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.

**Initial Ideation/Proposal**

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

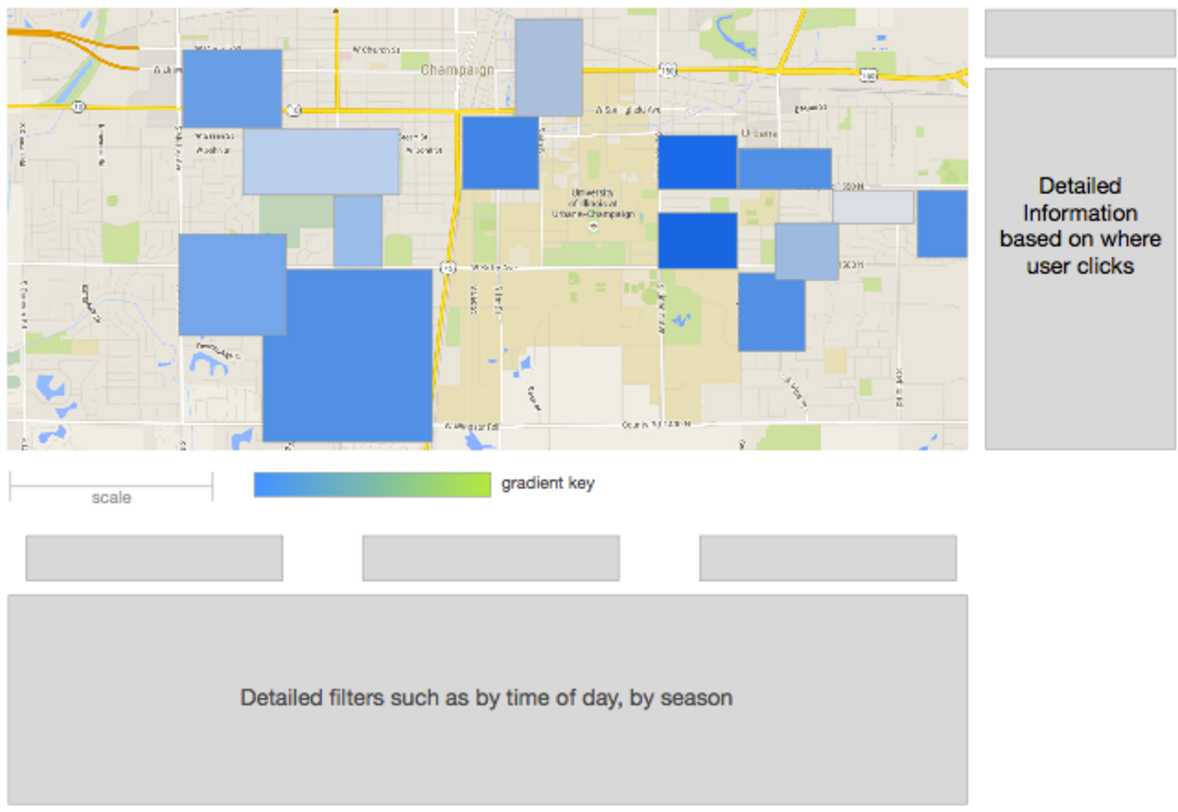
 We always wondered how urban planning plays a role in the success of a business, and in particular the business’ rating on Yelp. The two questions we aim to answer is “How much of a business is really just location,location,location?”, and “Does the type of restaurant in a location affect the rating it has?” 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) 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.

Figure 1.1

Initial Prototype of initial view of map

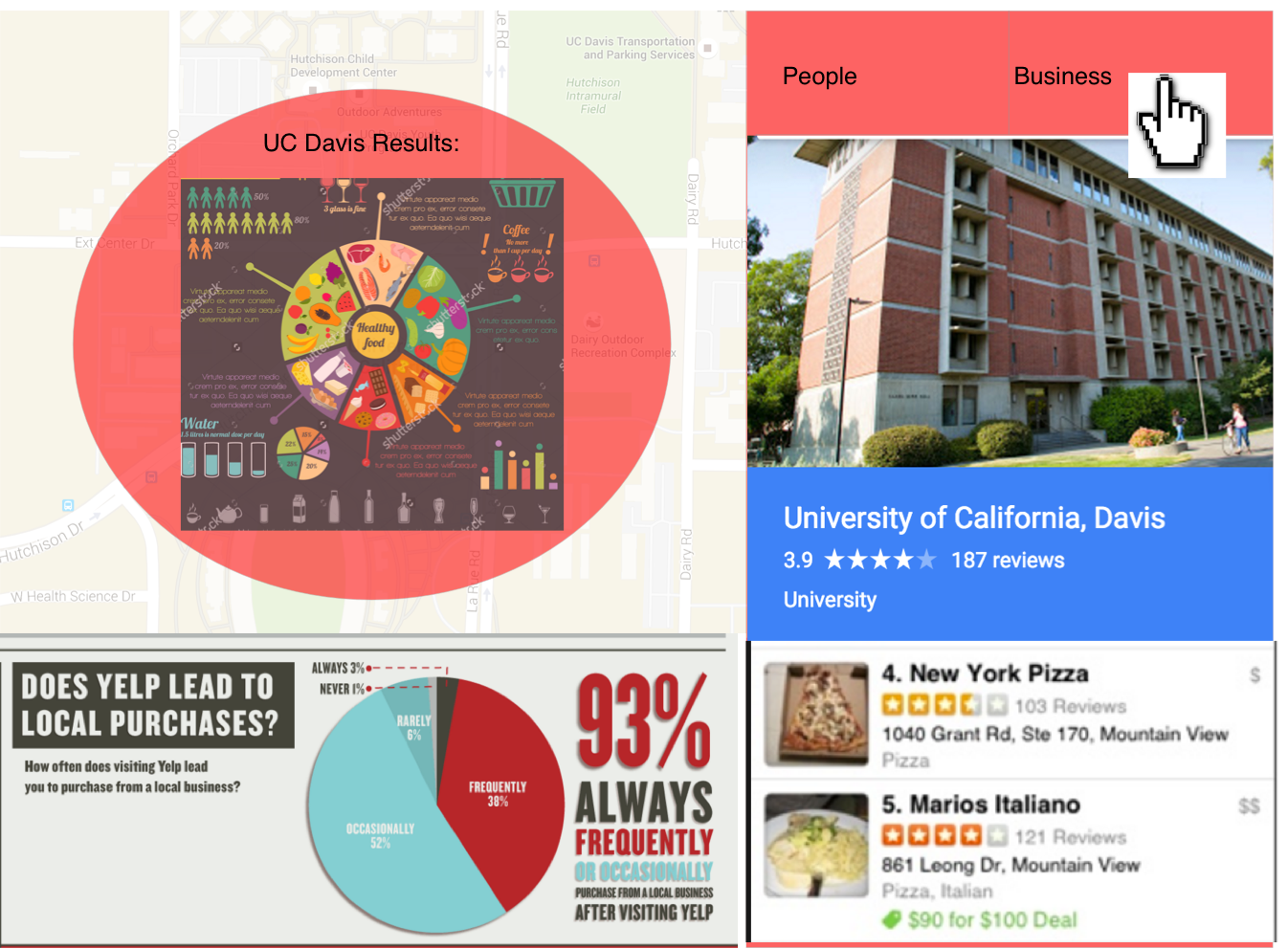


Figure 1.2

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 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.

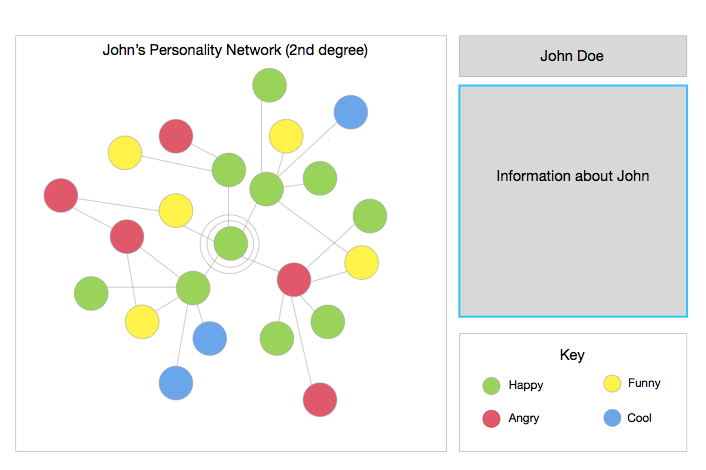


Figure 1.3

Relationship Graphs

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 was able to successfully make new and cleaner design choices. The first thing that we had to do was data parsing. That was one of the hardest part of the task because the files were extremely large and we had to find ways to make it useable. Originally, what we did is that we just took the first 1000 lines of the csv file. The problem with this was that the users, businesses and general data are not in order. An example would be that chipotle was listed on line one in business.csv but then the user reviewing it is in line 7000 hence we can’t just parse 5000 of code per file. We then decided to do some python code and command line to try 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 doesn’t change the fact that we want to run files of the same size. That is when we decided to think smartly and create schemas to condense our data. The first thing was that of the user data, for the user data we decided to take out friend list, because we notice that there are too many outliers to give us accurate information. For review.csv, we took out the actual reviews because we noticed that tip.csv gives us pretty much the same information but shorter. 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 get over 1gb worth of data condensed to less than 100mb!

From there, we were able to start implementing the map feature. Originally, we wanted to use Google maps API but then as we started using the API, we noticed that some of the features we wanted related to the animation might of costed money as well as it was hard to use. Because of the large dataset, Google Maps was lacking a lot of the features. We then did go on and did some research related to mapping since D3 isn’t exactly the most efficient tool to make maps. We then discovered mapbox.js which was extremely useful for coding the project. 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. The data of the neighborhood would be based on the input the dataset has. We eventually did implement this and was able to discover that it seems that the general trend is that yelp users in densely populated locations don’t put in detail information for things like “neighborhood”. 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.

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.

In terms of delegating the work for the project, since both of us are proficient in Javascript programming, we will take more of a peer programming approach to the project. In particular, Jackie would take pittsburgh and focus on visualizing Pittsburgh’s part of the data and Sridatt would take Las Vegas and both would take Urbana-Champaign. Jackie would mostly parse the business data while Sridatt would mostly parse the review data.

* Image of dataset with blue dots
  + Noticed that people in outskits in cities put more comprehensive data in the negihborhoods. This is probably because
* <https://www.mapbox.com/mapbox.js/example/v1.0.0/opacity/>