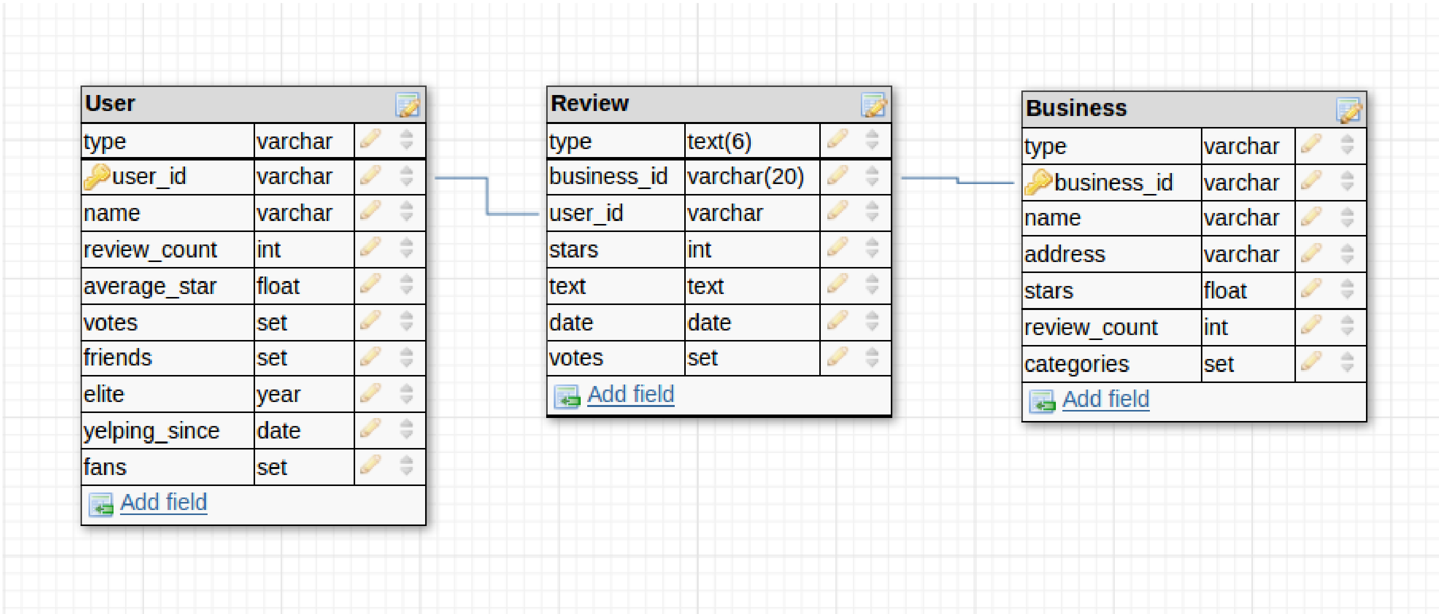
WEEKLY ISSUES REPORT 2
(10/03/2014)

Objectives:

- Select the most 3 promising features (give more priority on features that can be used in both SVM and Decision tree) add them to the list of current features “one by one” to compare the results.
- Read the book more deeply.
- Try random forest with 50, 100, … to see the differences.
- Discuss about the topic “Hidden Factors and Hidden Topics: Understanding Rating Dimensions with Review Text.” advertised on Yelp.
- …

What should we do? Follow the objectives discussed or concentrate on the paper you’ve just proposed.

Database diagram:



User	
type	varchar
user_id	varchar
name	varchar
review_count	int
average_star	float
votes	set
friends	set
elite	year
yelping_since	date
fans	set

Review	
type	text(6)
business_id	varchar(20)
user_id	varchar
stars	int
text	text
date	date
votes	set

Business	
type	varchar
business_id	varchar
name	varchar
address	varchar
stars	float
review_count	int
categories	set

List of potential features:

# No.	Name	Description	Entity
1	Review’s Length	The length of the reviews	**Review**
2	Review’s Votes	There are 3 kind of votes: funny, useful and cool, the number of each one indicated how many votes for that one	**Review**
3	Review’s Date	(e,g. '2012-03-14'), we can classify the day into 2 group: weekday (0) and weekend (1)	**Review**

4	User’s Number of Reviews	How many review a particular user wrote?	User
5	User’s Average Star	(e,g. 4.31), the average rating star of an user of all reviews	User
6	User’s Votes	Similar to votes for review but for users	User
7	Number of friends	How many friend this user have?	User
8	Number of fans	How many fan this user have?	User
9	Density of reviews	The rate (#reviews/time) indicate the frequency this user make review	User
10	Business’ Avg. Rating Star	The average rating for this business	Business
11	Business’ Review count	How many reviews for this business	Business
12	The categories of the business	Based on this category and the review count we can recognize if this business is good or not compare to other business on the same category.	Business

There are 3 possibilities about 3 features that we want to propose:

1. We choose (Review’s Length, Review’s Votes, Review’s Date) because these information are available on the Review entity, we don’t need to access to other entities for the reference.
2. We choose (Review’s Votes, User’s Average Star, Business’ Avg. Rating Star) because these information highly represented the quality of the review as well as the quality of the product.
3. Another possibility is we use the votes for the review to add the weight to the features (by multiplying the number of good votes with the current histogram), after that we build the model for each category (Food”, “Service”, “Ambience”, etc.) we have, because the characteristic of each category is different. And the three features we choose after filtering are: