Stage I Report

Chang Guo (cguo42@wisc.edu)
Yuncong Hao(hyuncong@wisc.edu)
Qun Zou(qzou5@wisc.edu)

1. The set of data sources that you have selected.

a) Data source: Los Angeles restaurants from Yelp and TripAdvisor for the structured datasets.

Our structured table including the following attributes for both of the data from the two sources:

- Zip code of the restaurant
- Phone number of the restaurant
- Rating of the restaurant
- Average price of the restaurant
- Number of reviews have been published for this restaurant

b) The top three user reviews for each of top 100 restaurants in Los Angeles from Yelp for our text data files.

2. A description of how you have extracted structured data from the two data sources.

We used Python selenium package for data crawling and Python CSV package for structured data output. We used the methods included in selenium browser automation library to open the specific websites we want and identify the starting character for HTML structure such as <p>, <class>, <h> for the data information we want and output the structured restaurant data as .csv file, with the reviews as .txt file.

3. The set of questions that you want to answer.

1. The most popular type of foods in Los Angelas.
2. The relationship between rating and average price of the restaurant.
3. Find the area in Los Angelas where users like to eat most.
4. Compare the ratings for the same restaurant on Yelp and TripAdvisor.

4. What is it that you want to extract from the text documents.

1. Overall descriptive word like “great”, “good”, “bad” word from the text to predict user’s rating score based on their text review.
2. Characteristic word for describing food, such as “tasty”, “spicy”, “service”, “looks good”, “atmosphere”, “parking” to analyze what characteristics users concern most for a good restaurant and what characteristics a restaurant should avoid in order to not get low ratings.

3. Coupon name like “Groupon” to find the relationship between user ratings, restaurant qualities and discount.

5. The names of open-source tools you have used in this project stage and a brief description of what they do.

We used Python selenium webdriver package and csv package for data extraction.

a) The selenium package is used to automate web browser interaction from Python. It allows us extract the data we want from the two website.

b) The CSV module implements classes to read and write tabular data in CSV format. It allows us to output the structured restaurant data we crawled in CSV format.