Jackie Zhang(998836564), Sridatt Bhamidipati (998993083)
ECS 163 Information Interfaces
March 13, 2016

Final Project Proposal

Introduction

Yelp is an application that allows users to review businesses they visit. As avid users of Yelp, we decided to use Yelp's dataset, which is in JSON format and contains various details in depth regarding businesses, users, reviews, and more. The dataset provides us with rich information regarding businesses from 10 different cities and 4 countries; initially we chose to focus on Pittsburgh, Urbana-Champaign and Las Vegas but because of efficient parsing of the data, we were able to do all U.S. cities. We also decided to write a python script to convert the JSon files into CSV because it is a format that goes extremely well with D3 and the frameworks that we use.

We always wondered how urban planning plays a role in the success of a business, and in particular the business' rating on Yelp. The two questions we aim to answer is "How much of a business is really just location,location,location?", and "Does the type of restaurant in a location affect the rating it has?" We also investigates the relationship between businesses and reviews.

Initial Ideation/Proposal

Below, we outlined what exactly we plan to do with the dataset initially and why it would be useful to see.

There are many ways to visualize this data since there are many categories that we can tap into but the main one would be the location. From the data "business" we want to use the key "address" to first get the address of all the business in the area. After that, we want to set parameters of how big the location is and potentially use google maps APIs to display the map on the webpage. We then want to get the ratings of the businesses in the area from the data called "review" part of the data. We decided to focus primarily on location. We would like to have a general layout of the businesses by integrating the Google Maps API and plotting addresses of the businesses from the "business" section of the dataset. Next, we want to compute the average rating of businesses in a general area and represent it by drawing a colored circle (or rectangle)

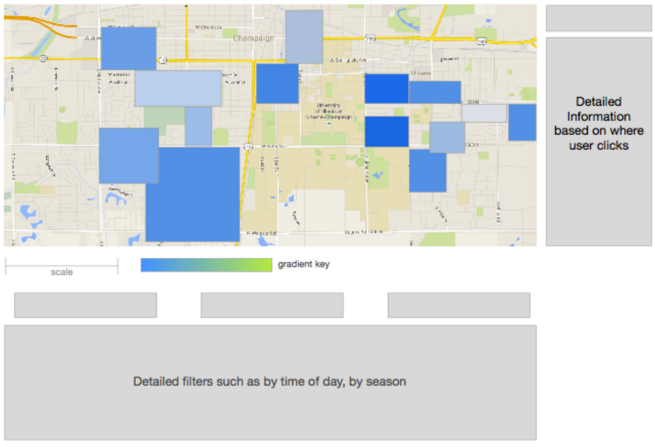


Figure 1.1
Initial Prototype of initial view of map

over the area. We can tweak the color of the circle to visualize more variables at once, such as a darker color for a more positively reviewed business area, and a lighter color for a negatively

reviewed business area. We would be careful when choosing our color palette to make sure the visualization is easy on the eyes and the background contrasts nicely with the circles without interfering. For the big large area view of the map we would have big transparent circles with colors that determines the average rating of an area. We will have a scale either towards the side or in the bottom to show where the user can see how the rating is for each place. We will also make sure the color would be in the lighter side and contrasting so it would be easier for the user to see the different areas if they overlap and also so that the colors won't interfere with the map itself.

The purpose of having an map is because it would be easier for the user to see patterns from each locations than just having an visualization of percent in a chart. It also would hopefully allow us to analyze what type of places would have a greater impact in terms of places like schools, party area, etc. We are hypothesizing that students are probably stricter in terms of ratings and also certain types of food would be more popular in some locations than other. An extra feature that would give us more detail information would be to have a time scale where we can go through the different seasons. It could give us insights where depending on the season of the year, one would give certain ratings. The way it will work will be as the user drag the time scale, the bubble would change color based on the average rating.

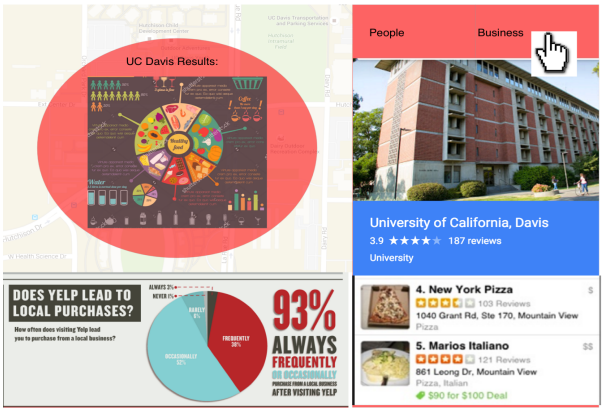


Figure 1.2
Business Data Graph

In terms of interactions for the visualization, one would be able to select and zoom in to a certain location. One would be able to get more information for each of the locations but if they

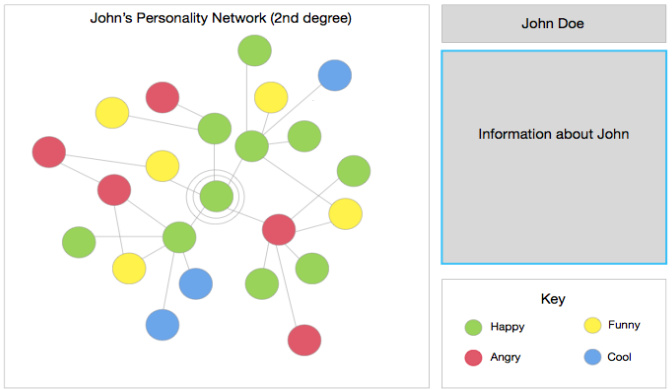


Figure 1.3
Relationship Graphs

want to just look at a place where an average review bubble is placed, they can as well. In the

more detailed view, we would be able to show the user a more detailed version and more specific information. There would be actual pictures of the restaurants when one hovers over as well as star ratings and the type of restaurant it is. Also when one hovers over a place, they would be able to get the time the restaurant is open and closed. The reason we are choosing to add more detailed versions is because it would allow the user to possibly have more insight on other potential ways that could affect a business rating. Who knows, maybe it is the appearance of a restaurant in a location or it could be as simple as when a restaurant is closed. If time permits, we would also add comments and other features that the dataset provides for us.

We also wish to take advantage of our detailed user data and analyze the personalities of Yelp users, hopefully to find interesting correlations with the users' personalities and other aspects of businesses on Yelp. We want to map the amount of reviews that they give in relation to their location, which would give us an insight on where yelp users are more prominent, or more willing to review businesses. We also would like to show the relationship between the number of reviews and the average ratings of a business, as well as show which categories of businesses attract which personality types. We plan to do this by utilizing the IBM Bluemix API to interpret the reviews on the business and use Natural Language Processing techniques to tell us exactly how the customer feels about the place.

Implementation Process

We learned a whole lot about using different APIs throughout this process and were able to successfully make cleaner design choices as we moved forward. The first thing we had to do was data parsing. This was a difficult task because the files were extremely large and sometimes were unable to open. Originally, we took the first 1000 lines of the csv file. The problem with this was that the users, businesses and general data were not in any particular order. For example, chipotle was listed on line one in business.csv and one of the users reviewing it was on line 7000 in review.csv. We then decided to do some python code and commands from the command line to split the files up. The commands that we used was: "split -1 200 {example}.csv" and "for i in *; do mv "$i" "$i.csv"; done" This worked, but our program still ran slowly due to the sheer amount of data. That's when we decided to condense our data. For example, we removed the "friends list" array from user data, which ended up saving a lot of space. For review.csv, we took out the actual text reviews, since we noticed that the shorter, concise text from tip.csv was interchangeable. We downloaded windows 10 into virtual box and then use a tool called delimiter to get the specific rows we want. We were able to condense multiple gigabytes worth of data to less than 100mb!

From there, we were able to start implementing the map. Originally, we wanted to use Google Maps API but we wanted to avoid the complications that the API doc seemed to bring. After looking through various map API's, we found that Mapbox.js fit our needs the best. We also used jQuery to make it easier to do JavaScript code. Most of the things we initially wanted "works" but the main issue was due to the density of the information and how some information actually isn't giving us any trend or any information we need.

Design Changes and Justification

The first design we change was on the initial graph. We initially wanted to cluster the map into separate "neighborhoods" using blue circles to highlight the different neighborhoods, with the hue of the circles representing the average rating of the neighborhood. We eventually

did implement this and were able to discover that yelp users in densely populated locations don't put in detailed information such as their "neighborhood", but rather left that category blank. It was interesting how we were able to see the trend where based on location, the quantity of information filled in per review is extremely different.

Because of this, we decided to cluster based on business location instead. We noticed that when there were many businesses in one concentrated place, and from that it was much easier to visualize different neighborhoods We use Mapbox to highlight the area with a polygon. We also added the extra features of changing the colors based on the amount of clustered points there are as well the feature where you can zoom in or out of the map. The reason we have to use different colors for different clusters is because it will allow users to see if they can zoom in further for more details. Once they zoom in to a certain point (when the markers are blue) they would be able to see the specific businesses.

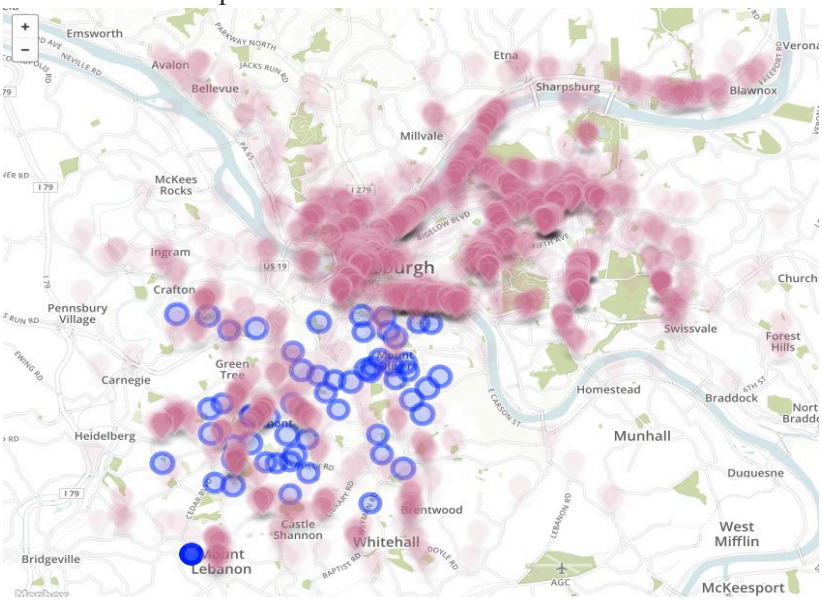


Figure 2.1: Original Clustering

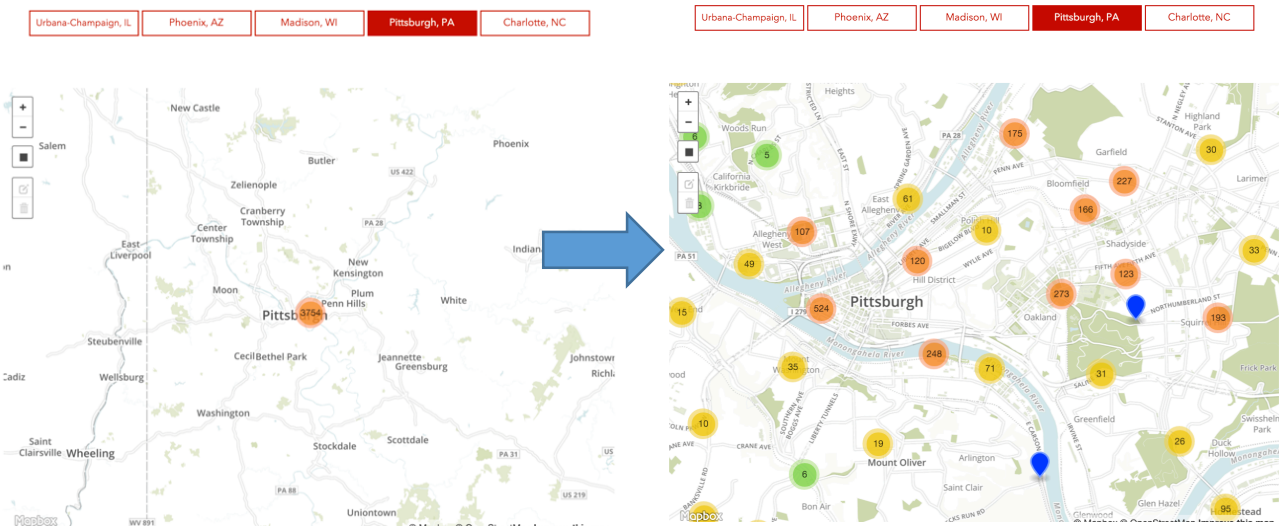
The next design change was taking out the friend "network" data. There are a couple of reasons for this. The first reason was that each user has a huge amount of friends and the data would be too dense. The second reason is that because most of the data is related to business, it would make more sense to focus on the businesses and their reviews for better flow. For the scope of this project, it did not look like something that is viable. Finally, as we were doing the project, we made a lot of UI changes so it is clean and easy to see as well as the color scheme is closely related to Yelp's normal color scheme.

Final Visualization Description

Our final visualization is focused primarily on businesses. The visualization would show a lot of relationships based on the following csv files we used: business data, check in data, review data, tip data, and user data. Below, we will map out the different features we have as well as what we were able to analyze from each feature.

Feature 1: Pan and Zoom / Clustering and remapping by Color/chunks

For this feature there is a plus and minus button in the top kind of like how a typical google map looks like so it is extremely intuitive. Originally, we did blue dots, we noticed that it was concentrated in the outer layer of the map and difficult to understand due to how clumped they were. From this, we interpreted that when there are less concentrations of business, there are more information versus in concentrated areas. We were also able to determine the different



"neighborhoods" in the map.

Figure 3.1: Feature 1 images

Feature 2: Top 10 Users by Reviews over time (Animation)

What it does is that it makes a marker for the specific user pop up. Then it will animate a path to all the user reviews over time. We focused on Pittsburgh's data. We were able to observe that there are 3 main clusters that the users go to. They would typically go to a certain neighborhood during the earlier time of the week and then after 2 or three days they would go to the neighborhood adjacent to it. We assumed that is because of the night life (bars, dinners, etc) that caused the users to go to specific locations.

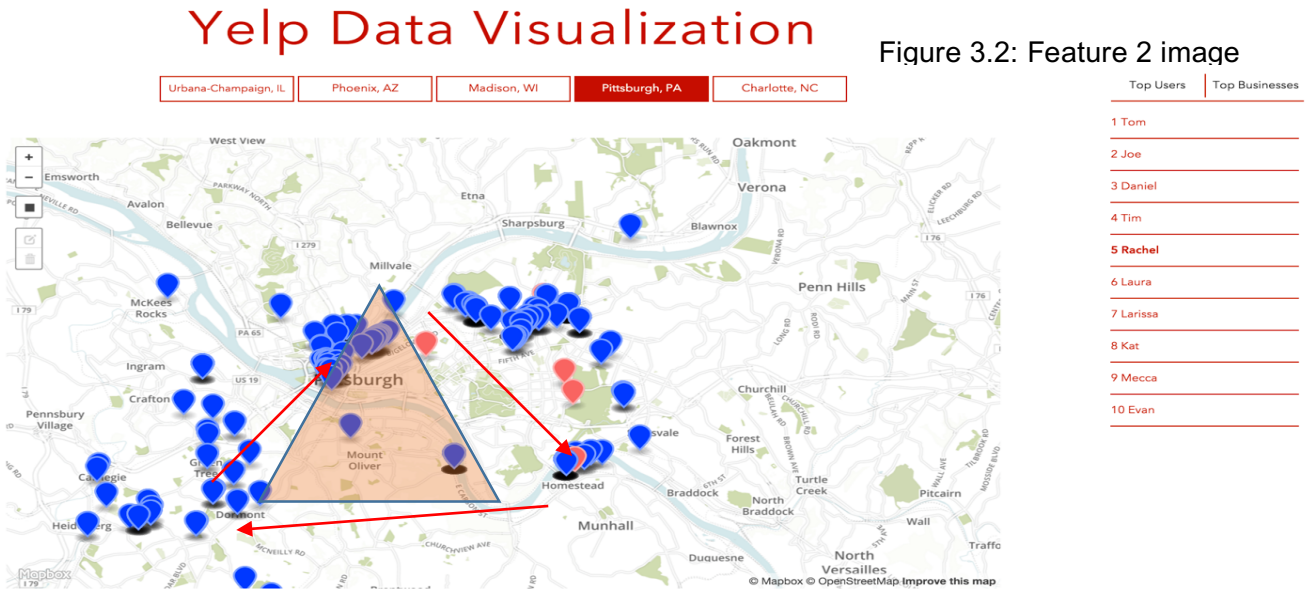


Figure 3.2: Feature 2 image

Feature 3: Selection by Rectangles and Querying Average Star Rating

We are able to search all the businesses from user input or by using the select bar by the side. With this feature, we were able to discover that in general, places that are less concentrated have higher ratings and places that are concentrated have lower ratings (Pittsburgh). This is probably because if one is in a less concentrated place and want to do a review, has a high possibility of being a good restaurant. Other trends trends that could be noticed are if businesses have a higher average star count nearby rivers or downtown type areas, or lower stars nearby railroad tracks, for example.

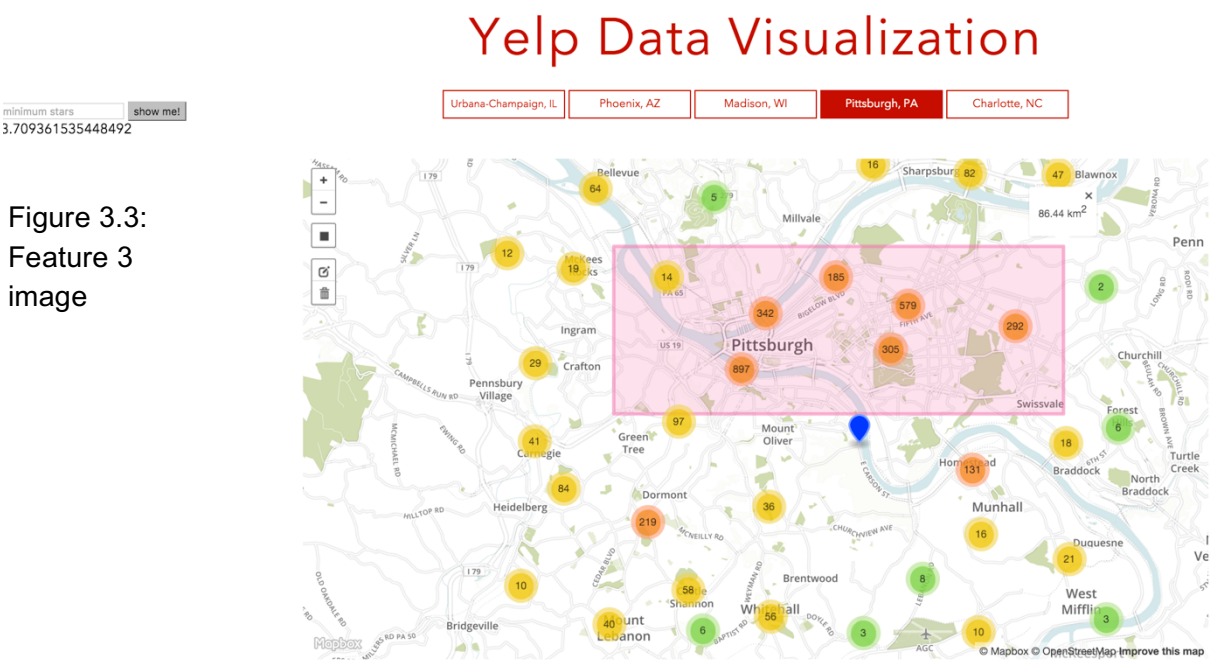


Figure 3.3: Feature 3 image

Feature 4: Top 10 Business Data→Ratings over Time (Line Graph)

In this graph we had the trend where the x axis is the months of the year and the y axis being the ratings of the companies. There don't seem to be an obvious trend but we do notice that for cafes in Pittsburgh, the ratings typically go up during the fall/winter season. We guess this is probably because Pittsburgh is sort of a college town hence, there are more students appreciating cafes to study in and relax.

Feature 5: Top 10 Business Data→Amount of Check-In versus Days (Bar Graph)

In this data we plot the amount of check in for businesses (y-axis) versus the day of the week (x-axis)it is. We notice an obvious trend for most places in all the cities. We notice that there is always a hike in the amount of check ins on Thursday-Saturday. We guess this makes sense since those days are the days to go out to party or drink and all these places are pretty vibrant with their night life.

Feature 6: Top 10 Business Data→Types of Reviews (Pie Chart)

This is a data pie chart that shows the quality of the reviews for the users. It is split into three: usefulness, coolness, and how funny it is. We only got the quality of reviews from top 10 most reviewed users. We thought there would be more of a difference in the values but in actuality, "usefulness" is the quality that was the most following coolness and funniness. This shows that useful reviews are the ones that tends to get users to the top. This also make sense because most

people who use yelp are for the basic information and not reviews like those from amazon or facebook.

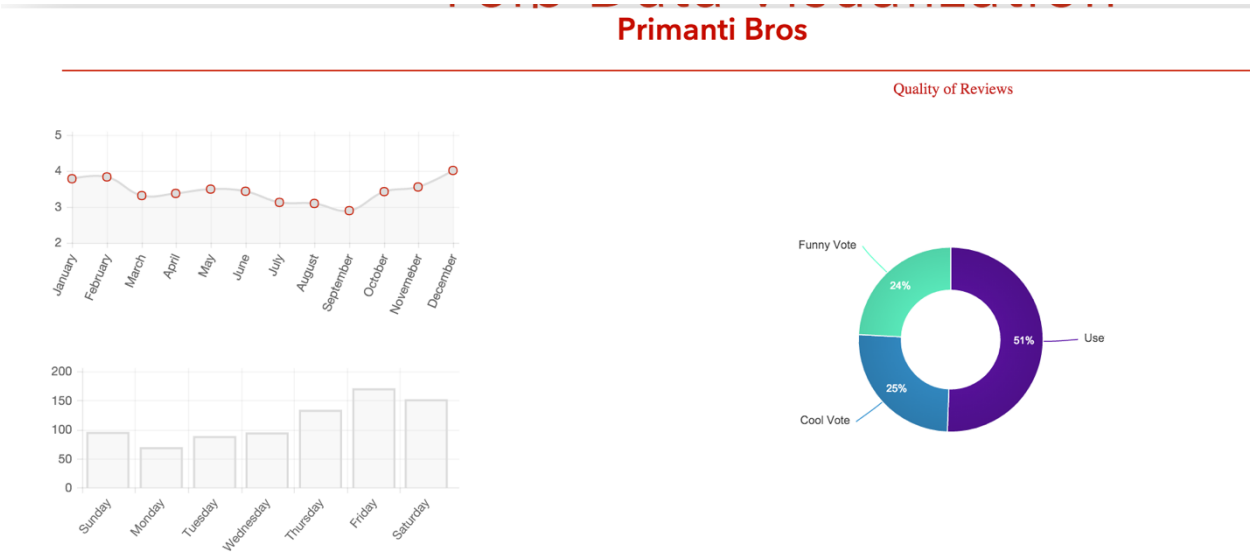


Figure 3.4: Features 4 - 6

Group Delegation and Extra Credit

In this project we did mostly peer programming. Jackie did most of the data parsing in terms of making sure all the data is usable and also made sure the variables can be used in the function. His main focus was the business data so that he also got the bulk of the d3 code skeleton functions to use. Sridatt implemented the functions and made most of the mapping functions. He also worked on the general UI/UX for the screen.

We implemented temporal animation of each particular top user, so that we could get an insight on their trends. The animation shows the user traveling from one place to another, depending on where they "tipped", in timely order.

We gave the user a lot of freedom in this project. The user can filter the plots on the graph by inputting a minimum star rating. In addition, the user can select his or her own query region and get an average rating count. This allows the user to dive deeper into analysis by picking different sizes, areas, condensed vs dispersed areas, etc.

In our attempt for getting the 20 points extra credit, we implemented more than 2 features and used practically every data available to make an interface that is definitely not intuitive to see from the million lines of code that was provided to us. We were able to go through different interfaces seamlessly with the data updated with every single screen we are on. We also worked on improving the UI with updated colors and layout in attempt to get better results.

Commands to run

Go to directory of the folder of the project

```
python -m SimpleHTTPServer 8000
```

Final Product
Below are images of our final product: