Salman Arif
Andrew Chuang
Jon Hori
Ryan Teehan

CS 122 Project Proposal
1/24/17

For this project, we plan to analyze data obtained from Yelp, both through the Yelp Dataset Challenge data as well as through web-crawling the Yelp website to extract information of interest. We will be focusing on restaurants in several major urban cities for this project (locations TBD). The goal of our project is to create a recommendation platform that can provide recommendations for restaurants based on restaurants the user already likes and dislikes. In addition, we plan to do some further analysis on a city by city basis or a neighborhood basis (within Chicago) to analyze some questions of interest. These include: how long does it take for restaurants to become popular in different neighborhoods/cities? Do different cities and neighborhoods have preferences for certain types of restaurants based on cuisine or other factors of interest?

The first task to complete will be to study the HTML of the Yelp website to determine how we will go about crawling for data. This may be difficult as Yelp shows restaurants either by search terms or by location, so we will have to be sure that we are creating a comprehensive dataset that includes all types of restaurants and also includes restaurants with low ratings. We plan to first focus on Chicago before possibly crawling for data on other cities. Additionally, we will likely need to create a Restaurant class that will hold all of the relevant information on each restaurant. Some of this information will include: location, neighborhood, price range, cuisine type, hours, ambiance, size, neighborhood, among other attributes of interest. We hope to have this done within the next week.

Following this, we will begin writing code to scrape the website, construct our restaurant class, and also determine what sort of algorithm we hope to use to make recommendations. Once we have an idea of what data we will have, we can determine how we want to connect different restaurants. This will likely require being able to find which users have reviewed different restaurants, and determine if there are patterns between restaurants. This should occur by the end of 7^{th} week. Following this, we may be able to expand our data set to include cities outside of Chicago, and add more restaurants.

After this, we will begin building the actual platform for users. Users will be able to input restaurants they enjoy, as well as characteristics such as price range, cuisine, and ambiance. This will result in a list of recommended restaurants, either within Chicago or in the city of their choosing if we choose to expand outside of Chicago. Additionally, if the user is looking for recommendations within Chicago, they can choose specific neighborhoods to look in.

Finally, part of the platform will include general information by neighborhood/city on what trends seem to emerge for that location or any other interesting points we find.