Project proposal from Xiaodong:

1. Learn belief propagation algorithm and apply it on review evaluation for users. After data wrangling, we have multiple tables with all information we need. Then we extracted each user’s evaluation in one category, for example, restaurant. We will do one step further analysis on subcategory, but for now we will just work on category. We have two simple ideas about how to match his reviews with his given stars. We have all the data now, but before proceeding to the data correction, we need to sort out those users table and evaluation tables where the user gives a non-match evaluation with the score.
2. This one is simply taking the average of one user’s evaluation scores, regard it as a baseline, added to what value this user gives to each restaurant he has visited. The new score will be regarded as this user’s evaluation to this restaurant. We had this simple idea because say, somebody likes to give low scores to others, most of whose score is 1 or 2. But sometimes he gave one Chinese restaurant a 4 star. Then this restaurant must be very satisfying his/her taste or he has entered a wrong value. To eliminating the possibility of wrong input, we will plot each user’s evaluation histogram, getting his RSM (root square mean), tell where his anomaly values stay and judge that. After that, we will update his score for each restaurant with the weighted original value plus his baseline value. (For example, this Chinese restaurant will be ceil (4+1) = 5 as his evaluation score).
3. Another way we are working on is the belief propagation algorithm suggested by Prof. Paris. This BP algorithm is calculating belief after updating messages until convergence. We have already generated word-clouds for different review stars. Say, we have word-cloud 1, 2, 3, 4 and 5 regarding with words in reviews of corresponding stars. Then we regard them as nodes and connect all of them to a user’s through DAG. The condition to use BP is there is no loops in graph. So now we need to calculate each message passed in BP to this user. Outside the word-clouds, we have all other attributes related with reviews, like restaurant information (location, type, served dishes….). So what we want to calculate is the influences on a user’s review based on all possible attributes.
4. About Spark, we have already installed Spark on our own laptops, spent a little time studying how Spark works. We plan to work on standalone mode first, test all the queries we have processed in MySQL part and see the performances like runtime. We are currently working writing Python scripts to process RDDs. Next stage we plan to test them on multiple machines, as we have got all processed data before, we can distribute them among different machines, each machine has part of data, we will use 2 or 3 threads to run part of queries we have run before and test again the performances. This is interesting as we would have chance to experience distributed system provides faster than normal MySQL.
5. About trend we want to predict: there are several things interesting to see. First one is how one restaurant would go, usually a newly open restaurant tend to provide less satisfaction to users as it is not experienced in managing. But there could be possibility that they lost enthusiasm as well it has run for a while. So it is very interesting to judge from custom’s review about this. Also, location is very important to run a restaurant business. How location influences business is fun to see from analysis of data. Third, seeing a review history from a user is fun. Some users will give higher and higher evaluations as he/she feels that previously he has given too tough reviews. So he/she might trend to give others higher and higher scores. All this can contribute to set up an interesting model for future analysis.