Yelp Business Trajectories - Group 3

Palak Arora (pa1154@nyu.edu)
Laura Buchanan (lcb402@nyu.edu)
Bhagya Ghumalla (blg332@nyu.edu)
Suhita Goswami (sg4565@nyu.edu)

Github: https://github.com/NYU-CS6313-Fall16/Yelp-Business-Trajectories-3/
Video: https://drive.google.com/file/d/0B2f0Vr8sYSLMMURfM0VjUG5TQWc/view
Demonstration: https://nyu-cs6313-fall16.github.io/Yelp-Business-Trajectories-3/

Problem:

The aim of our data visualization is to observe the impact, if any, of Yelp tips on the Yelp business rating. While Yelp data is analyzed for a number of purposes, we suspected that Yelp tips were often neglected in academic research when compared to analysis of dynamic Yelp rating, Yelp reviews, or Yelp user activity. We saw a potential opportunity to look into Yelp tips to find new insight that would have value to businesses hoping to improve their Yelp rating.

Questions:

To quantitatively investigate how Yelp tips, by their sentiment or count, influence business rating, we focused on two questions: is the average sentiment of tips and the rating of a business correlated? and does Yelp business rating increase with more tips written?

Using fundamentals in the field of Information Visualization, we knew that a clear way to inspect the presence of correlation between two variables is with a scatterplot. Therefore, we choose to use a scatterplot as the focal point of our visualization. To look at the particular language that was driving this sentiment, we also built a word cloud to display the words associated with each businesses' tips.

Dataset:

To answer these questions concerning Yelp tip influence, we employed data provided by the Yelp Dataset Challenge.[1] This is a publically available dataset is intended for academic research, and is made up of several district data frames providing information about business attributes, reviews, tips, photos, etc.

Attributes:

Business Dataset:

Business ID

Business Name

Price Range

Category

State

Business ID was used as our primary key on which we would merge the other required datasets. Business name was extracted so that it could be displayed in our final visualization. Furthermore, Price Range was displayed in our final visualization. Category was used as precursor for our derived feature *Broad Category* attribute. We used the attribute State to filter on businesses located in Nevada.

Reviews Dataset:

Business ID

Business Rating

In the Business Dataset, business ratings are averaged to the nearest 0.5 stars. Because we were interested in treating rating as a continuous variable in our analysis with scale finer than

[1] Yelp. "Yelp Dataset Challenge." *Yelp*. N.p., n.d. Web.

0.5 stars, we instead took the Business Rating from the Reviews Dataset. In this dataset, there is one data instance per review written. We used Business Rating to compute *Average Business Rating*.

Tips Dataset:

Business ID

Tip Text

From the Tips Dataset, we collected the Tip Text to compute the *Number of Tips Per Business*, *Average Business Sentiment*, *Word Sentiment*, *Inverse Log Word Frequency*, and *Set of Words in Businesses' Tips*.

Data Analysis:

Data Analysis was primarily performed with Python and the library pandas. From each of our three data sets, Business ID was extracted as our primary key on which we would merge data. In the Business Dataset, each instance in a unique business. With this dataset, we first filter on State == Nevada. We came to the decision to select this data two balance two opposing views; on one hand, found that using all available locations of Yelp data was cluttering our visualization and potentially added noise due to differing tip behavior in different areas; on the other hand, Las Vegas is the largest metropolitan area in the Yelp Dataset Challenge and therefore provided a large cohesive slice of data. (We chose to use State == Nevada rather than City == Las Vegas so that businesses in the Las Vegas cosmopolitan area that were outside of Las Vegas proper were included.)

Next, with the Business Dataset, we derived the Broad Category from the Category description provided for each business. Because we were using a scatterplot and were interested in displaying business category with color on our scatterplot, we knew it would be best to limit the number of categories so that the distinction portrayed by their color would be clear. Since there are thousands of unique values for the Categories attribute in the Business Dataset, we

composed the Broad Category attribute by searching for keywords. For example, it a Category description included "Restaurant" or "Food," we tagged that business with the Broad Category "Eateries." We decided on five Broad Categories: Eateries, Lifestyle & Entertainment, Health & Beauty, Auto, and Others.

As described above, we were interested in deriving Average Business Rating from the Reviews Dataset. Each instances in this dataset is an individual review. To compute the Average Business Rating, we grouped businesses by Business ID and then took the mean of the Business Rating.

In the Tip Dataset, each data instance is an individual tip. To collect the tips, we grouped by Business ID, counted Number of Tips Per Business, and concatenated the Tip Text. For preprocessing, we lowercased all characters in the per business Tip Text string and removed punctuation.

In addition to our cleaned dataset with each business as a data instance, we created a word look-up dataset. In this dataset, each word was a data instance, gathered from the set of words appearing in all pre-processed Tip Text strings. Then, for each word, we calculated the Inverse Log Word Frequency for each word from all Tip Text strings. Using the methods described in Turney and Littman 2003,[2] Word Sentiment was derived. Word Sentiment was a score between 1 and -1, with values closer to 1 denoting more positive words, and values closer to -1 denoting more negative words. Finally, using the set of words in each businesses Tip Text string, the average business sentiment was computed.

Initial Mockup:

Our initial visualization mock-up was very ambitious. We were interested in answering a variety of questions including:

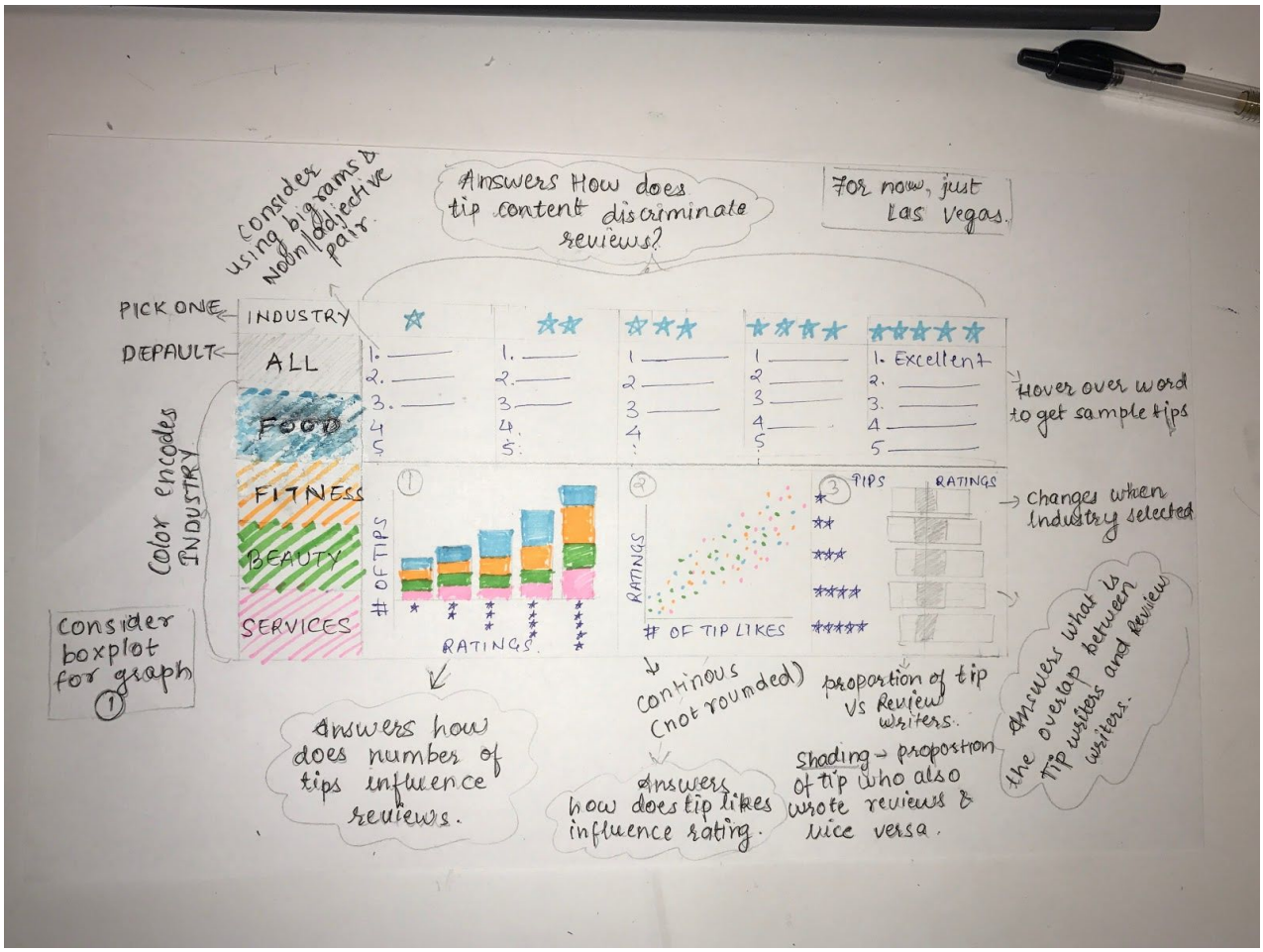
- How does tip count vary between industries?
- Does the business price range impact the tips and the rating of a business?

---

[2] Turney, Peter D., and Michael L. Littman. "Measuring praise and criticism: Inference of semantic orientation from association." *ACM Transactions on Information Systems (TOIS)* 21.4 (2003): 315-346.

- Are the frequency of tips and reviews correlated and do users who write reviews also write tips?
- How does the content of tips influence the rating of a business?

We envisioned an interactive visualization that would incorporate nearly every attribute available in the Yelp Challenge Dataset. Our first mock-up included these features:



However, it became clear that the complexity of the visualization would limit the insight that we could gained.

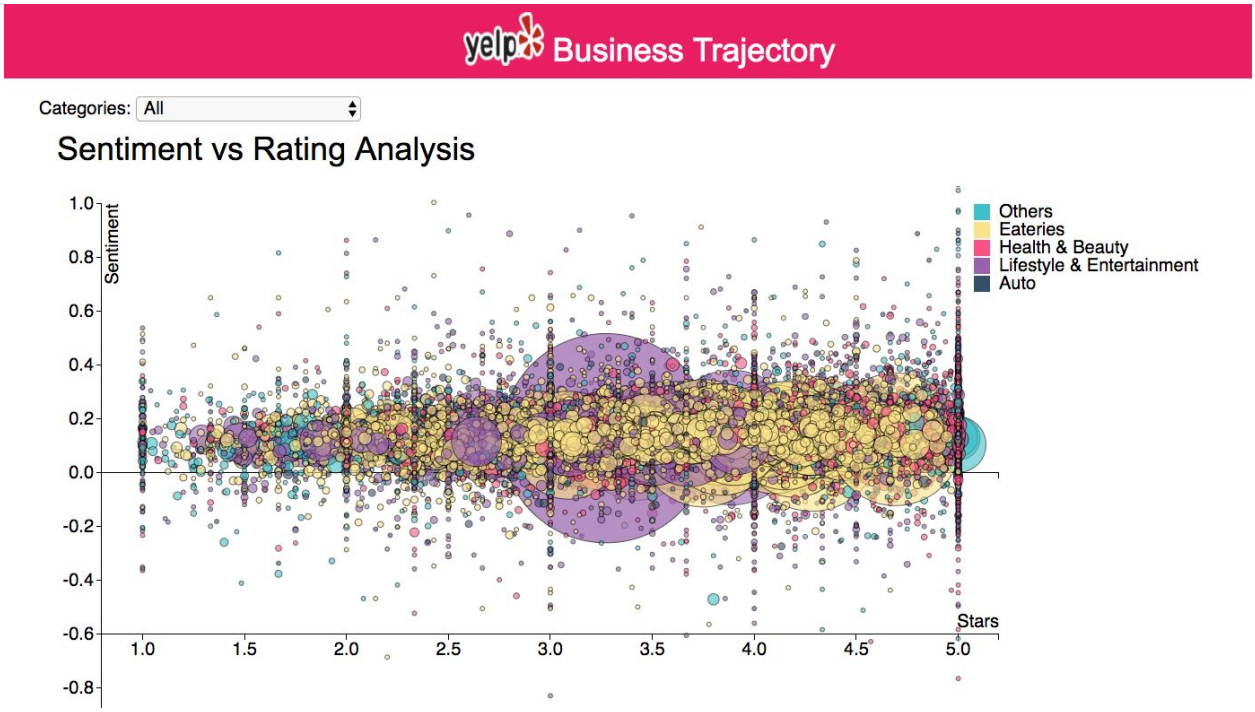
Intermediate Progress:

As we progressed with the project, we moved from a interface in which several graphs would be displayed, to a cleaner view with a line chart depicting Average Business Rating vs. Number of Tips Per Business with filters for Category, Price Range, Time (later excluded entirely), and Location, along with a word cloud displaying words unique to One through Five Star Ratings.

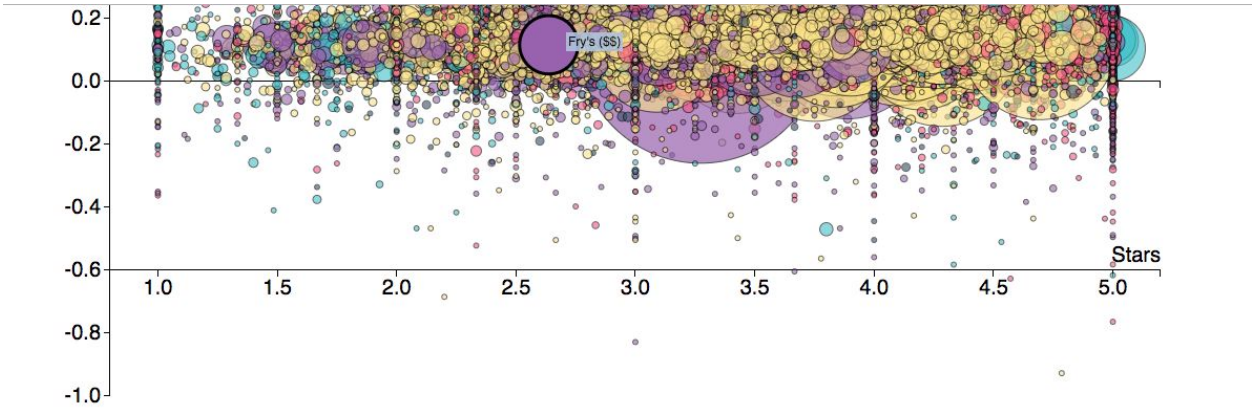
While we worked on this attempt, we again became frustrated with the complexity of this proposal and abandoned it. We decided that we would not filter on Price Range, Time, or Location. Furthermore, the word cloud envisioned was simplified to avoid the need for further data processing.

Final Visualization:

In our final visualization, we created a scatterplot that displays each business as a circle with a light gray border along the dimensions of Average Business Rating and Average Business Sentiment. The size of each business's point on the scatterplot indicates the magnitude of Number of Tips Per Business. The opacity for each point was set to 0.5 so that overlapping business points could be seen. The color of the point encodes the Broad Category. The colors for each Broad Category were chosen so that they were maximally differentiable while being aesthetically pleasing. On the upper left of the interface, the user may choose to use the dropdown menu to filter down to a particular Broad Category.



When the user hovers over the point for a particular business, the border of the business's point become bold and the Business Name appears in a small box tangent to the point. Instead of filtering on Price Range, we include the Price Range indicator ($-$$$$) along with the Business Name. Rather than filtering on Location, we include only businesses in Nevada.



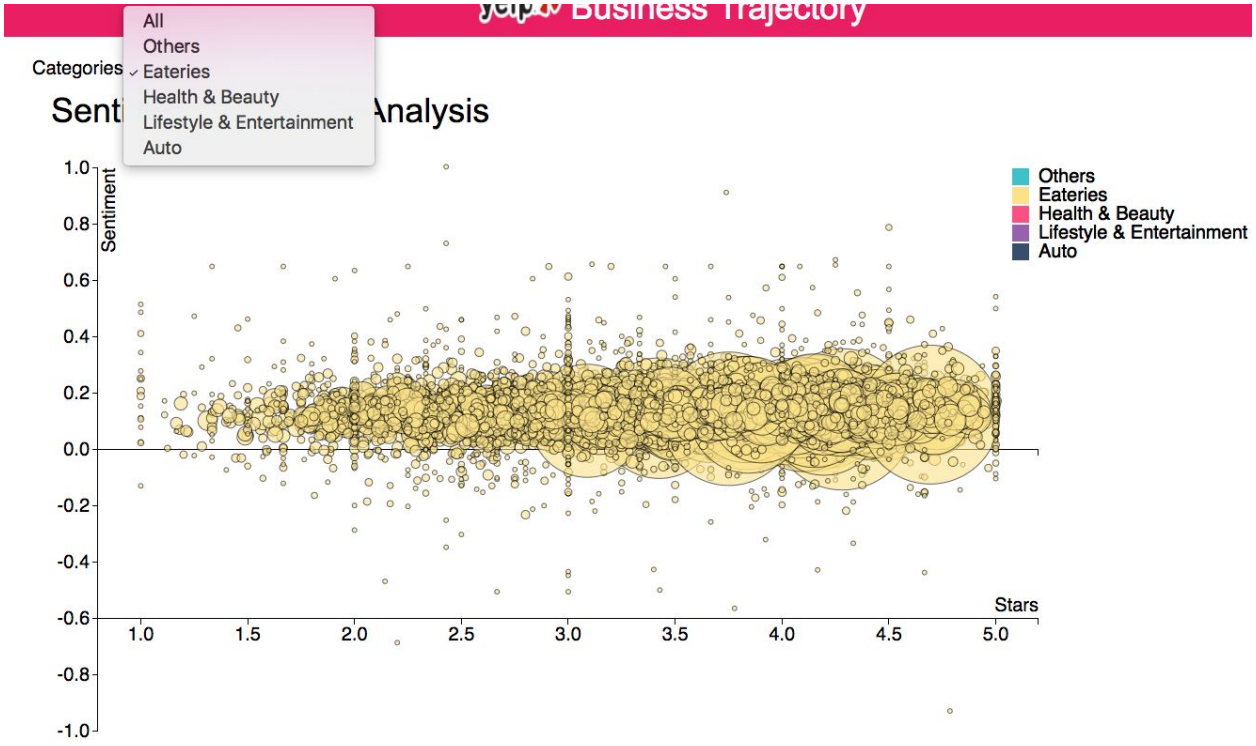
When the user clicks on a business's point, the words in that business's Tip Text string are searched for in the word sentiment look-up data frame. A word cloud is then displayed below the scatterplot. The title over the word cloud is the chosen business's name. Composing the word cloud are the set of words in the Tip Text string. Each word's font is scaled by that word's Inverse Log Word Frequency, and it colored with a diverging scale encoding Word Sentiment. A green (Sentiment = 1) to pink (Sentiment = -1) color range was chosen to be aesthetically pleasing without repeating any colors used to encode Broad Category, and thus cause confusion. The opacity of the words was set to 0.6 so that overlapping words were legible. The placement of the words were randomly generated using the d3-cloud[3] library, and thus did not encode information. For businesses with too many tips to display, 100 words from were randomly selected.

[3] Jasondavies. "Jasondavies/d3-cloud." *GitHub*. N.p., 27 July 2016. Web.

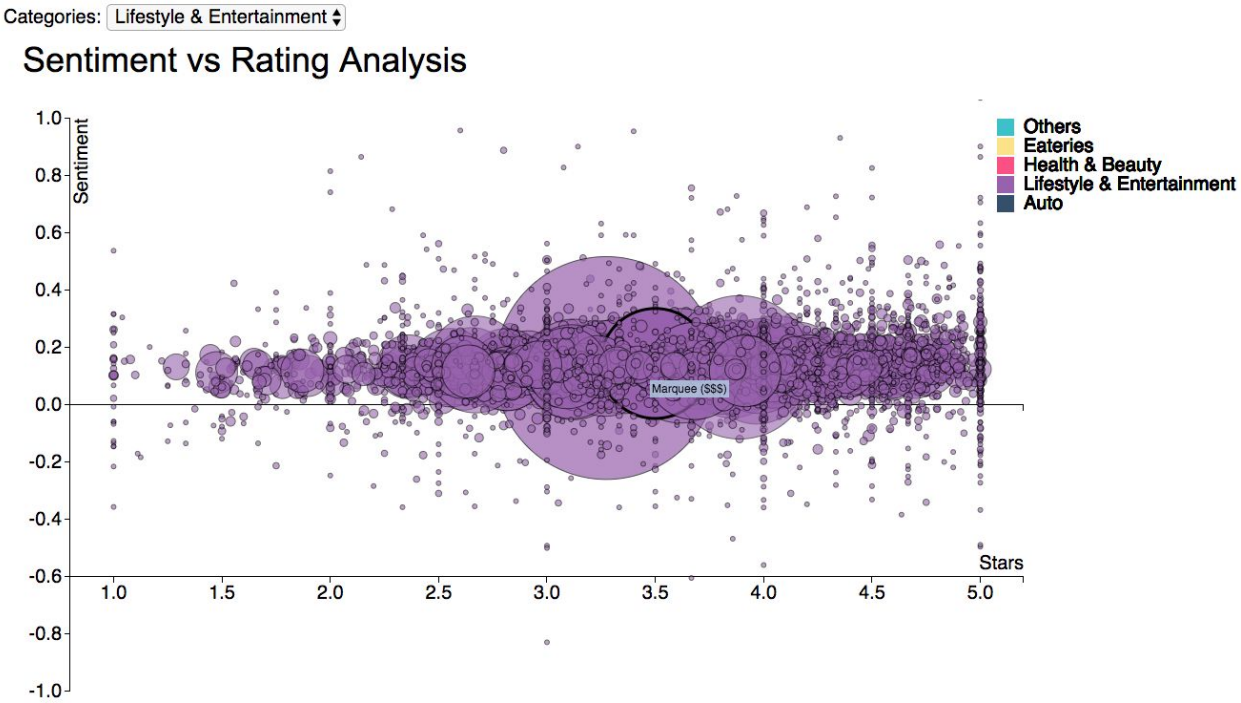
Findings:

Our Yelp Tip visualization did help use gain insight into the influence that tip count and sentiment has on Yelp business rating. While we did not find strong evidence against the null hypothesis (tips do not influence business rating), we did see several trends that encourage further investigation.

First, we found that Eateries behaved differently from the other Broad Categories. Eateries is visibly the largest of our five categories. Also, within Eateries, the larger the tip count for a business, the higher the probability that that business has a high rating. No Eatery with many tips has a high rating. Additionally, Eateries appeared to have a slightly higher average sentiment then the businesses in the other Broad Categories.



In the other Broad Categories, we witnessed that the outlier businesses, ones with a disproportionate number of tips, had moderate ratings. This is likely due in part to the averaging of many ratings, bringing the score towards the median rating value.



We found that the relationship between rating and sentiment was, for the most part, flat. However, we did find that the Average Sentiment for most businesses were positive. While the effect was small, this was in line with an underlying hypothesis of ours -- that tips would generally be positive.

Finally, we found that there were a number of rare words with strongly positive or negative sentiment appearing in some businesses' word clouds. We suspect that this easy to use tip viewing tool would be of interest to businesses to learn words that are rare and positive, and attempt to associate with their business, and words that are rare and negative to disassociate from.

Future Work:

There are several immediate improvements that could be made to our visualization. First, when a business point is selected with a mouse click, we would like the point to remain highlighted and labeled until unclicked or another business is selected. This would help the user

remember where in the scatterplot the business fell, and how many tips the business has. Also, in our Business Name - Price Range label, we would like the add the Number of Tips Per Business so that value could be read explicitly, rather than just encoded with the point size. Furthermore, we would like to add the business names listed alphabetically to the right of the scatterplot so that a business can be chosen and highlighted by name, rather than by location in the scatterplot.

One improvement to consider, but would require testing, was the suggestion to remove the light gray border around each business. On one hand, this would emphasize the distribution of businesses along Average Rating and Average Sentiment. However, this would make determining the border of each business point difficult. There are three alternative solutions. First, fewer businesses could be displayed so that there is little to no overlap. In our Nevada dataset, there are over 23,000 unique businesses. To cut down on this number, we could look at a smaller metropolitan area or a finer scale set of Categories. Second, our analysis could account for the skew in the distribution of ratings. Namely, businesses are most likely to the have a 4 star rating. We could have taken a random subsection of businesses with a constant distribution across ratings to minimize the amount of overlap in the 4 star range. Finally, we could have excluded business that had only one tip, or some other low threshold. Not only would this exclude many businesses; it would lessen the concentration of businesses with integer star ratings, and minimize wild positive or negative average sentiment from an individual strong sentiment tip.

To improve the layout of the word cloud, a more intelligent word placement generator could be used to ensure that no two words overlap at all. If that were the case, either the opacity could be set to 1, to make each word more salient, or the opacity could be used to encode another attribute such as word frequency.

While in many cases the sentiment displayed with each word in the word cloud followed the user's intuition, there were instances where the sentiment score for a word was surprising. This is largely due to the simplicity of the sentiment analysis chosen. A future goal is to apply a "smarter" sentiment analysis to the data in this visualization to gain richer insight from sentiment influence and improve word cloud impact.

Businesses with too many words to display in the word cloud were limited to a random set of 100 words. To improve upon this, either the dimensions of the word cloud could be increased to make space for all words, or a more interesting criteria, such as word frequency, could be used to determine which 100 words will be displayed.

One final feature that we would consider adding would be to allow selection of words in the word cloud to illuminate the context that a particular word is used. For example, when a word is selected, we could display the tip that word appears in for that business, along with the other tips that contain that word. This would help to user determine if a word is used similarly across tips, or a particular business inspires tip authors to use a word of interest in a unordinary way.