Analysis and Visualization for Yelp Dataset Using Google Maps API

Jieyan Chen, Wenyan Du, Boxuan Li, Miao Sun, Ruiying Wang

1 Introduction

Google Maps has built APIs that can be conveniently incorporated into all sorts of websites and applications. Yelp.com also utilizes Google Maps API. On the Yelp website, users can access the Google map-view of businesses with rich Yelp business information and reviews. However, users who locate businesses directly using Google Maps could not easily access Yelp reviews, as Google Maps provide only reviews and ratings assembled by Google. It is reasonable to believe that more users would access Google Maps rather than Yelp when looking for business information since Google Maps has five times more monthly users than Yelp. ***(Heilmeier Q2)***

2 Problem Definition

In this study, we aimed to incorporate the Yelp dataset into Google Maps API so that users could locate a business and access related Yelp reviews or information directly on Google Maps.***(Q1)*** If Google Maps users click on Yelp reviews and ratings and get interested in them, they are more likely also to become Yelp users.***(Q3, 5)*** It will also benefit Google Maps users as they can easily view rich information from Yelp while using Google Maps, avoiding tedious searching or clicking on Yelp website.***(Q4)*** The success of this project can be measured by analysis of surveys from potential users about user experiences and satisfactory on the intuitiveness and responsiveness of the website.***(Q5)*** Since Google Maps API provides a one-year free license, this project will not cost anything other than intellectual efforts from the team.***(Q7)*** Therefore, this project will not have any financial risks. In case that this project might not attract as many users as intended, it will still provide team members a good opportunity to improve our data analysis and visualization skills. Thus, this is another significant payoff besides increasing user numbers for Yelp as mentioned above. ***(Q6)***

3 Related Work / Survey

3.1 Yelp Dataset Analysis

As a subset of Yelp business, the Yelp dataset contains user and review information available as JSON files. It has been widely studied in different areas. One of the most studied data is the Yelp ratings. Yelp ratings have a profound effect on the success of a business. Studies [1][2] show that the web-based customer reviews have a significant impact on consumer purchases. Half more star rating helps restaurants to sell 19% more orders. Latent Dirichlet Allocation(LDA) model is a popular topic modeling to discover topics and subtopics. One study conducted LDA analysis to point out the demands of customers from the Yelp review dataset[3]. These topics are helpful because they can provide meaningful insights regarding what customers care about. Python (language used in the study) topic modeling libraries help users to find interesting topics quickly. The shortcoming of this study is that it uses unsupervised learning. Therefore, the accuracy of the predicted topics is not guaranteed. To overcome this, we can do a t-test of the star rating for positive and negative topics. Another study [4] built 16 different prediction models, including logistic regression, Naive Bayes Classification, Linear Support Vector Classification(SVC), etc., to predict the star ratings of users. Readers often prefer to look at the star ratings while not giving much weight to the review texts. By building multi-class classification machine learning models, the authors found that the logistic regression achieved the highest accuracy of 64% followed by Linear SVC.

Many studies also try to improve Yelp recommendation systems. Sawant et al. generated a recommendation system for Yelp users to easily find restaurants they may like. The study first used K nearest neighbors to cluster the business and users and then conducted different algorithms on the datasets. The "Cascaded Clustered Multi-Step Weighted BiPartite Graph Projection" algorithm performed the best among all. Sawant later built another recommendation system for all business kinds[6] and found that utilizing the regular properties and the network properties can make the predictions much more accurate. Yelp data are also used in Hybrid Recommender System to predict Yelp users preferences[13].

3.2 Geographic Data Visualization

Data visualization is the graphical representation of data. Good data visualization can help people understand the trends, outliers, and patterns better[14]. Sopan et al.[7] designed a Community Health Map, which enables users to visualize health care data in multivariate spaces and geospatially. Users can visualize the geospatial distribution of variables on an interactive map, which helps users efficiently explore huge geographic datasets. The interactive map allows users to use sliders and checkboxes to filter the map. The shortcoming of this study is that design a system from scratch requiring a lot of efforts. Users also need to learn the new system. (To overcome it, we propose to use Google Maps API. This way, UI design will be built on Google Maps API which is easier. Plus, the users are already familiar with the functions in the Google Maps.)

Artificial Neural Networks(ANN) is an emerging solution for pattern recognition[15]. Among the ANN models, Self-Organizing Map(SOM) is specifically useful in data exploration. Koua[8] used the SOM algorithm to explore a geospatial dataset. SOM helps uncover the structure and patterns of the datasets. Five main visualization techniques were explored: a cell visualization, projections, visualization of component planes, 2D and 3D surface plot of distance matrices. These techniques use distances, regions, and scale to represent the information.

3.3 Google Maps API

Mapping solutions are everywhere in our lives now. We use them to view events nearby, search for addresses and get driving directions. There are many mapping solutions, including Apple Maps, Yahoo! Maps and Bing Maps. The most popular one is Google Maps[9]. Google Maps API was released publicly in June 2005[9]. It provides builtin functions and classes to display locations on the map or show different routes and so on.[12]. These collections of Javascript classes can be called from a web page. Many examples use Google Maps data for applications, such as Housing maps which uses Craigslist housing data and Wikimapia for user descriptions of places.

Another example is an application developed by Gibin et al.[10] for thematic mapping in Google Maps. The application (GMap Creator) can read the project shapefiles onto a thematic map layer based on different data attributes. However, the solution is not scalable for extensive geographical coverages. Data pre-processing is needed to make the computation faster. In another study, Rahmi et al.[11] developed Android and web applications based on MySQL (as the primary data storage) and Google Maps API to help patients to find a suitable doctor quickly.

4 Proposed Method

Both Google Maps and Yelp own a tremendous number of users, with a significant overlap of the user group. To obtain sufficient information about a business, these users may refer to both applications, but this process may lead to inconvenience while collecting complete information by switching back and forth between Google Maps and Yelp. Moreover, for some popular businesses, it is very difficult for users to read through hundreds or even thousands reviews one-by-one and make a comprehensive conclusion. To overcome these problems, we decided to 1) integrate Yelp information into Google Maps. 2) conduct data analysis on reviews, and then select keywords and top reviews to display. 3) visualize analyzed data and statistics for business.

We proposed two ways to utilize Yelp dataset: 1) Download the Yelp Open Dataset. Process and analyze this dataset, then visualize useful information. 2) Use Yelp Fusion API to request data for visualization. These two tools are both provided through Yelp Developers. For Google Maps, we use Maps JavaScript API and Places API from Google Maps Platform.

4.1 First Stage

To incorporate geographic data of businesses from Yelp into Google Maps, we proposed to use the Yelp Fusion API. Yelp Fusion API request is a real-time query that does not rely on additional database structures, and there are no extra data processing steps after data acquisition. For each point of interest (POI) on Google Maps, we requested a subset of Yelp data with matching geographic information. However, this query returns a list of POIs with similar latitudes and longitudes. To locate the exact POIs, we proposed to match the business name and to add a limitation on radius and number of results to show, in addition to queries based on latitudes and longitudes.

Upon a mouse click of a POI, the original Google Maps API shows a tooltip containing information collected by Google, including name, address, and a link to view on Google Maps. We first disabled the showing of these Google text contents while enabling the display of Yelp contents and/or other desired Google contents in the tooltip. We achieved this by applying the "event.stop" function by Google Maps API. In order to show Yelp information in the tooltip upon mouse click on a POI, we utilized the "addListener" function through Google Maps API, which creates an event handler that can request data from Yelp upon mouse click on Google Maps.

We visualized Yelp and Google data together on a web page divided into two panels, the left shows text and image information, and the right shows the map. Upon mouse click on a POI on the map, the followings will show in the tooltip: 1) Google category sign; 2) Google POI name; 3) Yelp consumer price range; 4) address; 5) Yelp ratings; 6) Google ratings and; 7) A link to the detailed data analysis. Furthermore, upon mouse click, more Yelp and Google information about the POI will show on the left panel. On top of the left panel, we also proposed to have a "Place Searches" box that searches a specific POI on Google Maps. It can handle complicated search input, similar to what Google Maps does, and the map panel will navigate to the POI.

4.2 Second Stage

In the second stage of the project, we focused on detailed data analysis and visualization of both Yelp Open Dataset and Google data. It has three sections: general information, data analysis, and data visualization. Both general information and data analysis are shown in "Data analysis" page. This page will display by clicking "View data analysis" in the tooltip showed by mouse-click on each POI.

For general information, each POI contains multiple pieces of information including location and hours, website link, phone number, amenities, direction, highlighted reviews, recommended similar POIs, order food online, photos, etc. We first conducted an internal survey within our team about which pieces of information we should include for the analysis and visualization. We then summarized the results and selected the most voted pieces of information to display in the general information section. It turned out that the business name (Yelp), phone number (Yelp), link to Yelp page (Yelp), category (Yelp) and opening hours (Google) were the top 5 pieces of information that we preferred to incorporate in the web page. They are displayed on the top portion of "Data analysis" page.

For data analysis, we conveyed topic modeling and sentence analysis on Yelp review data (from Yelp Open Dataset). For a good demonstration, we analyzed reviews from Las Vegas, because it has the most number of businesses in Yelp. We included all review sentences to conduct a topic modeling and provided the keyword bags out of the top five topics. Then we implemented the sentiment analysis on all reviews, with three reviews with the highest sentiment scores and another three with the lowest scores. Basically, these reviews can represent the most positive and negative comments towards the business from the public, and provide sufficiently valuable information to customers. Both keywords with the highest occurrence in the reviews and the 6 reviews obtained from the sentiment analysis will be shown. We used Python Jupyter Notebook for the data analysis. "Scikit-learn" package "CountVectorizer" and "LatentDirichletAllocation" functions were used for topic modeling. "NLTK" package "SentimentIntensityAnalyzer" function was used for sentiment analysis. Because the dataset is large, the Jupyter Notebook ran slowly locally. We created a Databricks in the Azure environment to use Spark API MapReduce functions to accelerate the computation. The detailed Python and Spark codes can be found in the project codes. When showing these data on the web page, loading data may be delayed, because the file size of the analyzed data is huge. To solve this issue, we parsed the analyzed data by zip-code. In this way, when loading data, it directs to a file with a certain zip-code first, then reads a much smaller size file. It would be ideal to conduct the same analysis with Google review data. However, Google only provides five reviews to download for each business. Thus, we showed these five Google reviews, together with analyzed Yelp reviews.

For data visualization, we planed to visualize business statistics by region. Again, we used Las Vegas as a demo. From Yelp Open Dataset, we calculated rating distributions for the region, average rating for each type of business, total reviews for each type of business, price distributions for restaurants, total number of business in different zip-code, counts of restaurants supporting certain features. All these graphs are shown in a new web page upon clicking the Las Vegas logo below "View Business Statistics for Las Vegas" on the main page. This information will give users an overall understanding of the region's businesses. Also, on the main web page, we implemented a quick search function underneath the "Place Searches". Upon clicking on "Hotels", "Restaurants" or "Groceries", up to 20 pins will show on the map to pinpoint businesses in each category. The pins change as mouse moving elsewhere on the map and click the same button again. This is based on Google's "search nearby" algorithm from the center of the map. We improved the search by applying a more accurate calculation of the search radius. In the "Data analysis" page, we displayed the followings (providing the business has sufficient information from raw data for the analysis): 1) top 5 pieces of general business information voted by team members; 2) top topics in a word cloud format based on the Yelp comments on this business; 3) plot rating distributions for each business based on Yelp reviews, with a bar chart of the counts for each rating; 4) Top 3 Yelp positive reviews; 5) Top 3 Yelp negative reviews and; 7) Google reviews, which are representative reviews by Google users.

5 Experiments/Evaluation

To test our application, please visit the website: Yelp on Google Maps (Figure 1).

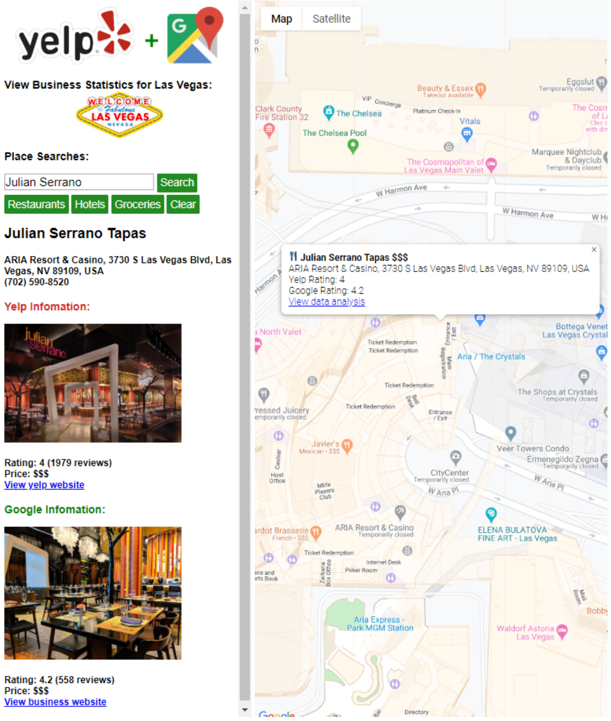
5.1 First Stage

Testing for this stage focuses on three aspects:

1) Test whether Yelp dataset and Google Maps information are shown as expected. Following the instructions on Yelp Fusion API, the request wasn't completed as expected. No pulled Yelp data was shown on the web page. Routing through a third party website has solved this issue. Tests have been done to choose a third-party website with a shorter query time.

For some POIs, there is no or little information from either Yelp or Google. In this case, we show "N/A" in the tooltip, indicating that the data is not available.

2) Test if Yelp and Google Maps information of the same POI matches. We checked manually by clicking different category POIs in different places and compared the latitude, longitude, name and rating information. Yelp data matches with Google data.



View Business Statistics for Las Vegas:

Place Searches:

Julian Serrano

Julian Serrano Tapas

ARIA Resort & Casino, 3730 S Las Vegas Blvd, Las Vegas, NV 89109, USA
(702) 590-8520

Yelp Infomation:

Rating: 4 (1979 reviews)
Price: $$$
View yelp website

Google Infomation:

Rating: 4.2 (558 reviews)
Price: $$$
View business website

Julian Serrano Tapas $$$
ARIA Resort & Casino, 3730 S Las Vegas Blvd, Las Vegas, NV 89109, USA
Yelp Rating: 4
Google Rating: 4.2
View data analysis

yelp +
WELCOME TO Fabulous LAS VEGAS NEVADA
Search
Restaurants
Hotels
Groceries
Clear
Map
Satellite
Beauty & Essex
Eggslut
Clark County Fire Station 32
The Chelsea
The Chelsea Pool
Vitals
The Cosmopolitan of Las Vegas Main Valet
Marquee Nightclub & Dayclub Temporarily closed
W Harmon Ave
W Harmon Ave
North Valet
Ticket Redemption
Entrance / Exit
Main Registration
Aria / The Crystals
Bottega Veneta Las Vegas Crystals
The Shops at Crystals Temporarily closed
Pressed Juicery Temporarily closed
Javier's Mexican · $$$
Veer Towers Condo
Ermenegildo Zegna Temporarily closed
CityCenter Temporarily closed
W Aria Pl
ARIA Resort & Casino Temporarily closed
ELENA BULATOVA FINE ART - Las Vegas
Bardot Brasserie French · $$$
Poker Room
Aria Express - Park MGM Station
Waldorf Astoria Las Vegas
Google

Figure 1: Main page of the project (click here for live demo). In the left panel, there are 1) Yelp+Google logo; 2) View Business statistics (click on "Las Vegas" Icon to show new page); 3) Places Searches; 4) Quick searches of restaurants, hotels or groceries; 5) Detailed business information from both Yelp and Google will be displayed when mouse clicking a POI. The left panel is the map area, after mouse click of a POI, a tooltip shows both Yelp and Google information (click on "View data analysis" will open a new page). Note: We use red color to indicate data from Yelp and green for Google.

3) Test the search function. The default search function from Google API detects the IP address and process searches in the local area. We changed the function so that we can make searches starting anywhere on the map. This function has been tested and validated.

5.2 Second Stage

Testing in the second stage includes three aspects.

1) Test if the information is accurate. We planed to manually check some POIs and also pull the data from our deployment, followed by comparing them with the original data to make sure our data was accurate and intact.

Through manual checking, we confirmed that the information was accurate.

2) Test the data analysis insights. We built two Natural Language Processing models: topic modeling and sentiment analysis. Each team member manually checked 10 POI results and summarized the findings. We need to answer two questions: the first one is whether the returned keywords from the topic modeling provide insights for customers to understand the POIs better, and if the top reviews returned by the sentiment analysis accurately represent the most positive and negative reviews from Google and Yelp.

After opening the "View data analysis" page, a data visualization of word cloud is shown. The word bags are from the topic modelling. The keywords make sense and can help customers to better understand the business. Top 3 Yelp positive reviews and top 3 negative reviews from Yelp are shown, and the reviews returned can represent their assigned categories. Five reviews from Google are also shown.

3) Test the data visualization. We would show six graphs on the Las Vegas business statistics page. Each graph needs to be nicely presented with the right scales and delivers enough information for users to understand the region. The quick search function will pin up to 20 businesses in designated categories. If mouse moves the map elsewhere and re-clicks the button, 20 pins will show in the new region. The "Clear" button will clear all pins. In "Data analysis" page, both keywords word cloud and rating histogram should show nicely.

After implementation, clicking the "Las Vegas logo" beneath "View Business Statistics for Las Vegas" shows six graphs that represent the local business information. The quick search buttons can be found under the search bar. Customers can quickly search for Restaurants, Hotels, or Groceries in the area. After clicking the quick search buttons, 20 pins are shown correctly. By clicking the "View data analysis" page of a POI, a new page pops out, showing more detailed information, including general information, a word cloud data visualization representing the keywords from topic modeling. A Yelp review star histogram is also shown for the rating distributions of this specific business. Other information such as analyzed top Yelp and Google reviews are also shown correctly. In summary, all data visualizations are validated as expected.

6 Conclusions and Discussion

We completed all the tasks at both stages. Team members are flexible to make changes to these tasks through out the process. The final product of this project, Yelp on Google Maps, captured information from both Yelp and Google Maps. It provides a comprehensive comparison between Yelp and Google data.

During this process, team members learned: 1) Data collection skills using Yelp and Google APIs. 2) Data analysis skills using Python. 3) Data visualization skill using D3 and Google Maps. Besides the skills, we also understand that communication is critical for group collaboration.

Both Yelp and Google provide APIs for data extraction. Compared with Yelp, data retrieving from Google is much faster and easier. Google APIs also has more contents and more types of API. It may be because that Google APIs use Google Cloud platform, which is a commercial platform. Thus, we used Google APIs more than Yelp API. On the other hand, Yelp provides a static dataset free to download. From this dataset, we performed multiple analyses and added valuable information to the project. However, it is very difficult to download batches of information from Google (both with and without API). For example, without purchasing the business Google APIs, it is only allowed to retrieve up to five google reviews for each business, which restricted the combination and comparison of Google Maps' and Yelp's reviews. Therefore, some data analysis results are limited to Yelp dataset. If in the future Google dataset becomes available to download, our project could be further improved with analysis on the data from Google.

We conducted a survey among team members when selecting information to be included for the general information section. It turns out that different people would have different preferences. In order to attract more users, it would be worth conducting surveys among a wider range of users to study what information is preferred.

7 Activities

Feb 17-28, Proposal (WD, RW, MS), Presentation Slides (JC), Video (BL).
Feb 19-Mar 6, Get familiarized with Yelp dataset analysis(WD, BL), geographic data visualization(MS, JC) and Google Maps API(RW).
Mar 6-16, First stage project construction and coding (WD, BL).
Mar 17-20, Midterm test for project functionalities (RW, JC, MS).
Mar 27, Progress Report (Draft:WD, MS. Revision: RW, JC, BL).
Mar 21-Apr 6 , Second stage project construction and coding (MS, JC, RW).
Apr 6-12, Functionalities Final test (WD, BL)
Apr 13-17, Final Report (all members).

8 List of Innovations

1) Took the advantages of both the popularity of Google Maps and the comprehensive information on Yelp. Incorporated Yelp dataset information directly into Google Maps. Compared Yelp and Google information and visualize the comparison.

2) Employed both Yelp Fusion API (a live dataset) and Yelp Open Dataset (a static dataset).

3) Utilized the powerful Google Maps' search function.

4) Applied two Natural Language Processing models.

5) Used Databricks Spark MapReduce functions to accelerate the data analysis computing speed.

References

1 Chen, Pei-Yu, Shin-yi Wu, and Jungsun Yoon. "The impact of online recommendations and consumer feedback on sales." ICIS 2004 Proceedings (2004): 58.

2 Anderson, Michael, and Jeremy Magruder. "Learning from the crowd: Regression discontinuity estimates of the effects of an online review database." The Economic Journal 122.563 (2012): 957-989.

3 Huang, James, Stephanie Rogers, and Eunkwang Joo. "Improving restaurants by extracting subtopics from yelp reviews." iConference 2014 (Social Media Expo) (2014).

4 Asghar, Nabiha. "Yelp dataset challenge: Review rating prediction." arXiv preprint arXiv:1605.05362 (2016).

5 Sawant, Sumedh, and Gina Pai. "Yelp food recommendation system." (2013).

6 Sawant, Sumedh. "Collaborative filtering using weighted bipartite graph projection: a recommendation system for yelp." Proceedings of the CS224W: Social and information network analysis conference. Vol. 33. 2013.

7 Sopan, Awalin, et al. "Community Health Map: A geospatial and multivariate data visualization tool for public health datasets." Government Information Quarterly 29.2 (2012): 223-234.

8 Koua, E. L. "Using self-organizing maps for information visualization and knowledge discovery in complex geospatial datasets." Proceedings of 21st international cartographic renaissance (ICC) (2003): 1694-1702.

9 Svennerberg, Gabriel. Beginning Google Maps API 3. Apress, 2010.

10 Gibin, Maurizio, et al. "An exploratory cartographic visualisation of London through the Google Maps API." Applied Spatial Analysis and Policy 1.2 (2008): 85-97.

11 Rahmi, Anisa, I. Nyoman Piarsa, and Putu Wira Buana. "FinDoctor-interactive android clinic geographical information system using firebase and Google maps API." International Journal of New Technology and Research 3.7 (2017).

12 Doshi, Pankti, Pooja Jain, and Abhishek Shakwala. "Location based services and integration of google maps in android." International Journal of Engineering and Computer Science 3.03 (2014).

13 Vladimir Nikulin. "Hybrid Recommender System for Prediction of the Yelp Users Preferences." Industrial Conferences on Data Mining (2014).

14 Danielle Albers Szafir, Steve Haroz, Michael Gleicher, and Steven Franconeri." Four types of ensemble coding in data visualizations. "Journal of Vision (2016).

15 Mukta Paliwal, and Usha A. Kumar. "Neural network and statistical techniques: A review of applications." Expert System with Applications. 36.1 (2009): 2-17.