(a)

```
In [5]: items.count()

Out[5]: 1701

In [7]: users.count()

Out[7]: 27555
```

the number of items:
1701
the number of users:
27555

The average overlap of **users** for **items** in the test set with items in the training set is **higher** than the average overlap of **items** for **users** in the test set with users in the training set.

(b)
We decided to use item-item similarity models to implementing the collaborative filtering.The user-based model has some limitations. One is its difficulty in measuring the similarities between users, and the other is the scalability issue. As the number of customers and products increases, the computation time of algorithms grows exponentially. The item-based model was proposed to overcome the scalability problem as it calculates item similarities in an offline basis. It assumes that a user will be more likely to purchase items that are similar or related to the items that he or she has already purchased.

(c)
The similarity between two items is measured by computing Cosine similarity. Denoting the set of users who both rate x and y, the correlation similarity is given by

$$C(\mathbf{x}, \mathbf{y}) = \cos(r_x, r_y) = \frac{\boldsymbol{r}_x^{\mathrm{T}} \cdot \boldsymbol{r}_y}{||\boldsymbol{r}_x|| \cdot ||\boldsymbol{r}_y||}$$