

"BOOK BUNDLES RECOMMENDATION FROM BOOK READERS RATINGS AND BOOK SALES DATA"

Data Science Track Program


























Capstone Project 2 : Final Report

Mehmet Oguz

10.01.2019

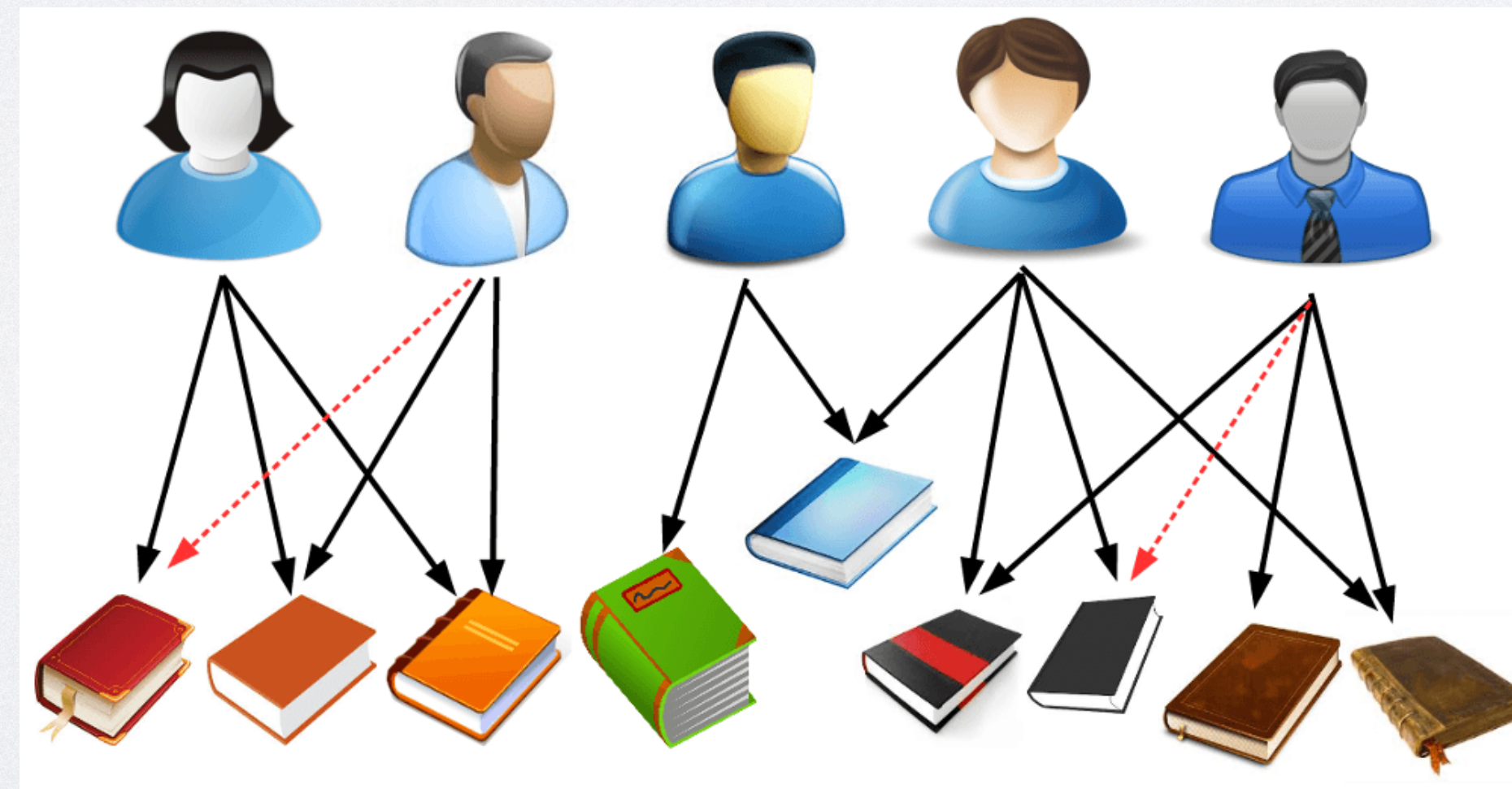
A RECOMMENDATION SYSTEM

A type of information filtering system that predict the rating or preference that a user would give to an item.

This analysis aims to recommend book bundles to the readers from book ratings and book sales data. We all had an online experience where a website makes some personalized recommendations.

- Youtube tells you “...viewers also watch...”,
- Amazon tells you “Customers Who Bought This Item Also Bought”,
- Udemy tells you “Students Who Viewed This Course Also Viewed”.



Data collection

I used explicit data that is provided intentionally by the users as rating. I checked below datasets.

- Cai-Nicolas Ziegler' Book-Crossings dataset;
- Julian McAuley's Amazon product dataset,
- Open library book dataset,
- Worldcat book dataset,
- Goodreaders dataset.

I used updated Goodreaders dataset in my project.

Goodreaders dataset

The data consists of three tables: ratings, books info, and users info.

- The books data set provides book details. It includes 10000 records and 24 fields: book id, Goodreads book id, best book id, work id, books count, isbn, isbn13, authors, original publication year, original title, title, language code, average rating, ratings count, work ratings count, work text reviews count, ratings 1, ratings 2, ratings 3, ratings 4, ratings 5, image url, and small image url.
- The ratings data set provides a list of ratings that users have given to books. It includes 5976479 records and 3 fields: user id, book id, and rating.
- The users data dataset provides the user demographic information. It includes 912705 records and 2 fields: user id, and book id.

Exploratory Data Analysis

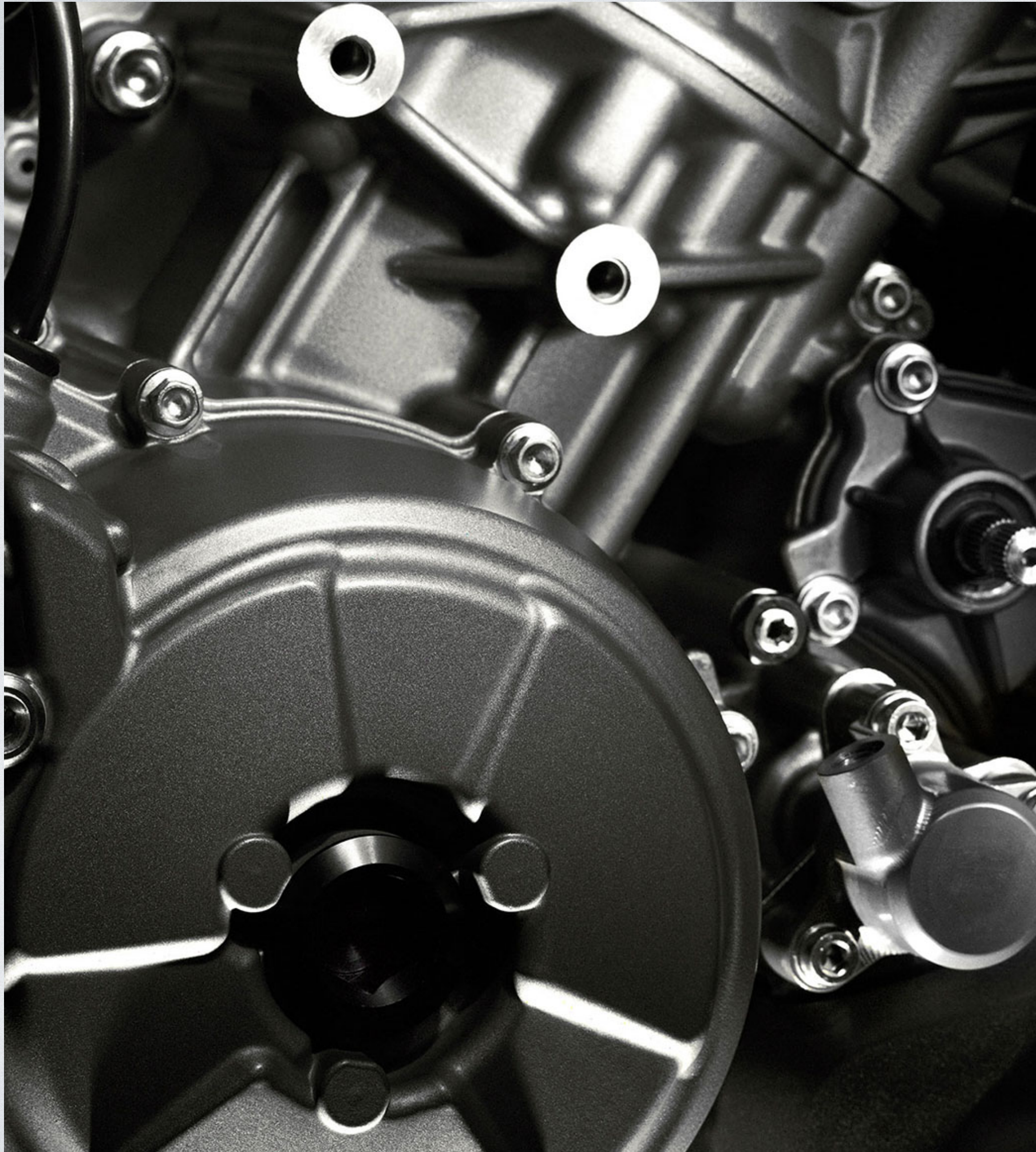
Missing and Wrong Data

- Many books ISBN number is missing (not suitable to fill it with 0 or median value.)
- Negative publication years,
- Very old publication dates.

Negative values filled with the median and left very old dates in database.

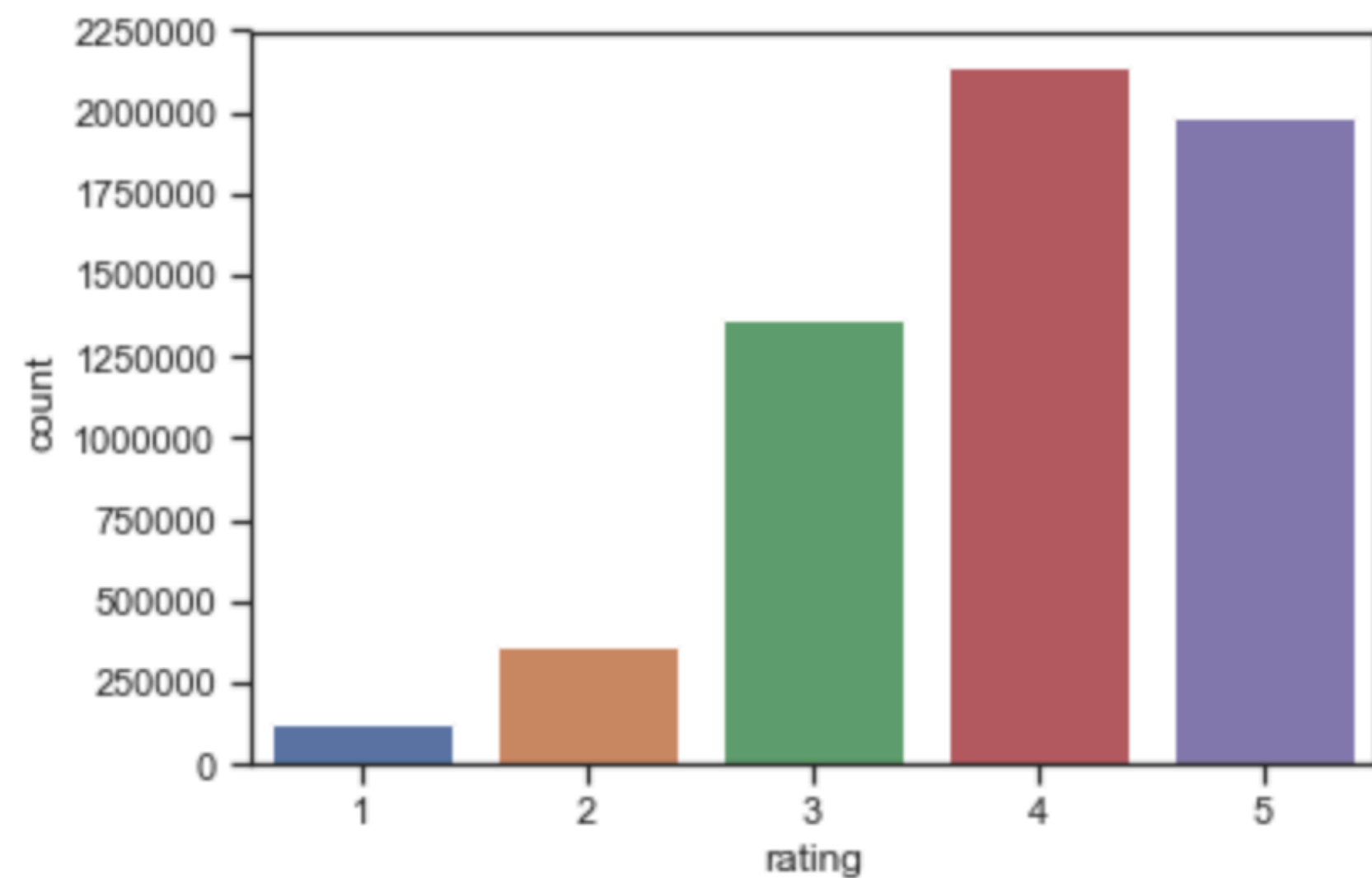
```
# See the column data types and non-missing values  
books.info()
```

```
<class 'pandas.core.frame.DataFrame'>  
RangeIndex: 10000 entries, 0 to 9999  
Data columns (total 23 columns):  
book_id                10000 non-null int64  
goodreads_book_id     10000 non-null int64  
best_book_id          10000 non-null int64  
work_id               10000 non-null int64  
books_count            10000 non-null int64  
isbn                  9300 non-null object  
isbn13                9415 non-null float64  
authors               10000 non-null object  
original_publication_year 9979 non-null float64  
original_title         9415 non-null object  
title                 10000 non-null object  
language_code         8916 non-null object  
average_rating         10000 non-null float64  
ratings_count          10000 non-null int64  
work_ratings_count     10000 non-null int64  
work_text_reviews_count 10000 non-null int64  
ratings_1              10000 non-null int64  
ratings_2              10000 non-null int64  
ratings_3              10000 non-null int64  
ratings_4              10000 non-null int64  
ratings_5              10000 non-null int64  
image_url              10000 non-null object  
small_image_url        10000 non-null object  
dtypes: float64(3), int64(13), object(7)  
memory usage: 1.8+ MB
```

Book ratings and Sparsity Level

```
#plotting count of rating  
sns.countplot(data=ratings_explicit , x='rating')  
plt.show()
```



```
In [27]: print ("number of users: " + str(n_users))  
         print ("number of books: " + str(n_books))  
  
number of users: 912705  
number of books: 10000
```

```
In [28]: #Sparsity of dataset in %  
         sparsity=1.0-len(ratings_new)/float(n_users*n_books)  
         print ('The sparsity level of Book dataset is ' + str(sparsity*100) + ' %')  
  
The sparsity level of Book dataset is 99.93451905051468 %
```

We have 10000 books and 912705 users.

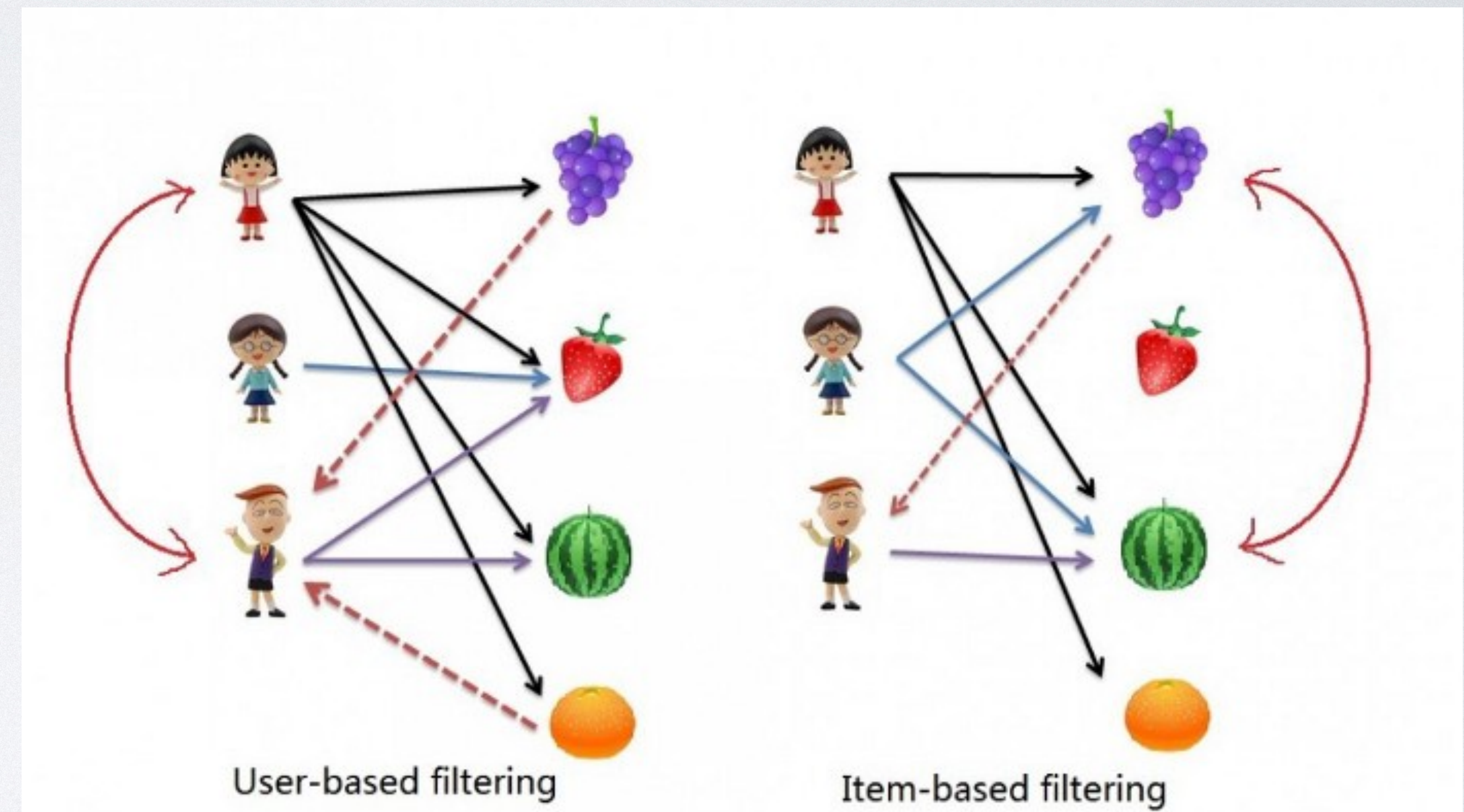
The sparsity level of our Book dataset is 99.9 %.

Number of books that are rated 1-5.

Recommendation system

Nearest neighbor model of collaborative filtering.

Collaborative filtering approach builds a model from a user's past behaviors like items previously purchased or selected and/or numerical ratings given to those items as well as similar decisions made by other users.



KNN MODEL

KNN is a perfect go-to model and also a very good baseline for recommender system development. KNN does not make any assumptions on the underlying data distribution but it relies on item feature similarity. When KNN makes inference about a book, it will calculate the “distance” between the target book and every other books in its database, then it ranks its distances and returns the top K nearest neighbor books as the most similar books recommendations.

Recommendations for Brain Rules: 12 Principles for Surviving and Thriving at Work, Home, and School:

1: A Whole New Mind: Why Right-Brainers Will Rule the Future, with distance of 0.8454638172875433:

2: How We Decide, with distance of 0.8603980043240806:

3: Drive: The Surprising Truth About What Motivates Us, with distance of 0.8629423447204392:

4: Talent is Overrated: What Really Separates World-Class Performers from Everybody Else, with distance of 0.8651073696099224:

5: The Upside of Irrationality: The Unexpected Benefits of Defying Logic at Work and at Home, with distance of 0.8722994549069018:

Perfect! This books are definitely should be recommended one after another

CONCLUSION

A collaborative filtering system does not necessarily succeed in automatically matching content to one's preferences. It requires a substantial number of users to rate a new item before that item can be recommended.

New users will need to rate sufficient number of items to enable the system to capture their preferences accurately and thus provides reliable recommendations.

New items need to be rated by a substantial number of users before they could be recommended to users who have similar tastes to the ones who rated them.

THANK YOU

Mehmet J Oguz
www.linkedin.com/in/mjoguz