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

Update 1

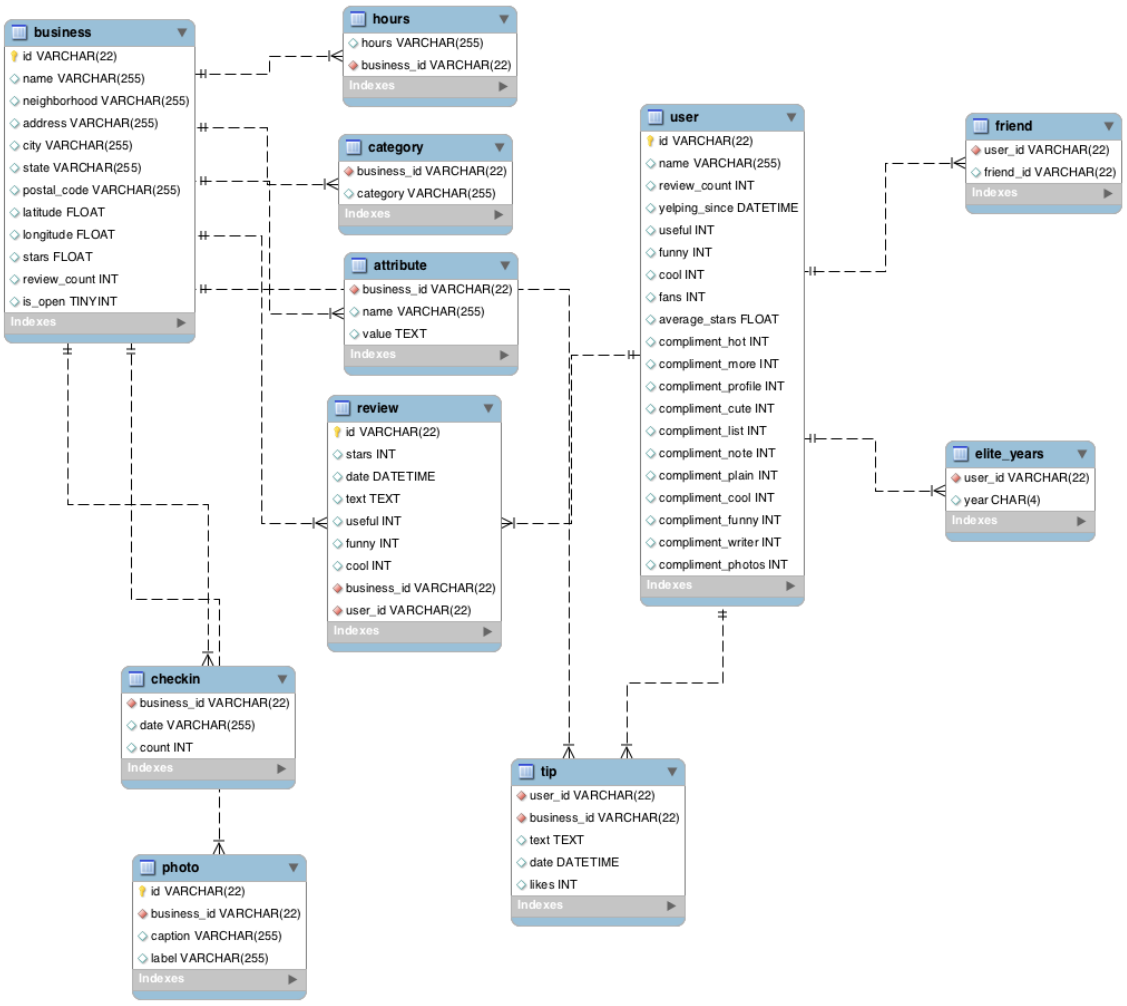
1. Problem Statement and Background

2. Methods

2.1 Exploratory analysis

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 particular data. Below is the Yelp Dataset schema.

Yelp Data Schema



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

In order to answer our original questions, we had to narrow down and verify what attributes we needed from our data set.

The following tables are used to answer question 1:

Business
id
name
neighborhood
city
longitude
latitude
stars

Review
business_id
user_id
stars
date
text
funny
useful

Tip
business_id
text
likes

The following tables are used to answer question two:

Business
id
name
neighborhood
city
longitude
latitude
stars

Category
business_id
category

To get started, we created local databases and imported the given SQL file using MySQL Server and PyCharm in our development environment. Using Jupyter Notebooks, we have created some basic visualizations and calculations in order to better filter and preprocess our data.

2. Extraction Update

Our original 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?

Number of Pittsburgh neighborhoods:
Number of categories/cuisines:
Categories/cuisines:

In order to tell authentic from trendy, we have joined the business, category and attributes table.

We have picked restaurants with more than 200 reviews: