ANALYSIS AND VISUALIZATION OF YELP DATA

Anjali Ramchandani
Department of Computer Science and Engineering
University of California, Riverside
aramc001@ucr.edu

Gowtham Tumati
Department of Computer Science and Engineering
University of California, Riverside
gtuma002@ucr.edu

Lovepreet Singh Dhaliwal
Department of Computer Science and Engineering
University of California, Riverside
ldhal001@ucr.edu

Sudip Bala
Department of Computer Science and Engineering
University of California, Riverside
sbala027@ucr.edu

Abstract

There is a rising competition amongst restaurants that employ social media and other tactics to gain popularity. Consequently, review platforms like Yelp, have become one of the tools for businesses to gauge their reach. Also, consumers are increasingly using these platforms to decide where to spend their money. Naturally, a good reputation among users is the key for survival for restaurants. In this project we utilize Yelp platform to perform trend analysis. From the data available for multiple types of businesses, we select restaurant businesses and categorize based on cuisines.We demonstrate how spatial correlation plays a huge factor in popularity of a few cuisines. To show this we built an interactive model with geo-spatial visualizations to categorize the location-concentration of the popular restaurants/cuisines.

1 Introduction

1.1 Background

Yelp is a website (and hosts mobile apps) for crowd-sourced reviews about businesses. It has also evolved into a social network, where users can follow other users on Yelp or their linked social profiles on various other platforms and host social events. Consumers are using these platforms to decide where to eat, where to shop, where to repair things etc. and create a personal network to share their finds. Since the era of Web 2.0, this is not surprising and it is in fact desirable as it makes global connectivity more accessible. With the advantages, there are some aspects that users need to be careful about. The entire business ecosystem is transformed to handle the changes due to this hyper-connectivity. Businesses are affected by how they are viewed on social media as it is

Preprint. Under review.

now rare that a consumer would not consult reviews or other social media topics before using any service. Although all types of businesses face similar challenges when it comes to gaining relevance in social media, restaurants are the major business that are hugely impacted by consumers' social media presence. A research indicates that a one-star increase in independent restaurants' reviews led to 5 to 9% increase in their revenue[1]. According to [3], out of the top 10 major businesses on Yelp, restaurants rank first. Due to the popularity of Yelp as a reviews platform, we will utilize Yelp platform to perform trend analysis, that is which food items are popular in particular locations and which "qualities" make the food item or a restaurant stand out. In the coming sections we will investigate the methods use to establish this relation and the challenges faced in each process.

1.2 Problem Statement

Our goal is to demonstrate the role of geographical location in trends related to reviews, cuisine, popularity of various restaurants. We also want to build a visualization interface to be able to easily make out new challenges. We have demonstrated this using reviews in US restaurants spread over all the 50 states. There are multiple factors to our problem

- Gather reviews data and pre-format them for analysis and spatial queries
- Filtering data based on consumer inputs using interactive filters.
- Implementing a nearby search using kd-tree queries.
- Visualizing maps data for helping businesses and consumers.

1.3 Plan

For the first task we will use Yelp's academic dataset. Formatting is based mainly on the fact that we will use data only for restaurants and not any other businesses. At this step we simply pre-filter based on the categories required later for our search and we retain the location information. At the next step we take in user-inputs such as their choice of cuisine and the star rating that they would prefer and further reduce the size of the set. We had earlier planned to use real-time data, but since we are limited to using the academic dataset, we have demonstrated a simple proof of concept in python. For performing a $k - d\,tree\,search$, there were plenty options available in python. SciPy provides a $spatial$ module that gives an efficient version for knn search with k-d tree. For data visualization we compared the alternatives between D3.js and Leaflet.js. We studied that for maps based visualization, we have no better alternative than leaflet.js, hence we are using a combination of leaflet.js and ggplot for our purpose. Since we have developed the app in python, there is a package known as $folium$ which is built as wrapper over leaflet.js.

2 Related Work

A lot of research an verification surrounds Yelp reviews and other platforms and their location correlation. Tayeen (2019)[4] did an extensive verification of reviews platform by using two datasets; a Yelp dataset for business information and reviews, and another Location dataset that gathers location-based information in a city or an area. Che and Xia (2020)[5] built a predictive model to predict the sucess of restaurants based on Yelp reviews. Five models are created respectively from five machine learning methods which include regression models, Naive Bayes, Decision Tree, and Neural Network. Prithvirajan (2015)[10] build a model to analyze how well or poorly the star ratings (on a scale of one star to five stars) associated with these reviews on Yelp tally with the sentiment derived from the textual portion of the consumer review. All of the research is based on some predictive analysis and corresponding ML models. Our goal in this project is to validate the location dependence of the restaurants, explore geovisualization techniques and perform data analysis on reviews and location data.

3 Dataset and features

3.1 Dataset

We had planned to use Yelp's $fusion\,API$ for real time data extraction. Although we did a proof-of-concept once, our app failed to run later due to licensing issues. Hence, we decided to limit the implementation to the academic dataset. The Yelp academic dataset is a subset of their businesses, reviews, and user data for use in personal, educational, and academic purposes. All these subsets are available as JSON files. The dataset is limited in the number of businesses and reviews it provides. (it is limited to only recommended reviews). As described by their website, it is limited to following number for each dataset:

- 8,021,122 reviews
- 209,393 businesses
- 200,000 pictures
- 10 metropolitan areas

```json
{

    "business_id": "tnhfDv5Il8EaGSXZGiuQGg",
    "name": "Garaje",
    "address": "475 3rd St",
    "city": "San Francisco",
    "state": "CA",
    "postal code": "94107",
    "latitude": 37.7817529521,
    "longitude": -122.39612197,
    "stars": 4.5,
    "review_count": 1198,
    "is_open": 1,
    "attributes": {
        "RestaurantsTakeOut": true,
        "BusinessParking": {
            "garage": false,
            "street": true,
            "validated": false,
            "lot": false,
            "valet": false
        },
    },
    "categories": [
        "Mexican",
        "Burgers",
        "Gastropubs"
    ],
    "hours": {
        "Monday": "10:00-21:00",
        "Tuesday": "10:00-21:00",
        "Friday": "10:00-21:00",
        "Wednesday": "10:00-21:00",
        "Thursday": "10:00-21:00",
        "Sunday": "11:00-18:00",
        "Saturday": "10:00-21:00"
    }
}
```

Listing 1: business.json example from Yelp academic dataset

From these datasets, we will utilise the second type, that is business.json. We wanted to utilize reviews.json as well to analyse the sentiment of these reviews, also as a correlation to location data, but we are limited by the reduced dataset. Therefore we do not have the flexibility for using an ML model as a lot of error is added by the incomplete data. So going forward with business.json, we can run our app based on the *review count* provide by the json file. Each file is composed of a single object type, one JSON-object per-line. An example is shown in listing 1.

3.2 Feature Selection

We are not using an ML model for sentiment analysis but if the scope of this project is extended to sentiment analysis, exactly these categories can be employed for feature selection for classification models or SVM models. However, a more suitable dataset or real time dataset should be used. Along with that the review.json file provided by Yelp should be added. Since our dataset is reduced, sentiment analysis is considered as a future aspect hence we do not discuss reviews.json any further. Now the resultant filtered json is described in listing 2.

```
{

    "business_id": "tnhfDv5Il8EaGSXZGiuQGg",
    "name": "Garaje",
    "address": "475 3rd St",
    "city": "San Francisco",
    "state": "CA",
    "postal code": "94107",
    "latitude": 37.7817529521,
    "longitude": -122.39612197,
    "stars": 4.5,
    "review_count": 1198,
    "is_open": 1,
    "attributes": {
        "RestaurantsTakeOut": true,
        "BusinessParking": {
            "garage": false,
            "street": true,
            "validated": false,
            "lot": false,
            "valet": false
        },
    },
    "category": "Mexican",
    "hours": {
        "Monday": "10:00-21:00",
        "Tuesday": "10:00-21:00",
        "Friday": "10:00-21:00",
        "Wednesday": "10:00-21:00",
        "Thursday": "10:00-21:00",
        "Sunday": "11:00-18:00",
        "Saturday": "10:00-21:00"
    }
}
```

Listing 2: business.json objects after initial filtering

The important filtering aspect here is that ”*categories*” is changed to ”*category*”. We categorize the restaurants by cuisine type only. Hence all other categorization is ignored. Hence we build a new parameter for category and replace the older categorization. Any restaurants that do not categorize themselves based on cuisines result in null category, so we filter these out and remove any duplicate data.

4 Results and Analysis

4.1 Visualization

Leaflet is the leading open-source JavaScript library for mobile-friendly interactive maps. It allows users to use layers such as Tile layers, WMS, Markers, Popups, Vector layers (polylines, polygons, circles, etc.), Image overlays and GeoJSON. Since we are using python, we opted for a wrapper known as $folium$. Folium makes it easy to visualize data that's been manipulated in Python on an interactive leaflet map. It enables both the binding of data to a map for choropleth visualizations as well as passing rich vector/raster/HTML visualizations as markers on the map. Interacting with Folium is much more easier than leaflet.js, as there are many low-level advanced settings such as CORS which are tremendously made easier. In our case, we simply interact with the data inline rather than running a web service. The first visulaisation of the prefiltered restaurants is shown in figure 1.

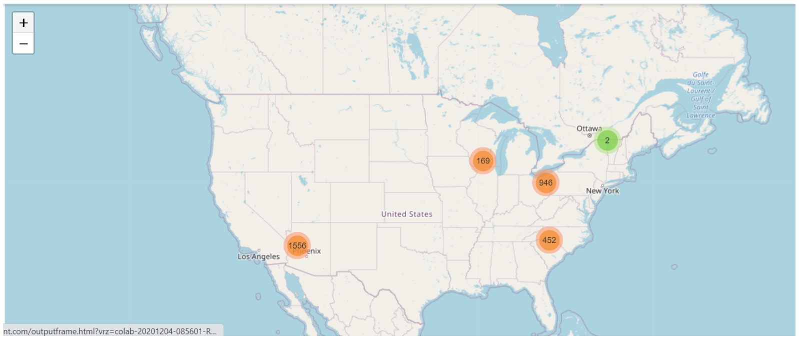


Figure 1: Filtered Restaurants across United States from the Dataset

To understand spatial correlation, we demonstrated location behaviour with multiple steps. The first step is the heat map shown in figure 2. It contains the exact data as in step 1 pre-filtering. The difference here is the heat map is varied based on number of restaurants in a particular area. This shows some metropolises have more concentration of restaurants than others.

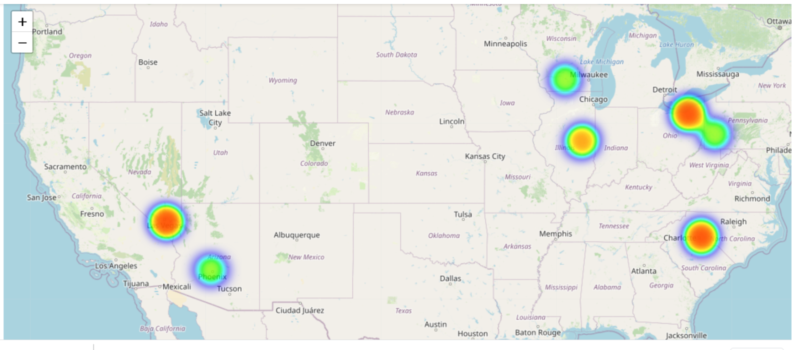
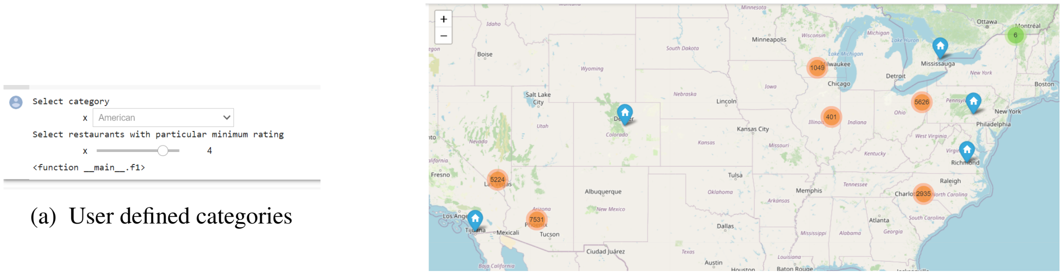


Figure 2: Heat Map displaying the concentration of filtered restaurants in each location.

We turn to sptep 2 now where users can filter out the restaurants based on categories and star ratings. This is illustrated in figure 3 and figure 4.


Select category

x American

Select restaurants with particular minimum rating

x 4

<function __main__.f1>

(a) User defined categories

(b) Result dataset

Figure 3: User Interactive Filter and Results

4.2 Spatial queries

k-d Tree is a modified Binary Search Tree(BST) that can perform search in multi-dimensions and that's why k-dimensional. This makes it an ideal search for spatial data which is multi-dimensional. The k-d tree differs from the BST because every leaf node is a k-dimensional point here. It is one of the most popular data structure for spatial queries besides r-trees. k-d tree query is based on this data structure. For the third step of performing nearest neighbor search, we use a knn query for nearest restaurants confirming to users' filters. The result of this query is shown in figure 5.

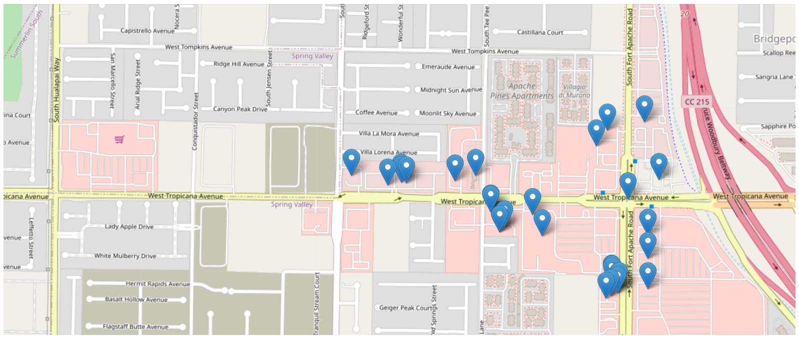
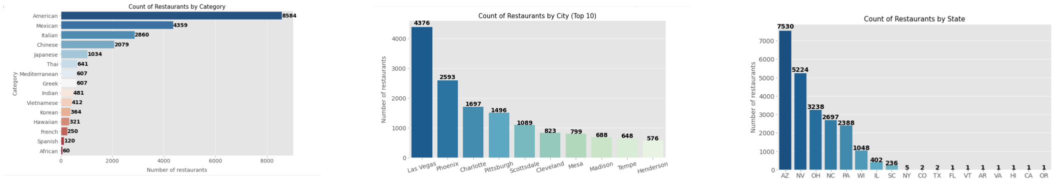


Figure 4: Result of the K-D Tree Search Query

We used a recursive algorithm to build the k-d tree. If we consider a $single$ knn query over M points limited by area of interest, the overall complexity of building the k-d tree is $O(M[log^2M])$. The advantage of using k-d tree instead of simple collections, especially for location data is that the time complexity is reduced to $O(logM)$ time instead of linear time. Therefore the total complexity for a single knn search is $O(M[log^2M])$. For N such queries, the time complexity is $O(M[log^2M] + NlogM)$

4.3 Exploratory Data Analysis

We explained in the last section how visualisation with heat maps helps us understand that some cities have more density of restaurants compared to other cities. We also showed how we can use spatial search to get faster knn results. However, that is a very high level visualization. The categorization that we have for restaurants and the values that we have from Yelp's business.json file can be used to provide stronger results for location correlation. These fields can be used to understand the correlation at various resolutions: at city level, state level or at each category. Note that for a finer are-wise resolution we have already demonstrated user-interactive filters and k-d tree search. Now we will demonstrate the behavious at slightly higher resolution.



(a) Count of restaurants per category

(b) Count of restaurants per city

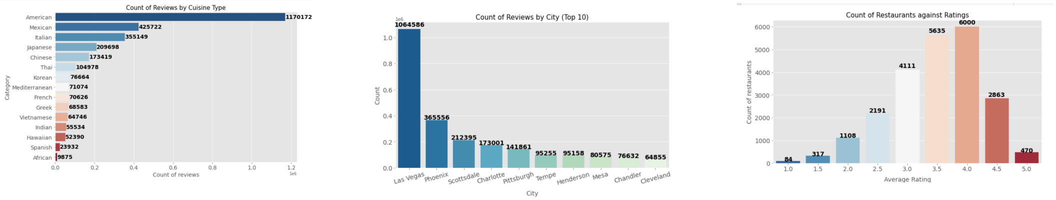
(c) Count of restaurants per state

Figure 5: Count of restaurants with different demographics

Figure 5 shows the cont of restaurants based on city, state and each category. We can clearly observe some metropolises have larger concentration of restaurants. We also observe that majority of restaurants categorize as "American" cuisine type. It should be noted that these methods are elementary and are verified. However, a few results in the data analysis will come across as incorrect.

For example, the restaurants with user filters in CA is just 1 and that in AZ $7.5k+$. This does not match the reality, but this issue arises due to the limited dataset as explained earlier.

We have seen the behaviour of restaurants in correspondence to location. Our next step is to observe the behaviour of reviews based on locations. In concept, the behaviour of the reviews should be the same as those of restaurants. Since the reviews directly vary in accordance with restaurants, we should observe the same pattern. This is seen in figure 6, where we have used the same demographics to plot reviews. However, more analysis is needed in this regard. The Yelp academic dataset is biased at some states and for some metropolises it does not provide enough data. We worked on 0 restaurants in New York and Los Angeles. The dataset is sufficient for showing the location correlation and proving the correctness of libraries and methods. Finally in figure 6(c) we observe the relation between star rating and number of restaurants. Since most consumers prefer 4 and above rating, we filter almost half of the dataset.



(a) Count of reviews per cuisine type

(b) Count of reviews per city

(c) Count of Restaurants against ratings

Figure 6: Count of reviews with different demographics

5 Future Work

Given the scope of this project we implemented crucial methods to justify location patterns in reviews for restaurants. However there are many subtleties involved in extracting data and finding all the possible relations. We accept that a lot can be improved and built as an extension to this project. Some of the aspects that can be improved are:

- **Validation methods**: We have used plots and visualization on maps as a method for proving of location-data. We used restaurant behaviour as well as reviews behaviour for analysis. We can add more methods to validate our analysis. This is not in the scope of this project but predictive classification models can be used[5]. We can assign weights to restaurants in relation to location and reviews and use predictive models to find the performance of nearby restaurants and validate if they follow the same patterns.
- **Relation between dominant reviews location and dominant restaurants**: As mentioned in section 4.3, the behaviour of reviews is similar to the behaviour of restaurants based on location. But we cannot conclude that this is the case since we were performing analysis on reduced dataset. Even if using this data, if we analyse the relation between places where we have maximum reviews and places where we have maximum highly rated restaurants, the results might be contrasting. For example a restaurant having 300 reviews and average rating of 3 might be better than a restaurant receiving 50 reviews and an average rating of 4.
- **Finding the error introduced by reduced dataset**: We have mentioned in earlier section the errors introduced by the Yelp academic dataset. Despite that the academic dataset is heavily experimented on academic purposes. This begs the question that whether the errors introduced by the dataset have any influence in the correctness of our methods. We can do a piece-by-piece analysis of same methods run on academic dataset and real-time data (if we resolve licensing issues) and quantify the errors in each case.
- **Sentiment Analysis**: We mentioned earlier that we are ignoring the reviews.json file provided by Yelp, that contain the actual user reviews. If we build a model for categorizing these reviews[10], we can perform sentiment analysis on the desired quality of food for each category of cuisines.

6 Conclusion

We have successfully implemented a model to visualize the location dependence of restaurants as well as their reviews on Yelp using interactive maps. We also implemented a knn search that is based on user interactive filters based on categories and star rating. We analysed all possible relations between location and reviews and in the process realized some errors introduced by the limited dataset. We came to conclusion based on the dataset that reviews vary proportional the concentration of restaurants in an area. We also concluded that approximately half of the restaurants receive star ratings 4 and above irrespective of the location. We also identified a few key areas like use of real time data and sentiment analysis that can be added to extend the scope of this project.

References

1. Luca, Michael. "Reviews, reputation, and revenue: The case of Yelp. com." Com (March 15, 2016). Harvard Business School NOM Unit Working Paper 12-016 (2016).
2. Q. Xuan et al., "Modern Food Foraging Patterns: Geography and Cuisine Choices of Restaurant Patrons on Yelp," in IEEE Transactions on Computational Social Systems, vol. 5, no. 2, pp. 508-517, June 2018, doi: 10.1109/TCSS.2018.2819659.
3. 32. B. She, X. Zhu and S. Bao, "Spatial data integration and analysis with spatial intelligence," 2010 18th International Conference on Geoinformatics, Beijing, 2010, pp. 1-6, doi:10.1109/GEOINFORMATICS.2010.5567628.
4. A. S. Md. Tayeen, A. Mtibaa and S. Misra, "Location, Location, Location! Quantifying the True Impact of Location on Business Reviews Using a Yelp Dataset," 2019 IEEE/ACM International Conference on Advances in Social Networks Analysis andMining (ASONAM), Vancouver, BC, Canada, 2019, pp. 1081-1088, doi: 10.1145/3341161.3345334.
5. Y. Chen and F. Xia, "Restaurants' Rating Prediction Using Yelp Dataset," 2020 IEEE International Conference on Advances in Electrical Engineering and Computer Applications(AEECA), Dalian, China, 2020, pp. 113-117, doi: 10.1109/AEECA49918.2020.9213704.
6. S. B. Hegde, S. Satyappanavar and S. Setty, "Sentiment based Food Classification for Restaurant Business," 2018 International Conference on Advances in Computing, Communications and Informatics (ICACCI), Bangalore, 2018, pp. 1455-1462, doi: 10.1109/ICACCI.2018.8554794.
7. A. Nasser, D. Hamad and C. Nasr, "Visualization Methods for Exploratory Data Analysis," 2006 2nd International Conference on Information and Communication Technologies, Damascus, 2006, pp. 1379-1384, doi: 10.1109/ICTTA.2006.1684582.
8. Y. H. Dehkordi, A. Thomo and S. Ganti, "Incorporating User Reviews as Implicit Feedback for Improving Recommender Systems," 2014 IEEE Fourth International Conference on Big Data and Cloud Computing, Sydney, NSW, 2014, pp. 455-462, doi: 10.1109/BDCloud.2014.51.
9. Ruhui Shen, Jialiang Shen, Yuhong Li and HaohanWang, "Predicting usefulness of Yelp reviews with localized linear regression models," 2016 7th IEEE International Conference on Software Engineering and Service Science (ICSESS), Beijing, 2016, pp. 189-192, doi: 10.1109/ICSESS.2016.7883046.
10. M. Prithivirajan, V. Lai, K. J. Shimand K. P. Shung, "Analysis of star ratings in consumer reviews: A case study of Yelp," 2015 IEEE International Conference on Big Data (Big Data), Santa Clara, CA, 2015, pp. 2954-2956, doi: 10.1109/BigData.2015.7364134.
11. A. Salinca, "Business Reviews Classification Using Sentiment Analysis," 2015 17th International Symposium on Symbolic and Numeric Algorithms for Scientific Computing (SYNASC), Timisoara, 2015, pp. 247-250, doi: 10.1109/SYNASC.2015.46.
12. Bao, Fan, and Jia Chen. "Visual framework for big data in d3.js." 2014 Ieee Workshop on Electronics, Computer and Applications. IEEE, 2014.
13. S. Madle and D. Das, "Near by Services using Spatial Computing," 2019 International Conference on Wireless Communications Signal Processing and Networking (WiSPNET), Chennai, India, 2019, pp. 62-65, doi: 10.1109/WiSPNET45539.2019.9032849.
14. A. Eldawy and M. F. Mokbel, "The Era of Big Spatial Data: Challenges and Opportunities," 2015 16th IEEE International Conference onMobile DataManagement, Pittsburgh, PA, 2015, pp. 7-10, doi: 10.1109/MDM.2015.82.