Project Proposal:

Collaborative filtering challenge for restaurant recommendation

Background

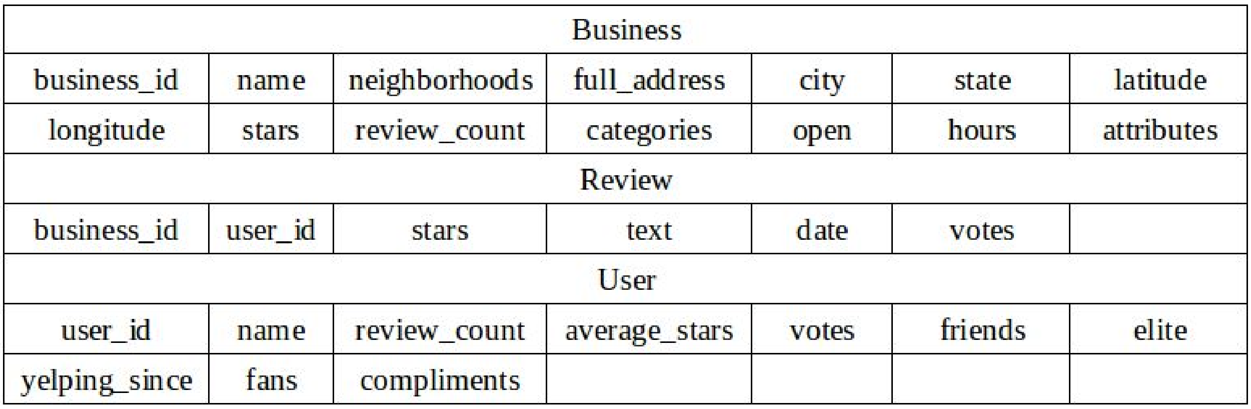
Nowadays automatic recommendation system takes a more and more important role in online commerce while it is still a new domain. Encouraging researches in this domain will not only help us understand better the customers and their behaviors but also help companies make good decisions. Participants can have a deeper comprehension of collaborative filtering and have knowledge on how the recommendation system in our daily life works. This project is about designing a restaurant recommendation system with the help of Yelp datasets.

Material and method

Datasets : yelp datasets available here **(link)**
Size of original data: **2.2M** reviews and **591K** tips by **552K** users for **77K** businesses.
Available dataset are : business data, users data and review. **(link)**

While not all the data will be used, we'll firstly preprocess these data and extract a training set, a development set and a test set from original data. Maybe one problem we will meet is how to find the true ratings for all(or lots of) restaurants for all users as to evaluate the result of participants. The possible solution is finding a large amount of users who all have rated lots of restaurants, the intersection must be quite large. Then one work in preprocessing is to find this subset of users and restaurants and to cover up some(10% according to BellCore CHI 1995) of the truths for participants to predict.

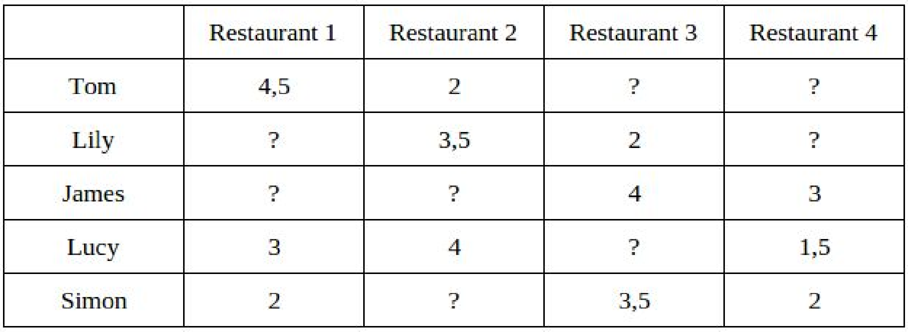
We intend to use all features which will make the dimensionality quite big, but it's participants' job to reduce the dimensions using for example information gain and find out those most helpful features. The features in data is like Table 1 below:



Business						
business_id	name	neighborhoods	full_address	city	state	latitude
longitude	stars	review_count	categories	open	hours	attributes
Review						
business_id	user_id	stars	text	date	votes	
User						
user_id	name	review_count	average_stars	votes	friends	elite
yelping_since	fans	compliments				

Table 1. Features in each dataset

The main task is to create a model for collaborative filtering based on users reviews datasets collected over Yelp plate form in many countries. The participants will have to think out all possible ways to mine useful information in the available data to create their model. The objective is to predict all ratings for every user on the restaurants which the user haven't rated and then pick up top 10 restaurants with highest rates to propose it to this user. Table 2 shows the main problem in this challenge.



	Restaurant 1	Restaurant 2	Restaurant 3	Restaurant 4
Tom	4,5	2	?	?
Lily	?	3,5	2	?
James	?	?	4	3
Lucy	3	4	?	1,5
Simon	2	?	3,5	2

Table 2. Main problem is to predict the values of those question marks

For the quantitative metrics of assessment, we will evaluate results according to their accuracies of ratings on restaurants which we have covered up of each users with calculating mean squared error as final results.

Preliminary results

Recommendation problem using collaborative filtering is a little different from standard machine learning problem. We can't use general supervised learning algorithm to solve it as there are too much data to predict while we have much less data to train a model. That's the difficult part of this challenge, we have to dig into dependencies between users and restaurants and make good use of similarities and correlations among them to make predictions. So a little knowledge of collaborative filtering is required.

Previous Yelp recommendation system experience can be found in paper "Yelp Recommendation System Using Advanced Collaborative Filtering"[1]. He performed a similar technique on Yelp dataset to predict latent rating of users on restaurants. He used root mean squared error to measure the results, he used a baseline method implemented by Yehuda Koren[2]. His final results are in Table 3 below:

Algorithme	RMSE
Baseline	1,4742
knn	1,5274
knn with restaurant clustering (Type of food)	1,3091
knn with restaurant clustering (Type of food and style)	1,1235
SVD	1,3822

Table 3. Previous experience by Chee Hoon Ha