3.1 Deliverable Descriptions

Team Formation: List each team member, along with their department and degree program

Problem identification: Find (or create) a dataset of interest to you. Provide a detailed description of the data. This should be at least 1 page in length. The description may include:

• Number of discrete data objects

• Format (csv, Excel, JSON, XML, JPG, ...)

• Source

• Reason why the data is interesting from a user’s perspective

• Reason why the data is interesting from a database perspective

• High level description of any other work / analysis done on this data set

Yelp Dataset

<https://www.yelp.com/dataset>

Yelp Dataset Documentation

<https://www.yelp.com/dataset/documentation/main>

Yelp ideas:

1. Restaurant categories (Category, restaurant name(ID))
2. Restaurant location (State, city, neighborhood, restaurant name)
3. Category with time combination (at what time people prefer what food)
4. Query abnormal user rating count (percentage) -> guess Yelp rating on restaurant (refer to IMDB rating algorithm)
5. At each time which type of restaurant is most popular
6. For each neighbor in a specific city which type of restaurant is most popular
7. Gender distribution?
8. Calculate review and tip amount for each user -> relate to gender
   1. Man -> more tip
   2. Woman -> more review
9. Calculate the average useful/funny/cool votes for each user
10. Count the frequency for each user go back to a restaurant

ABNORMAL TABLE

User | Average Rating | Since | Rating

**1 Problem Description**

Yelp publishes crowd-sourced reviews on local business, such as restaurants, which heavily influence customers’ decision making, and in turn business’s performance. However, the published reviews, along with the ratings, are not informative enough, often providing misleading messages and hurts participants’ interests. Our team focuses on producing more informative Yelp restaurant reviews and ratings by analyzing the dataset from different perspectives. We used the Yelp open data set which is a subset of Yelp’s restaurants, reviews, and user data (<https://www.yelp.com/dataset>). We extracted the restaurant categories (category, restaurant name(ID)), restaurant location (State, city, neighborhood, restaurant name), category with time combination (at what time people prefer what food), reviews details, and users information. We plan to query abnormal user rating count (percentage) to deduce Yelp’s rating algorithm on restaurant (similar to IMDB rating algorithm). We also want to calculate review and tip amount from each user and relate these data to the users’ gender, and we want to calculate the average useful/funny/cool votes for each user and count the frequency for each user go back to a restaurant. By do the through analysis of the datasets, we hope to correct the misleading official information and provide more insightful ratings for Yelp users.