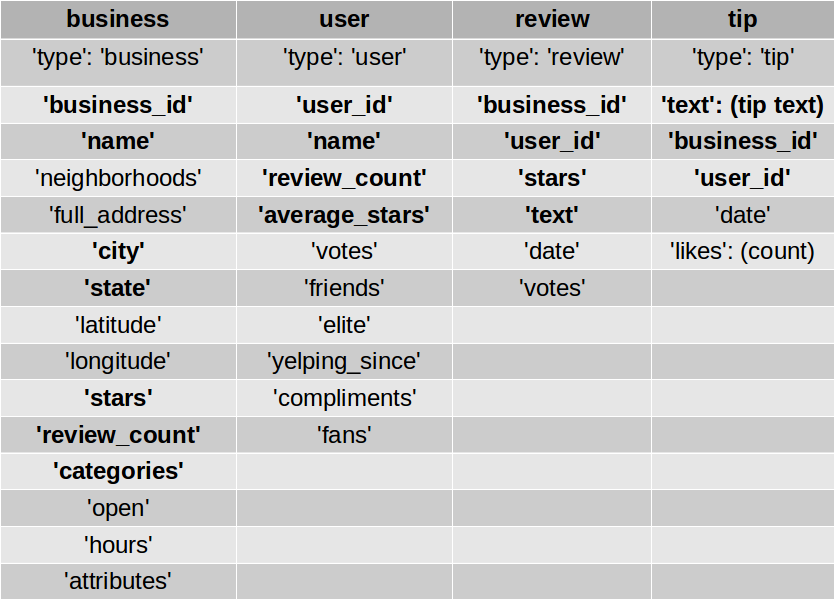
**Capstone Project: Sentimental analysis and cuisine based recommendation system based on yelp ratings, reviews and tip data on restaurants within the US**

Data source: <https://www.yelp.com/dataset_challenge>

Yelp.com is a website/mobile application with the capabilities of submitting reviews by users for various businesses across the world. In particular, it's a popular application for users who are interested in finding if a particular restaurant/cuisine offer the best quality of food and service. This project proposes to do a sentimental analysis of the reviews, ratings, and tip data tied with the demographic information within the interested US cities.

Clients: Generic user of the yelp.com site.

Fields within the four json files within the yelp dataset:



Interesting questions to be looked into:

1. User demographics (Male/Female)

2. Gender differences in the reviews and tip text within cities with most reviews

3. Correlation between tip text and reviews.

4. Recommendation system, Yelp ratings and (user data)

Additional datasets/libraries to be looked at:

1. Gender determination from US names

2. demographic/population data within the interested cities (yelp data) in US

**Phase 1: Preliminary data preparation**

Determine gender based on the US names database. This data will be used to see any differences in the review/tip differences based on gender.

Demographic/population data will be gathered for the interested US cities (within the yelp data). This data will be used to see demographic preferences to particular cuisine types in various cities.

The yelp data will be filtered to get only the restaurant data from US cities. A common data frame will be created from the four yelp data sets (businesses, users, review, and tip).

**Phase 2: Predict yelp rating and build a recommendation system**

Features from variables in datasets:

1. name

2. city

3. latitude

4. longitude

5. stars (target variable?)

6. review\_count

7. categories

User dataset:

8. name

9. review\_count

10. average\_stars(?)

11. votes

12. friends

13. elite?

14. yelping since

15. fans

Review dataset:

16. stars(target variable?)

17. text

18. date

19. votes(?)

Tip dataset:

20. text

21. date

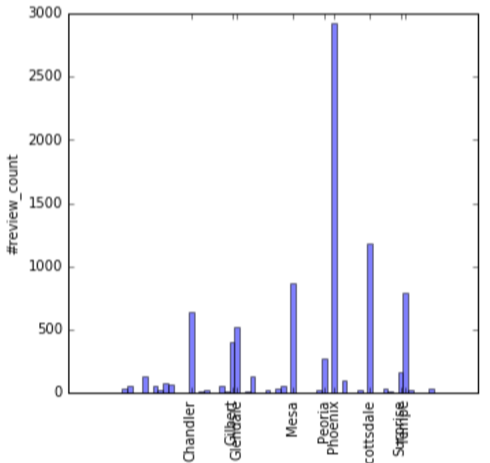
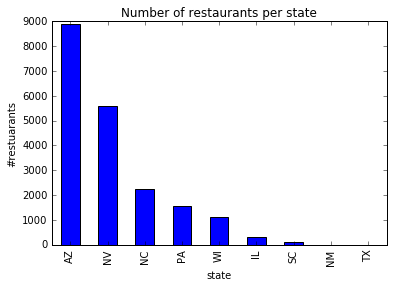
22. likes

Direct features:

Business dataset:

Feature#1: City

The US city in which the restaurant is located in. Within Yelp data, the restaurants are located in 9 states. Following graph shows the distributions of #restaurants within each state. Within each state there are multiple cities in which these restaurants are located in.



Feature#2: Latitude/Longitude

The latitude and longitude of the restaurant's location. This feature might be important for spatial differences in the reviews.

Feature#3: Review\_count

This is the #reviews per restaurant. This could be a key feature because it's directly related to the reviews and the ratings it might get.

User dataset:

Feature#4: name/gender

This is the names for each user. After it is mapped to the gender based differentiation, it might be a useful feature because we can extract male/female differences in reviews and how it might affect the extracted rating.

Feature#5: review\_count

This might be useful for examining the user's activity in submitting reviews in yelp. We could get a weightage based review rating based on an individual user's profile.

Feature#6: average\_stars

Not sure if this is something that would be considered as a target variable.

Feature#7: #friends

This might be useful for examining the user's social network and their activity on yelp.

Feature#8: elite

This shows the user's elite status with posting the #reviews on an annual basis. We can put a weightage to their elite status and

Feature#

Expand on features

15-20 features

sentiment – business – review

user – reviews, male/female ratio, sentiment

star reviewers -

time stamps

time series (tip count)

scikit learn

Initially, the user data will be analyzed to build a graph for generating a weight-age for the users with most reviews and also how they are linked.

For building yelp rating and the recommendation system, the target variable would be rank and different features would be included. The example features include:

user weight-age data

sentimental review data based on gender

demographic location data

tip review data

The generated yelp rating would be from a value of 1-5.

For the recommendation system, for a given geographic location and cuisine type the system would generate a list of restaurant places with the star ratings.

I would use supervised machine learning algorithms: multiple linear regression and random forests to predict the ratings.

For the multiple linear regression, individual linear classifiers will be built for each feature and later combined to get the target.

For the random forest approach, the model would be used to predict list of important features by running the tree multiple times (iterate several 100 times).

AUC?

For both the models, the training datasets would be divided to 10 different non-overlapping datasets for randomizing. The models would be trained on 9 datasets and tested on 1 dataset.

Cities: