

Yelp Extensive Data Analysis

Team Members:

- Nic Walter
- Naijia Wu
- Mengting Wang
- Mengxuan Li

Abstract:

Our goal is to use a Yelp dataset to create a visualization which provides multilayered data about local business locations in a single city. We plan to create an interactive map, word cloud, a bar chart, and line chart that exposes hidden commonalities and differences across user experiences of local businesses in a single area.

Data:

- <https://www.kaggle.com/yelp-dataset/yelp-dataset/code>
- https://github.com/ermiasgelaye/Yelp-Challenge#Google_Heat_Map_and_Visualization
- <https://www.kaggle.com/ambarish/a-very-extensive-data-analysis-of-yelp>

Data descriptions:

This dataset is a subset of Yelp's businesses, reviews, and user data. We found the dataset in Kaggle. There's information about businesses across 8 metropolitan areas in the USA and Canada.

5 subdatasets exist within this dataset:

1. business.json : Contains business data including location data, attributes, and categories.[\[ggmap, heatmap on location, restaurant rating\]](#)
2. review.json : Contains full review text data including the user_id that wrote the review and the business_id the review is written for. [\[wordcloud, text analysis-sentiment analysis/ topic modeling\]](#)
3. user.json : User data including the user's friend mapping and all the metadata associated with the user.[\[for social network analysis\]](#)
4. checkin.json : Checkins on a business.
5. tips.json : Tips written by a user on a business. Tips are shorter than reviews and tend to convey quick suggestions.[\[wordcloud\]](#)

Exploratory data analysis:

- Regional level:
 - We are interested in creating heatmaps of all restaurant locations; the highest review count; the highest star rating restaurants; and explore what are the most popular categories of restaurants; reviews of specific categories of restaurants in various cities; cities that have most business parties, and do a deeper dive into these findings.
- Individual level:

- For each individual, what are the relationships between user's elite, compliment, average_stars and restaurant ratings
- Find out the critical links or highly influential people within the community

Visualizations:

1. Interactive Map
 1. Possibilities include the option to filter by various metadata in user tips or comments (e.g. food poisoning, rude service, speedy order fulfillment), heat mapping by comments of a particular type
2. Word Cloud
 1. Possibilities include most common comment types within a certain area by content and sentiment, most frequently used words by business type, most frequently used words for positively rated businesses vs negatively rated businesses
3. Bar chart
 1. Possibilities include most commented upon businesses in the area, proportion of comments that list positive vs negative comments
4. Line chart
 1. Possibilities include the change over time of comments of a certain type (e.g. do comments about food poisoning increase or decrease over time in this area?)