Analysis of Geo-tagged Yelp Data

Chunxi Wang, Zhiyu Wang, Chuan Xing Zheng

Boston University

Abstract

Today, one of the most influential leading consumer review sites is Yelp. Yelp contains a large dataset that computer scientists can use to analyze the social relationships and trends between the star rating (from 1 to 5) and reviews, the number of reviews, the type of business, and the number of customers. Our plan is to store, query, and analyze Yelp's large dataset. The main purpose of this project is to use various data analysis and mining tools to attain optimal accuracy of classifying and summarizing the data. Since the Yelp dataset only contain 10 cities around the world, we decided to use Las Vegas as our main city to do the analysis. Optimally, we are finding interesting patterns between the type of restaurant and the star rating of the restaurants in Las Vegas. The data analysis tools we used in this project were K-means++ cluster, Dynamic Time Warping, Singular-Value Decomposition, Latent Semantic Analysis and Random Forest Classifier.

Dataset

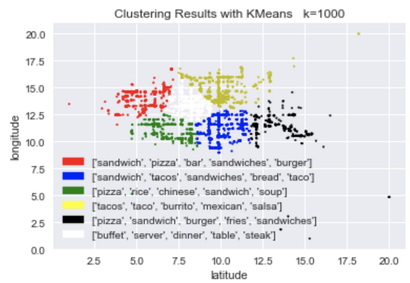
The data-sets we retrieved and worked with is the Yelp data-set and were downloaded from from the following link: https://wwDataset w.yelp.com/dataset. This data has been made available by Yelp for the purpose of the Yelp Dataset Challenge. Specifically, the challenge dataset contains the following data:

- 481K business attributes, e.g., hours, parking availability, ambience
- 1.6M reviews and 500K tips by 366K users for 61K businesses
- Aggregated check-ins over time for each of the 61K businesses

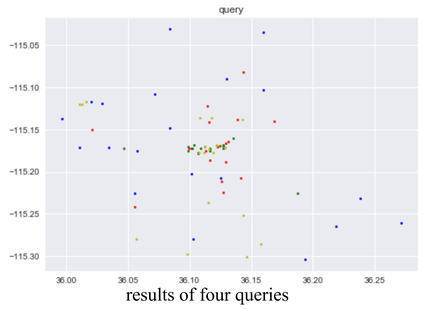
The dataset only contain 10 cities around the world:

- U.S.: Pittsburgh, Charlotte, Urbana-Champaign, Phoenix, Las Vegas, Madison
- Canada: Montreal and Waterloo
- Germany: Karlsruhe
- U.K.: Edinburgh

Dimension reduction



Label each cluster by restaurant reviews

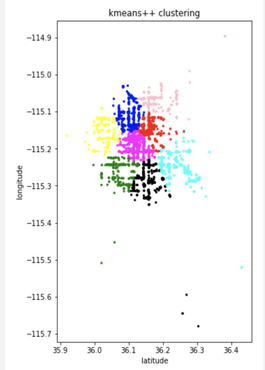


results of four queries

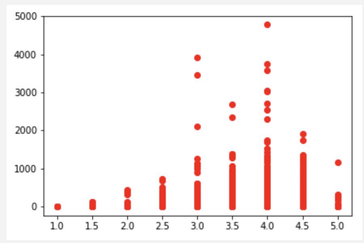
Spacial data Analysis

In this part, we found that the average rating has this tendency: the restaurants located about 10 miles from Downtown got the highest average stars.
Besides that, to know where to open restaurant in Las Vegase, we used random forest to predict the star a restaurant will get according to its location, open time and price range.

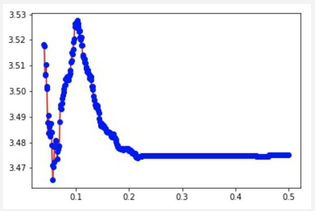
1.K-mean clustering result:
The average star rating for each cluster or color: Blue = 3.33, Green = 3.56, Red = 3.35, Cyan = 3.19, Magenta = 3.60, Yellow = 3.46, Black = 3.56, Pink = 3.49.



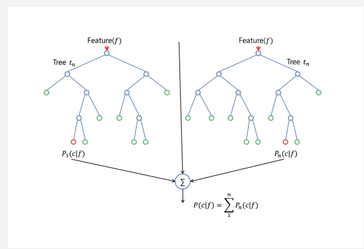
2. Comment number and stars.
X coordinate is the stars. Y coordinate is the number of the reviews. Each red point in the graph represents one restaurant.



3. Average stars and distance to downtown.
The downtown of Las Vegas is located in (36.1146,-115.1728). The distance is the Euclidean distance between restaurants and downtown point.



4. Random Forest classifier.
Random forests or random decision forests are an ensemble learning method for classification, regression and other tasks, that operate by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes (classification) or mean prediction (regression) of the individual trees.



5. Open time mapping.

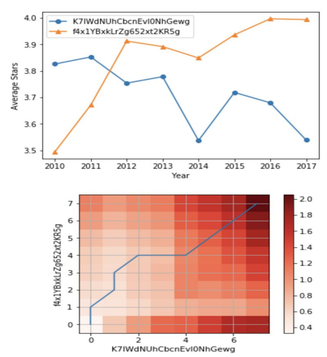
```
8:30-22:30 0
11:00-0:00 1
0:00-0:00 2
10:00-16:00 3
8:00-0:00 4
11:00-21:00 5
11:00-22:00 6
7:00-2:00 7
10:00-23:00 8
9:00-0:00 9
8:00-22:00 10
5:00-23:00 11
10:00-22:00 12
11:00-20:00 13
11:00-21:30 14
```

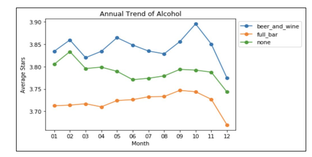
6. Accuracy output

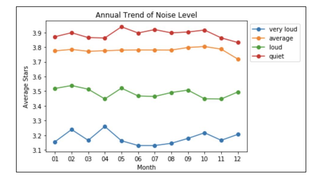
```
#accuracy print.
print("the size of the test set is", len(test))
print(count,"predictions out of ", len(test),"test are correct.")
print("So, the accuracy is", count/len(test))

the size of the test set is 200
140 predictions out of  200 test are correct.
So, the accuracy is 0.7
```

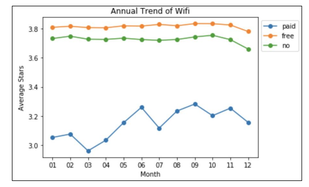
Time series data analysis

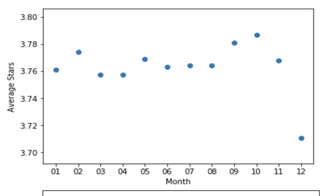
- Visualized the time series of average star rating for the restaurants.
- We ended up with only seven restaurants with at least 4,000 reviews and plotted their time series of average star rating from 2005 to 2017.
- picked the two restaurants with the same amount of years. We used dynamic time warping (DTW) with the Euclidean distance to find similarity between the two temporal sequences.
- We can see that there are higher star ratings in February compared to all other early months in a year. The reasons behind this might be because there are more holidays around this time of the year (e.g. Birthdays and Valentine's Day). Surprisingly, the ratings are not high during the summer months (June, July, August). We were expecting higher ratings during this time of the year because people tend to go out more. On the contrast, the highest rating received is during the Fall season (September and October). And the lowest rating received is during the month December.
- We used DTW with Euclidean distance again to measure the similarity between the temporal sequences of the alcohol features.
- Optimal warping path between the "beer_and_wine" and "none" sequences are monotonously similar throughout the year until when it gets to the very end of the year.

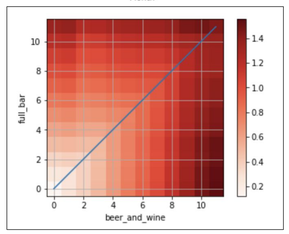


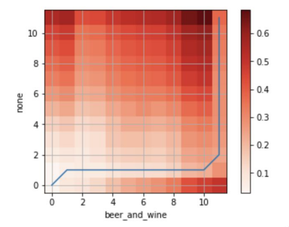












Conclusion

First, we get use K-means++ to cluster the restaurants into 8 parts, each cluster has different average rating star. We also get the relationship between average star rating and the distance away from downtown. We built two machine learning modules to predict the star of a restaurant based on its location, price range and open time. One module used Neural network, the other is based on random forest. Because only 2,775 records are provided, neural network can not provide effective result, and random forest could make 70 percent accurate prediction.

Second, we utilize LSA to analyze restaurant reviews. For labeling each cluster, we calculate the frequency of terms in each cluster and choose the five most frequent words in reviews as label. After transforming the restaurant-terms matrix by SVD, we can retrieve the most similar restaurants according to our query by calculating the distance. We get some interesting patterns by testing some queries. For example, expensive bars which serve sushi can often be found near downtown while quiet bar which serve pizza can be found everywhere in Las Vegas. This model is really useful for answer the query from people who want to find some specific place to eat.

At last, the temporal sequences were used with the Dynamic Time Warping method and we didn't see any major patterns for the evolution of star ratings between the 7 restaurants in Las Vegas that received more than 4,000 reviews from 2005 to 2017. Moreover, our DTW plot show that the optimum warping path did not show a close relationship between them, specifically for the 2 restaurants that consist with the same amount of years. Subsequently, we used DTW to analyze the annual temporal sequences for restaurants that serve alcohol. The 3 features that we compared in the alcohol attributes were "beer_and_wine", "full_bar", and "none". The stars received by the restaurants with the feature \say{beer_and_wine} is most similar to the feature "none".. On contrary, it is not similar to the feature "full_bar". This was an outcome we didn't expect because we thought restaurants that serve "beer_and_wine" should be similar to restaurants that serve "full_bar"; you can have beer and wine in full bar too.

Reference

https://simonpaarlberg.com/post/latent-semantic-analyses/
http://repository.cmu.edu/cgi/viewcontent.cgi?article=2022\&context=dissertations
https://cseweb.ucsd.edu/jmcauley/cse255/reports/wi15/Wa'el_Farhan.pdf
http://cs229.stanford.edu/proj2015/301_report.pdf
http://scikit-learn.org/stable/
http://www.numpy.org

GitHub

https://github.com/ZhiyuWang95/CS562_Geo-tagged_data