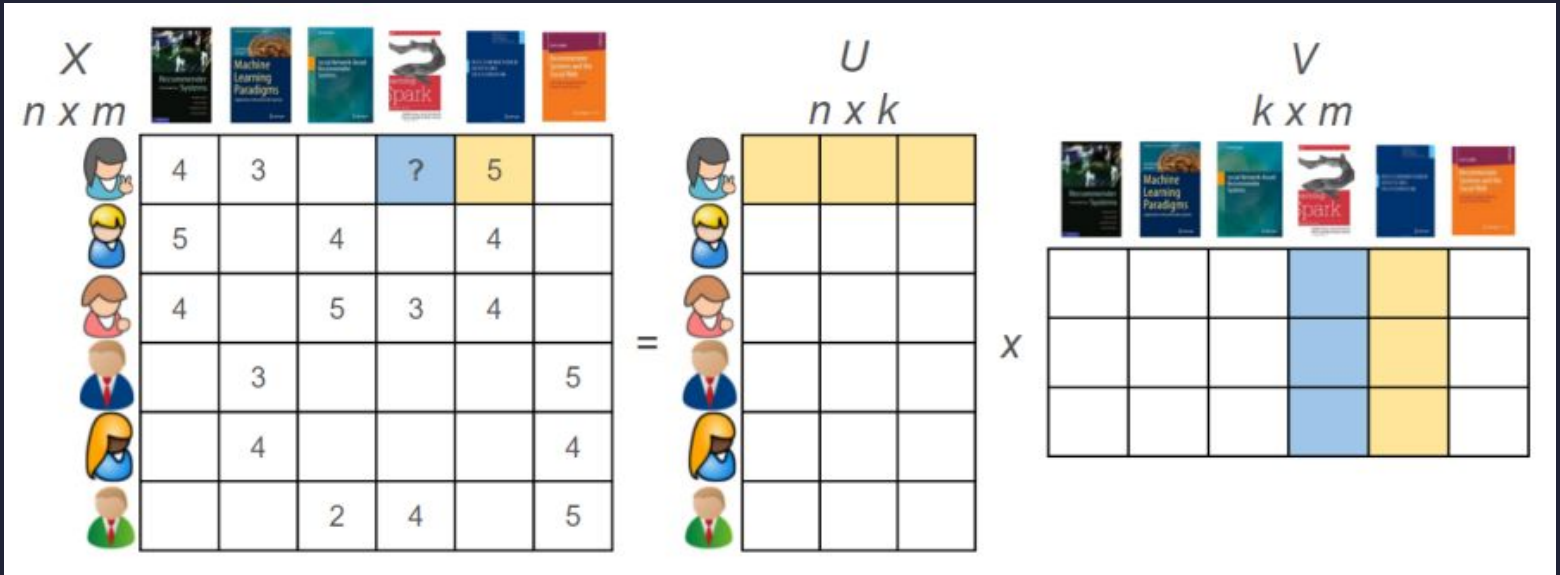
2. Language Limitation: Focuses only on English, excluding non-English books.

3. Lack of Exploration: Relies on past interests, hindering new discoveries.

Collaborative Filtering Model

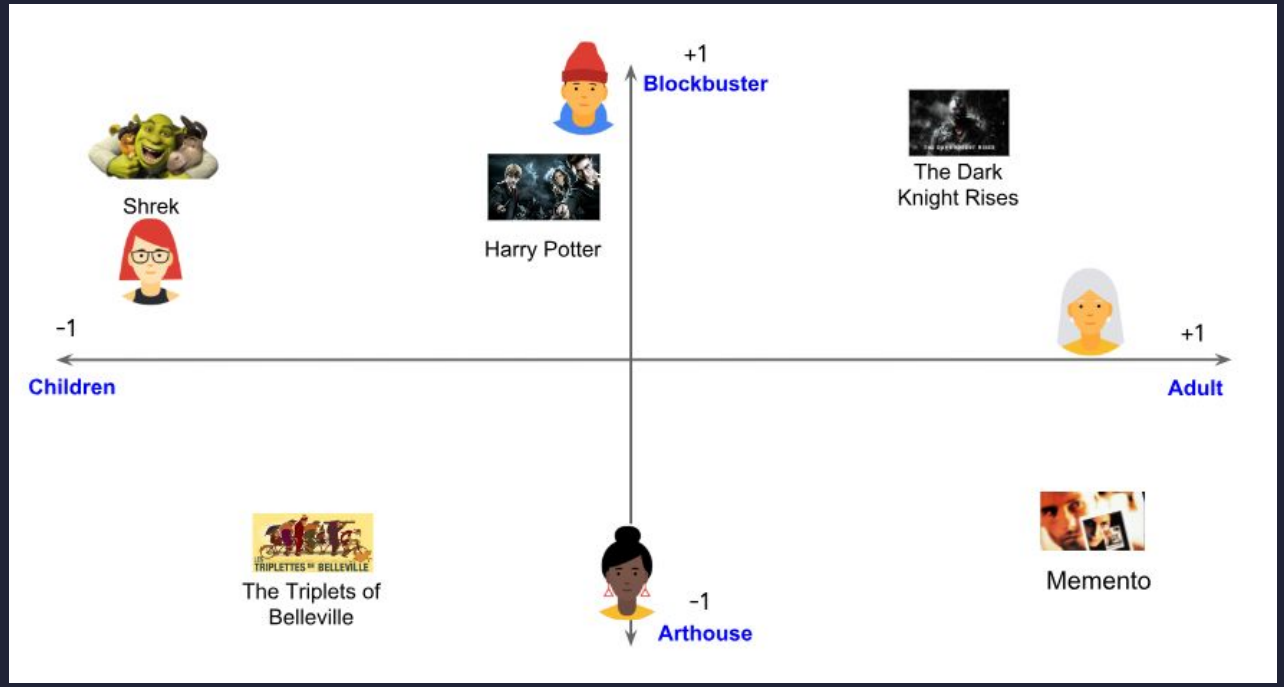


Interaction Matrix



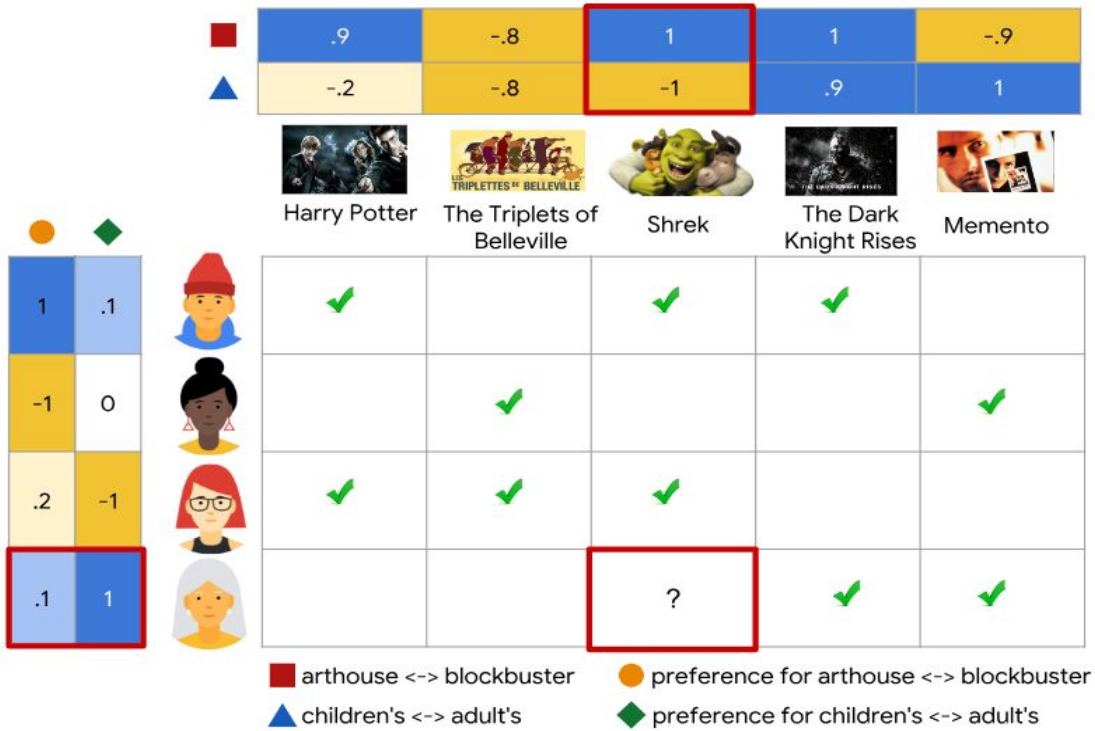
Available from: https://fritz.ai/recommender-systems-with-python/

Latent Features



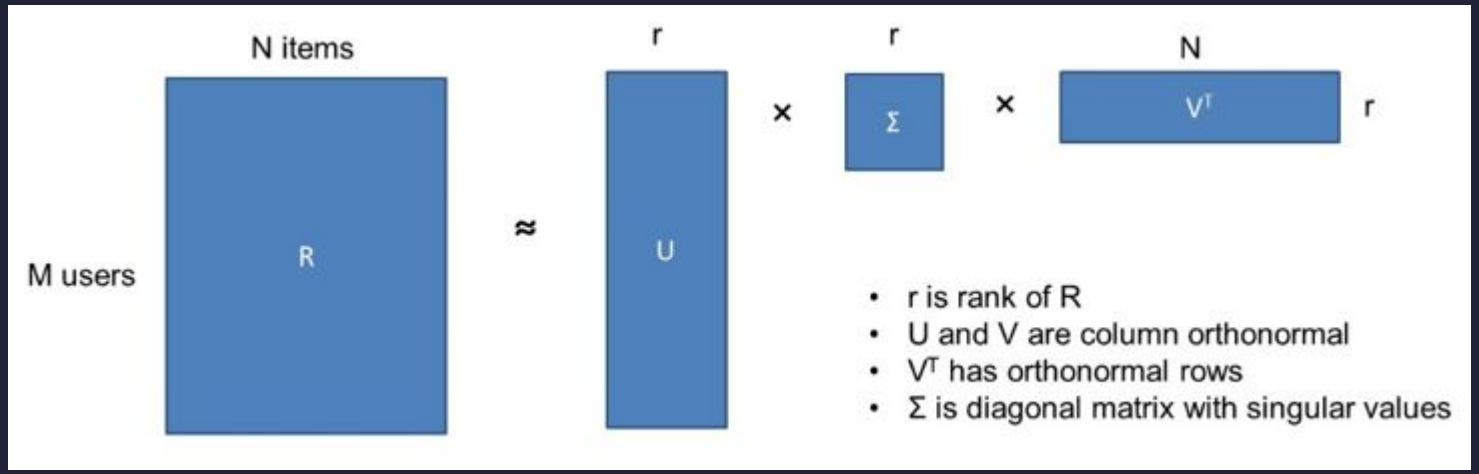
Available from: https://developers.google.com/machine-learning/recommendation/collaborative/matrix

Latent Features



Available from: https://developers.google.com/machine-learning/recommendation/collaborative/matrix

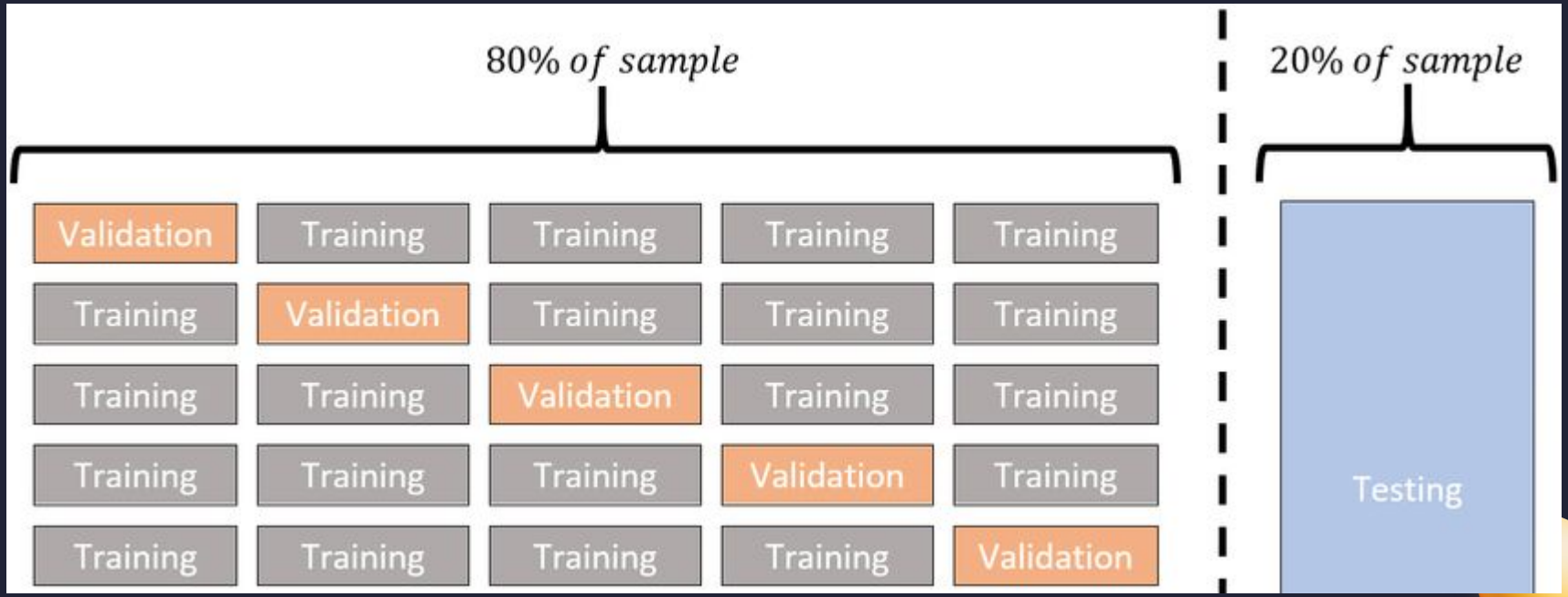
Matrix Factorization (SVD method)



Available from: https://www.dataminingapps.com/2020/02/singular-value-decomposition-in-recommender-systems/

U: Users and their preferences, where each row corresponds to a user and each column to a latent feature capturing preferences.
Σ : A diagonal matrix containing singular values, indicating the importance of each latent feature.
V^T : Books and their characteristics, with each row corresponding to a book and each column representing a latent feature indicating its characteristics.

Evaluation



Available from:
https://www.researchgate.net/figure/Representation-of-a-5-Fold-Cross-Validation-resampling-approach-where-the-model-is_fig1_350878579

05

```
Evaluating RMSE, MAE of algorithm SVD on 5 split(s).

                  Fold 1  Fold 2  Fold 3  Fold 4  Fold 5  Mean    Std
RMSE (testset)    1.6075  1.6208  1.6151  1.6181  1.6078  1.6139  0.0054
MAE (testset)     1.2452  1.2524  1.2469  1.2491  1.2400  1.2467  0.0041
Fit time          1.55    1.53    1.54    1.50    1.54    1.53    0.02
Test time         0.09    0.09    0.20    0.09    0.21    0.13    0.06
RMSE: 1.6194
MAE:  1.2516
```

Results

Average(s)

- RMSE : 1.6139
- MAE : 1.2467
- Fit time : 1.53s
- Test time : 0.13s

06

Limitations

Limitations of the Dataset:

1. Missing data: Absence of age and demographic data limits insights.

2. Limited Feature Set: Few attributes restrict analysis and model performance.

06

Limitations

Limitations of Collaborative Filtering using SVD:

1. Cold-Start Problem: SVD can't recommend new books without prior data.

2. Neglect of Side Features: Ignores user demographics and book attributes.

3. Computational Expense: Resource-intensive, affecting scalability.

07

- SVD Success: The Collaborative Filtering System using SVD can predict user preferences reasonably.
- TF-IDF Challenges: The Content-Based System struggles to match user interests.
- Improvement for TF-IDF: Incorporating additional text features such as book overviews, genres, and content previews could enhance its effectiveness.

Conclusion

Thank You!