Week 06 - Lab Session Results

February 3, 2025

Familiarize Yourself with the Dataset

In the lab sessions, we will work with the "All Beauty" category of the Amazon Review Data, and we will use the 5-core subset. You can download the dataset and find information about it here: https://nijianmo.github.io/amazon/index.html

Exercise 1

Download and import the 5-core dataset.

Exercise 2

2.1

Sort the dataset entries (ascending) by user id (`reviewerID`), product id (`asin`) and rating timestamp (`unixReviewTime`). Then, check the dataset for missing ratings (`overall`) and duplicates (cases where the same user has rated the same item multiple times) and clean them, if any. For duplicates, keep the last entry only. How many observations does the cleaned dataset have?

```
Observations in the cleaned dataset: 4092
```

2.2

Create a test set by extracting the *single* **latest** (based on the timestamp `unixReviewTime`) positively rated item (rating ≥ 4) by each user. Remove users that do not appear in the training set. How many observations does the training and test set have?

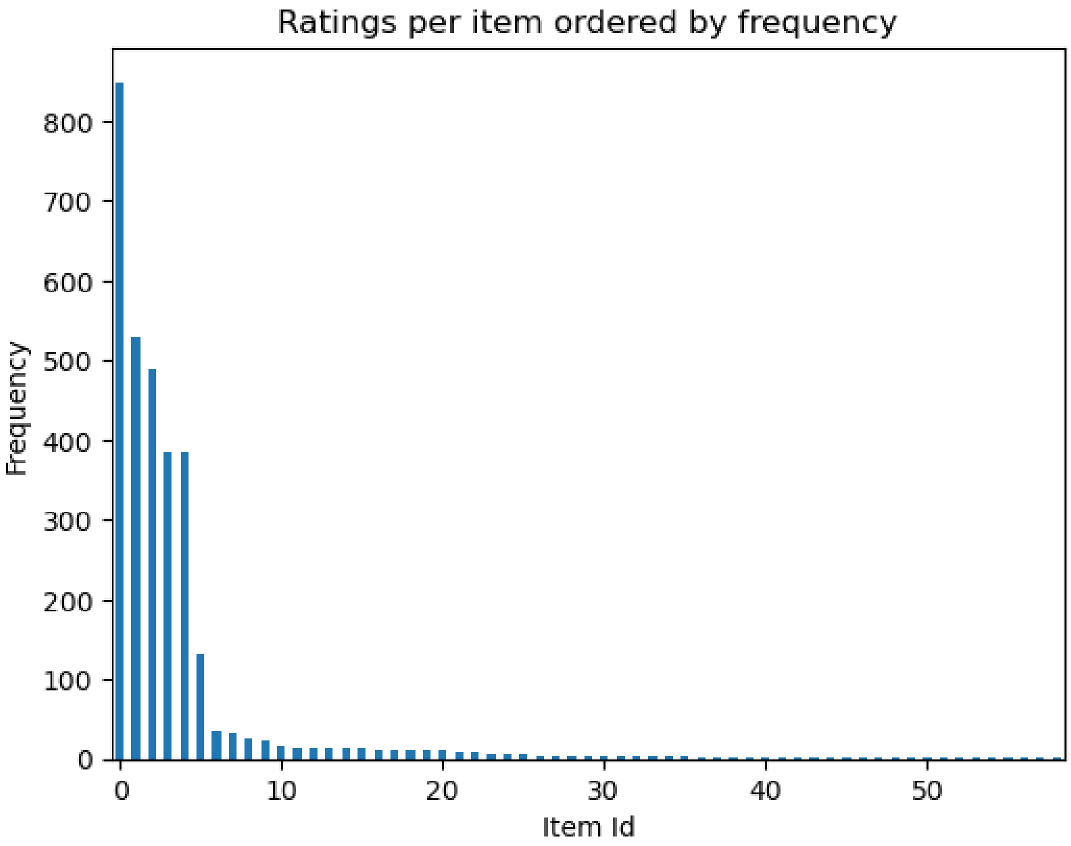
Note: After the test set is extracted, the *cleaned* dataset is now the trainset. (i.e., the trainset is completely separate from the test set, not a superset of it)

```
Observations in training set: 3133
Observations in test set: 949
```

Exercise 3

Compute the number of ratings per item in the training set. How does a barplot of the number of ratings ordered by decreasing frequency look like?

Reflect on how it will affect the prediction process of a recommender system if only a small fraction of the items are rated frequently.



Collaborative Filtering Recommender System

Exercise 1

In this exercise, we are going to predict the rating of a single user-item pair using a neighborhood-based method.

1.1

- Represent the ratings from the training set in a user-item matrix where the rows represent users and the columns represent items.
- Fill unobserved ratings with 0.
- Compute the cosine similarities between the user with `reviewerID`=‘A25C2M3QF9G7OQ’ and all users that have rated the item with `asin`=‘B00EYZY6LQ’.

- What are the similarities and what are the ratings given by these users on item ‘B00EYZY6LQ’?

```
                cosine similarity  overall
reviewerID
A1F7YU6O5RU432           0.079243      5.0
```

```
A1R1BFJCMWX0Y3           0.245145      3.0
A1UQBFCERIP7VJ           0.058634      5.0
A22CW0ZHY3NJH8           0.207883      3.0
A2LW5AL0KQ9P1M           0.275810      4.0
A2PD27UKAD3Q00           0.000000      5.0
A2WW57XX2UVLM6           0.000000      4.0
A2ZY49IDE6TY5I           0.682835      4.0
A39WWMBA0299ZF           0.000000      5.0
A3M6TSEV71537G           0.000000      5.0
A3S3R88HA0HZG3           0.000000      4.0
A914TQVHI872U            0.245145      5.0
AOEUN9718KVRD            0.105670      3.0
```

1.2

Predict the rating for user 'A25C2M3QF9G7OQ' on item 'B00EYZY6LQ' based on the ratings from the 5 most similar users, using a weighted (by cosine similarity) average. You do not need to account for the mean rating per user for the weighted average. What is the prediction? Round the predicted rating to 3 decimal places.

```
Predicted rating: 3.875

Top 5 most similar users:
```

```
[10]:                 cosine similarity  overall
      reviewerID
      A2ZY49IDE6TY5I           0.682835      4.0
      A2LW5AL0KQ9P1M           0.275810      4.0
      A1R1BFJCMWX0Y3           0.245145      3.0
      A914TQVHI872U            0.245145      5.0
      A22CW0ZHY3NJH8           0.207883      3.0
```