CMSC12300
Group project proposal
Fulin Guo, Jiaxu Han, Ellen Hsieh, Ying Sun

Project Proposal

Description of dataset

In the group project, we plan to use the "Yelp" data from https://www.yelp.com/dataset. The size of this dataset is about 3.6GB when compressed and 8.69GB when uncompressed. There are six JSON files in total. Each file is composed of a single object type, one JSON-object per-line.

The business file (business.json) includes business data including business id, name, address, various attributes, hours, and categories. The review file (review.json) contains full review text data including the user_id that wrote the review and the business_id the review is written for. In addition, the file contains the stars that the reviewer gives to the business, the date, the text and various votes received. The user file (user.json) includes user's friend mapping and all the metadata associated with the user, such as user_id, name, review_count, friends, and so on. The check-in file (checkin.json) contains the business_id and check-in dates. The tip file (tip.json) contains tips written by a user on a business. Finally, photo file (photo.json) contains photo data including the caption and the specific classification of the photo.

In general, this dataset includes 6685900 reviews, 192609 businesses, 200000 pictures, and 10 metropolitan areas. It also covers the 1223094 tips by 1637138 users, over 1.2 million business attributes like hours, parking, availability, and ambiance aggregated check-ins over time for each of the 192609 businesses.

Research questions and hypotheses

Plan 1: Using "Yelp" dataset to understand the "food desert" problem

The “food desert” refers to an area that lacks access to affordable, good-quality, and healthy food. It plays an important role in the vicious cycle in the low-income population that eventually leads to poor health outcomes. The food desert problem is a complicated issue and potentially requires multi-level interventions from the industry, local government, and community programs. Here, the research question is trying to understand how the income level in a specific area relates to the review of the restaurants or supermarkets in that area. We hope that the results would contribute to the policymaking process associated with this issue.

In addition to the Yelp data described above, we will use individual income tax statistics (zip code data) (https://www.irs.gov/statistics/soi-tax-stats-individual-income-tax-statistics-2016-zip-code-data-soi) to estimate the income level of a certain area. More specifically, we will use zip code data collected in 2016.

There are two hypotheses associated with this research question. The first hypothesis is that restaurants in a low-income area have lower ratings and more negative reviews and tips whereas restaurants in a high-income are had higher ratings and more positive reviews and tips. This hypothesis may suggest that people living in low-income actually do not like the food around them, but they probably do not have better options. The second hypothesis is that restaurants in a low-income area and high-income area do not differ in terms of the ratings and reviews, but it differs in terms of what constitutes a “good” restaurant. This hypothesis may suggest that people living in a low-income neighborhood and high-income neighborhood may have different expectations and standards for what is a good restaurant for them.

We plan to use various visualizations to present the results of the data analysis.

Plan 2: Predicting popular and successful business

What makes a business popular and successful? We hope to use Yelp data to analyze what characteristics (for example, location, hours, categories, RestaurantsTakeOut, BusinessParking) are associated with popularity (review stars, review count, positive words in review) of a restaurant. We might use MapReduce to calculate the counts of positive words in the review for a particular restaurant. Then we might use Lasso regression/PCR to detect the factors that lead to the popularity of a restaurant.

Plan 3: Classification of users

Some users might like to give positive comments to restaurants while others like to give negative comments to restaurants. Some users might like to give star only while others might like to provide both grade and some comments for restaurants. Some users might like to write particular types of words in comments while others might like to write other types of words. The research question here is to find ways to classify users. We hope to use unsupervised learning to classify users. Potential methods include topic model, map reduce, and k-means clustering.

Plan 4: Friends' influence on reviews

Nowadays, more and more restaurants are available for customers to choose from when they think of eating out. The overwhelming choices lead people to seek for others' recommendation to narrow down their choices. In the past, friends' recommendation might be the main criterion to consider when thinking of going out for tasty food. However, now people can just easily get others' recommendation from the review forum like Yelp. Therefore, the research question that we are interested in is whether the friends' recommendations and reviews still have a significant influence on the user of Yelp.

The Hypothesis would be the reviews of the user on Yelp are influenced by his/her friends' reviews. Using Yelp dataset, we can know the user and who his/her friends are on Yelp. Then, we will look into their review and their friends' reviews according to different restaurants or businesses. After that, we can use sentimental analysis to analyze the text of those reviews and observe how they are related to each other.