Team Espresso

Kim Pham, Zhuqian Zhou, Aidi Bian, Sihan Wang

Defining the Problem

Right now we know the **most popular hotels by total clicks and rankings**.

However, we do not know most popular hotels by individual types of users.

If we want to reach a wider audience, we must assume there will be different types of users. If we can figure out those different types of users, we can learn to optimize the best hotel recommendations or **most popular hotels by user**.

What hotels should be recommended to what type of users?

Our Solution

1. Identify types of users (cluster)

2. Identify types of hotels (cluster)

3. Build item-based recommender system, based on collaborative filtering

Data Pre-Processing

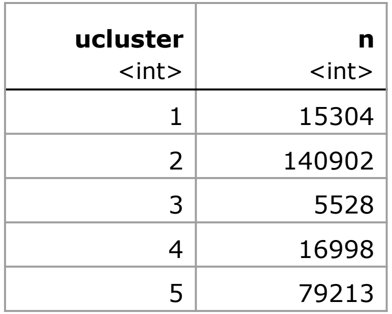
Data Pre-Processing

1. Import both TripAdvisor datasets and join datasets by “hotel_id”
2. Split into test and train data sets by dates before and after "2019-1-20"
   a. Why time? We need past data as the training set and future data as the test.
   b. Why “2019-1-20”? We need a relatively larger training set (1,050,306 obs.) versus the test set (108,295).
3. On train dataset:
   a. Calculate:
      i. Total Clicks = click_booking + click_hotel_website + click_price
      ii. We don’t count the “click_view” since this is not a profitable action
   b. Scale

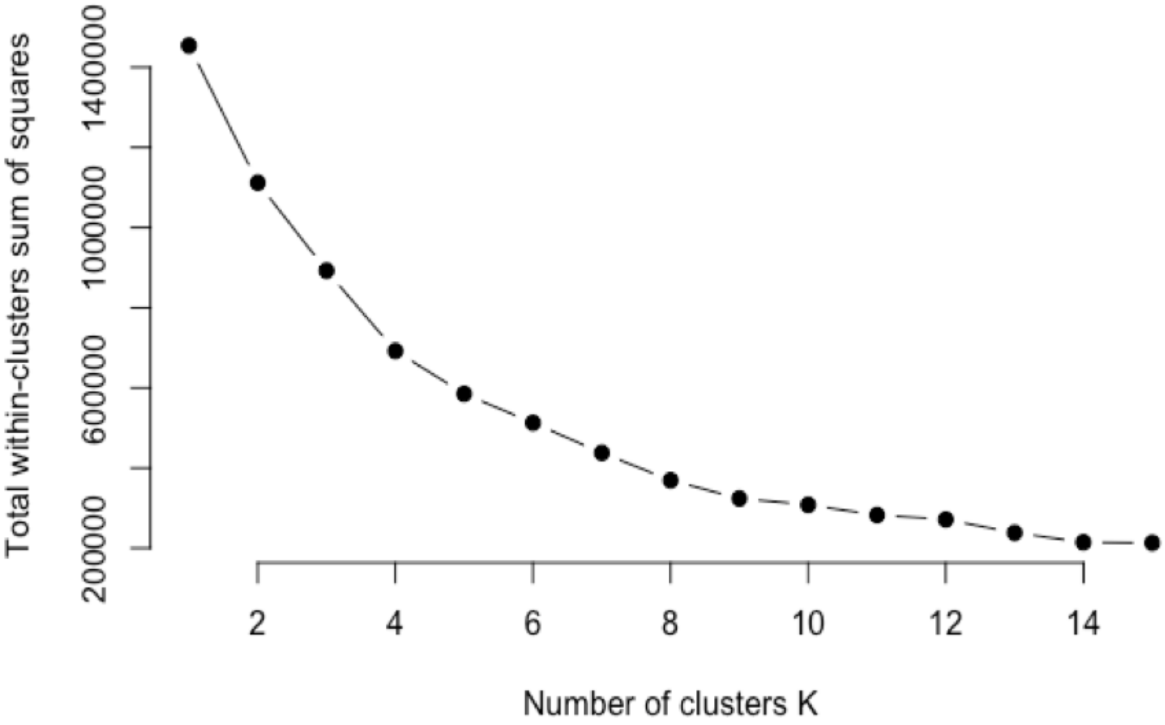
User Clusters

User Clustering

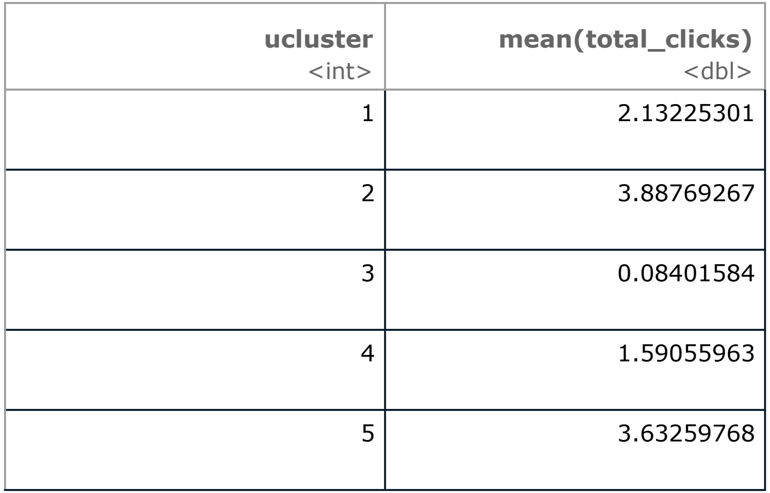
kmeans, cluster by k=5



ucluster <int>	n <int>
1	15304
2	140902
3	5528
4	16998
5	79213



Mean Total Clicks by Cluster

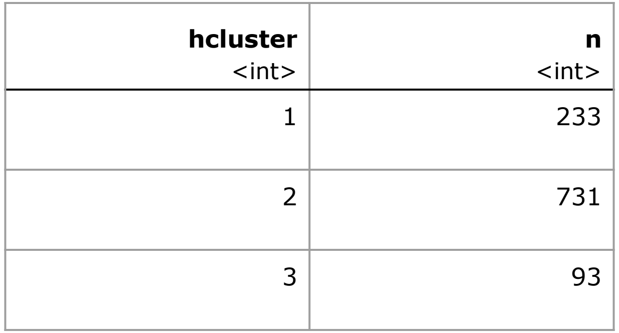


ucluster <int>	mean(total_clicks) <dbl>
1	2.13225301
2	3.88769267
3	0.08401584
4	1.59055963
5	3.63259768

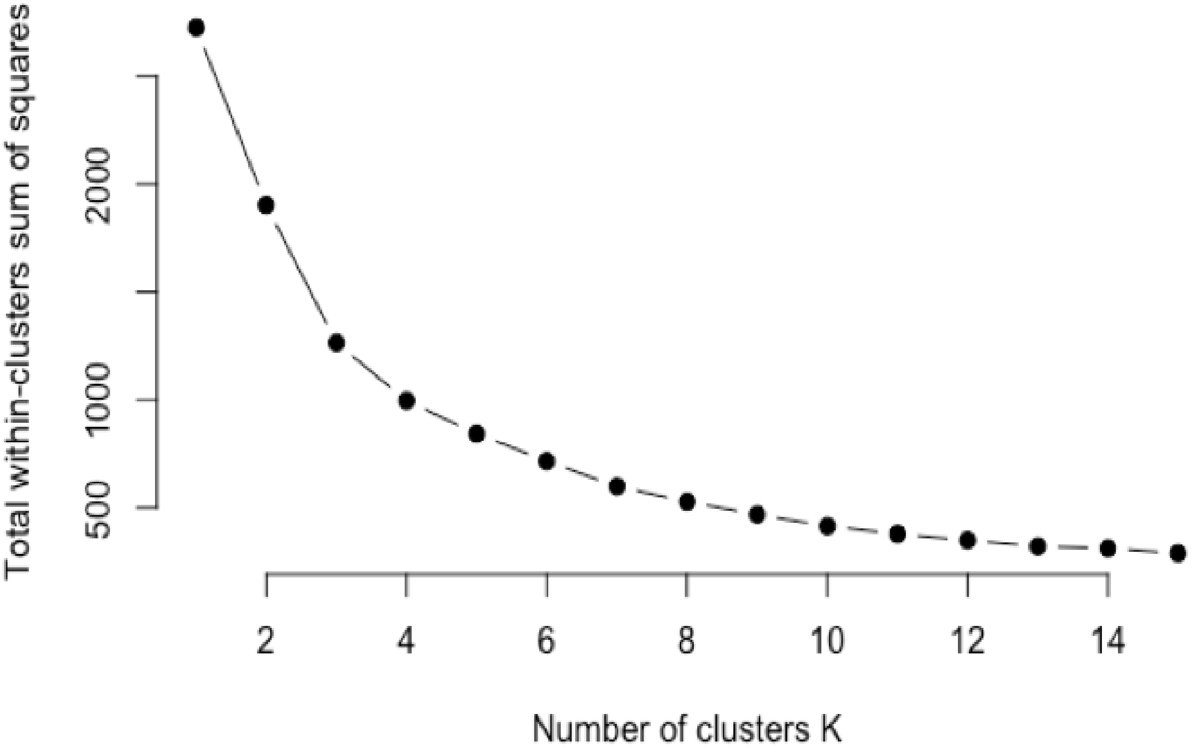
Hotel Clusters

Hotel Clustering

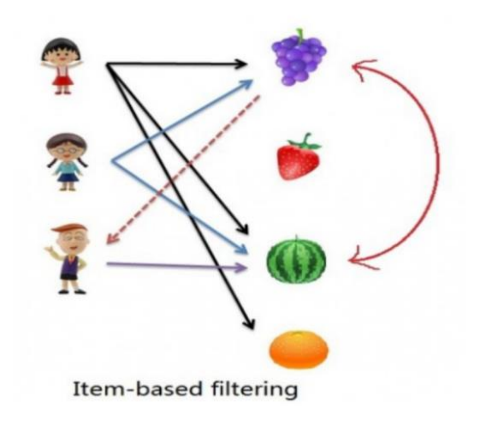
kmeans, cluster by k=3

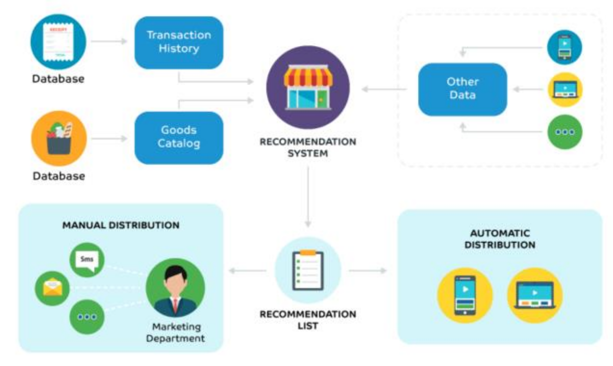


hcluster <int>	n <int>
1	233
2	731
3	93



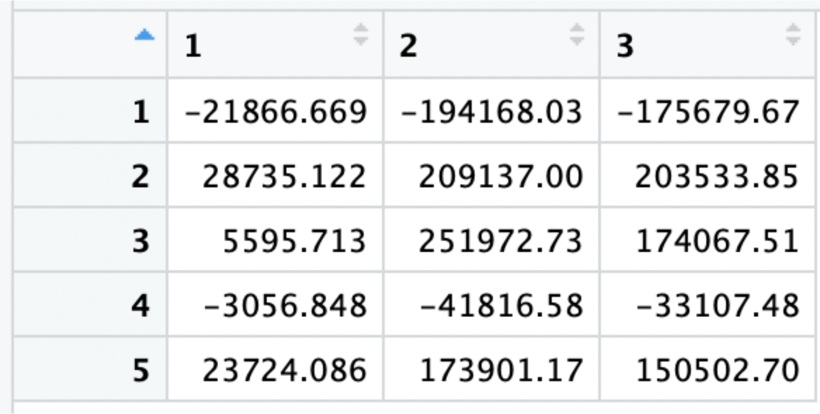
Build Item-Item Recommender System



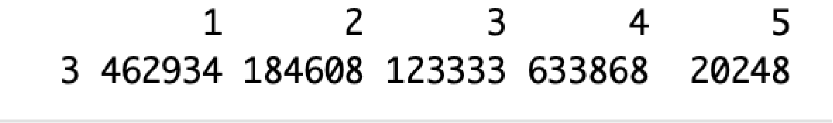


User-Item Matrix

Number of user clusters = 5
Number of hotel clusters = 3

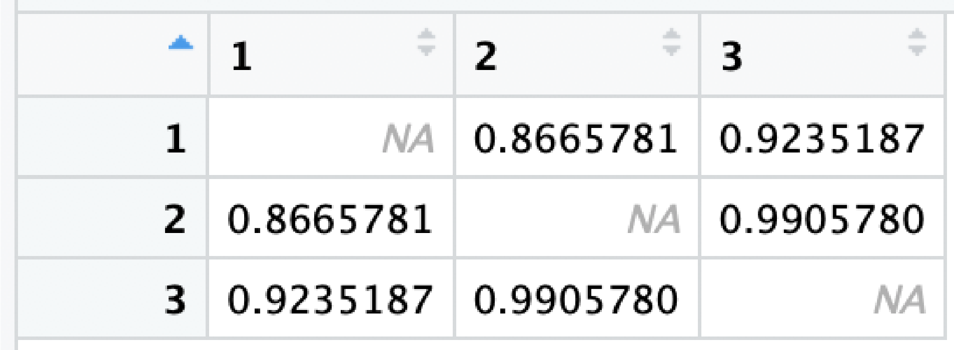


	1	2	3
1	-21866.669	-194168.03	-175679.67
2	28735.122	209137.00	203533.85
3	5595.713	251972.73	174067.51
4	-3056.848	-41816.58	-33107.48
5	23724.086	173901.17	150502.70



```
       1      2      3      4      5
3 462934 184608 123333 633868  20248
```

Item-Item Similarity Matrix

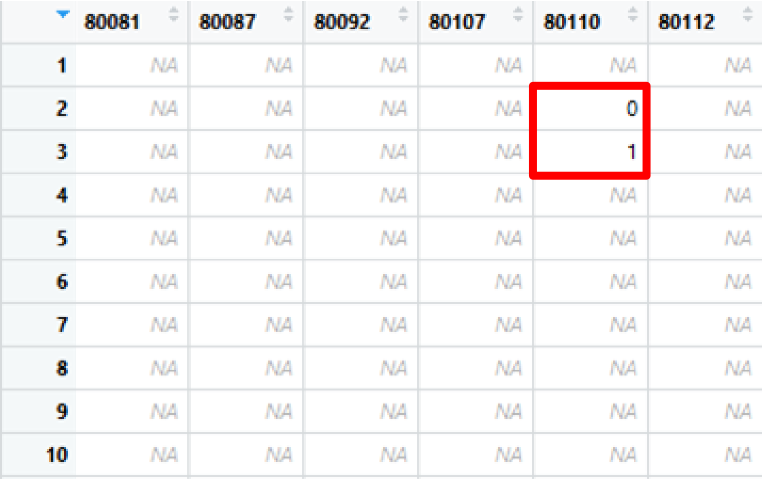


	1	2	3
1	*NA*	0.8665781	0.9235187
2	0.8665781	*NA*	0.9905780
3	0.9235187	0.9905780	*NA*

Accuracy=0.0067=12333/(462934+184608+123333+633868 +20248)
The accuracy is very low.
(May due to the limited number of hotel clusters)

User-Item Matrix

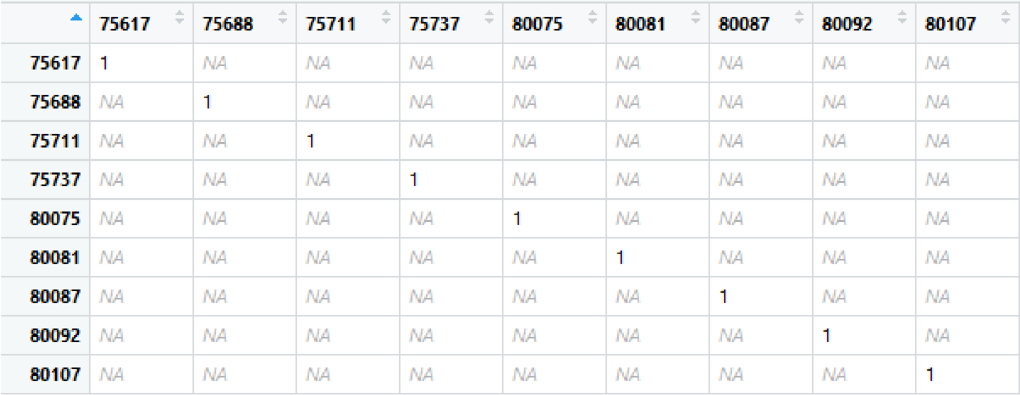
No clusters at all (a lot of N/A !)



	80081	80087	80092	80107	80110	80112
1	NA	NA	NA	NA	NA	NA
2	NA	NA	NA	NA	0	NA
3	NA	NA	NA	NA	1	NA
4	NA	NA	NA	NA	NA	NA
5	NA	NA	NA	NA	NA	NA
6	NA	NA	NA	NA	NA	NA
7	NA	NA	NA	NA	NA	NA
8	NA	NA	NA	NA	NA	NA
9	NA	NA	NA	NA	NA	NA
10	NA	NA	NA	NA	NA	NA

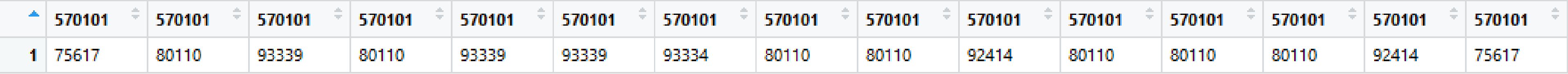
Item-Item Similarity Matrix

Still a lot of N/A !



	75617	75688	75711	75737	80075	80081	80087	80092	80107
75617	1	NA	NA	NA	NA	NA	NA	NA	NA
75688	NA	1	NA	NA	NA	NA	NA	NA	NA
75711	NA	NA	1	NA	NA	NA	NA	NA	NA
75737	NA	NA	NA	1	NA	NA	NA	NA	NA
80075	NA	NA	NA	NA	1	NA	NA	NA	NA
80081	NA	NA	NA	NA	NA	1	NA	NA	NA
80087	NA	NA	NA	NA	NA	NA	1	NA	NA
80092	NA	NA	NA	NA	NA	NA	NA	1	NA
80107	NA	NA	NA	NA	NA	NA	NA	NA	1

Recomm. (but haven’t got to the part of prediction yet)



	570101	570101	570101	570101	570101	570101	570101	570101	570101	570101	570101	570101	570101	570101	570101
1	75617	80110	93339	80110	93339	93339	93334	80110	80110	92414	80110	80110	80110	92414	75617

Limitations

- To explore more methods that can reduce the challenges
- To provide recommendation in collaborating filtering a wider range of applications
- Consider the quality and privacy aspects
- The recommendation accuracy of clustered hotel data is very very low (less than 1%); We also tried to build a model with un-clustered hotels, but did not get the final cross validation matrix due to the big amount of data (R broke down when running this)

What we learned

We learned . . .

- clearly defining problems and outlining solutions helps us keep on track
- about different recommender systems (item-based versus user-based)