Northeastern University
CS6220 – Data Mining Techniques
November 2017, Emily Dutile, Vyshaal Narayanam, Xiwen Song, Yu Tian

Update 1

1. Problem Statement and Background

As stated in our proposal, we look to address the following to questions:
(1) What topics are discovered frequently in reviews and do they correlate to a positive or negative review? What should a restaurant focus on to make their rating/reviews better?
(2) What neighborhoods in Pittsburgh have the best cuisine selection? Are there some areas that are more authentic, trendy or upscale than others? Based upon reviews of the top restaurants in Pittsburgh, can we recommend restaurants to a user?

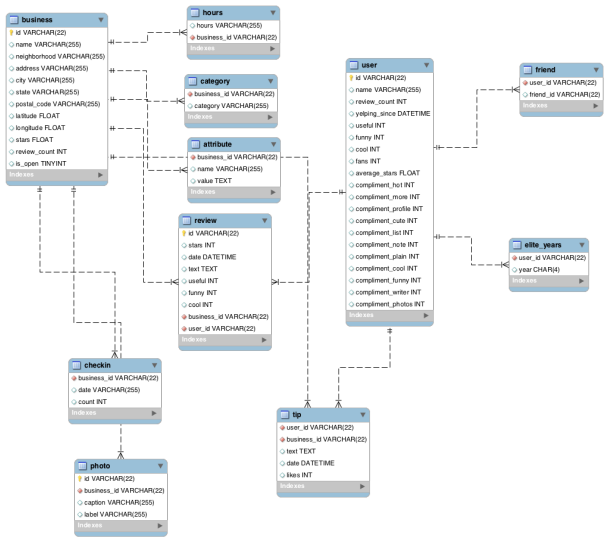
2. Methods

2.1 Data Collection and Transformation

With the Yelp Open Dataset[1] consisting of 4.7 million reviews, 156,000 business, and 12 metropolitan areas, we needed to filter out a signification amount of data and perform some further exploration in order to set potential thresholds to either include or exclude some data.
To get started, we created local databases and imported the given SQL file using MySQL Server and PyCharm in our development environment. This allowed us to begin writing queries in order to better understand our dataset and possibly explore other questions. From here, we narrowed down and verified what attributes we wanted to use in order to help answer our original questions. Using Jupyter Notebooks, we have created some basic visualizations and calculations in order to better filter, preprocess, and understand our data.

Below is the Yelp Dataset schema.

Yelp Data Schema

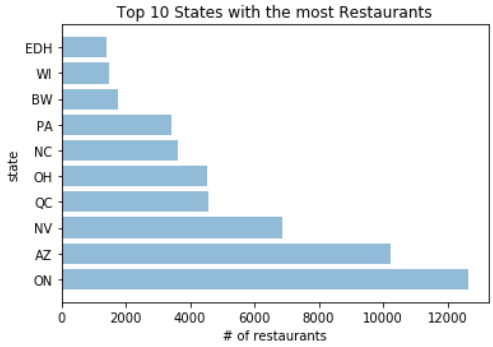


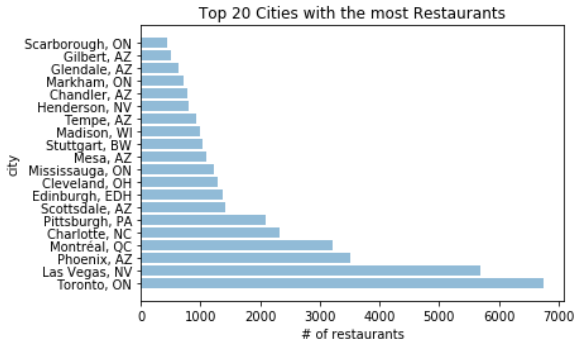
[1] https://www.yelp.com/dataset

2.2 Exploratory - Reviews

2.3 Exploratory - Pittsburgh Restaurants

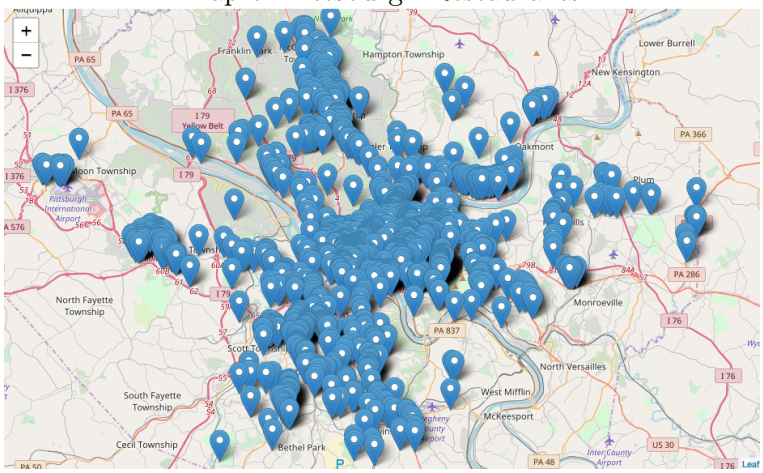
Since Boston didn't appear to be in the data set, we first performed some analysis on what other city we'd like to look at in order to perform clustering and recommendation on.



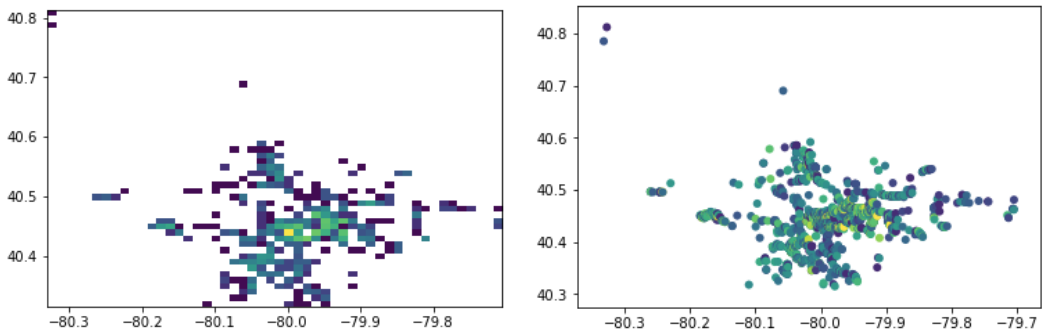


After running some queries and looking at the cities and states with the most restaurants, we decided to choose Pittsburgh due to the number of restaurants (2089), the significant number of neighborhoods (53), and the variety of cuisines (43 different cuisines). Although Phoenix and Las Vegas had more restaurants, Phoenix had no neighborhoods and Las Vegas had fewer neighborhoods. Interestingly enough, Pittsburgh only had 4 restaurants labeled as touristy, so we will look to see if there are particular neighborhoods that are more 'upscale' or 'divey', rather than authentic and touristy.
In order to get a better idea of the restaurants within Pittsburgh, we created a simple visualization which plots the latitude and longitude of every restaurant.

Map of Pittsburgh Restaurants



Although interesting, this doesn't tell us too much other than it appears to be very dense around the middle of the city. Our heat map of restaurant density and the most popular restaurants in Pittsburgh (shown left and right respectively) do show that a lot of the neighborhoods in the center of the city are more popular.



From here, we decided to further explore the Pittsburgh neighborhoods and first discover the top 10 neighborhoods with the most restaurants. Following this, this lead to creating a map of the top 10 neighborhoods with the highest average review count. In other words, we were curious what neighborhoods are the most popular among Yelp reviewers.