Gaining Insight into Yelp Data

Nehal Kamat[*] University of Colorado Boulder
Aadish Gupta[†] University of Colorado Boulder
Sachin Muralidhara[‡] University of Colorado Boulder
Keerthi Pai[§] University of Colorado Boulder
Naif Alharthi[¶] University of Colorado Boulder

ABSTRACT

The Yelp Dataset Challenge is an academic challenge for students and teachers alike to provide valuable insights into the trends, cuisines and social profiles of various businesses across the world. The dataset contains information for 11,537 businesses and 8,282 check-in sets, 43,873 users, 229,907 reviews for these businesses. Our task was to draw up 5 different high-level visualizations that could provide viewers a sneak-peek into the social circles of business reviewers, growth in popularity of businesses as well as pin-point the most famous fast-food joints across the US and Europe. We have developed a dashboard to represent the data visually using popular JavaScript charting and visualization libraries, namely D3.js and Leaflet.js.

1 INTRODUCTION

Finding local restaurants, businesses and retail outlets using online services has become something of a norm in the past few years, and Yelp is a leading provider of local business information to its users. With the goldmine of local business information that Yelp houses, there are multiple questions that can be asked and answered regarding this data, and we approach a few of these questions with visual representation of Yelp's enormous dataset. Yelp's dataset consists of a large number of businesses spread across the USA and Europe, as well as 8,282 check-in sets, 43,873 users, 229,907 reviews for these businesses. We attempt to develop a dashboard that houses 5 different visualizations, each offering a high-level view of different segments of the dataset and, in the process, telling a unique story to the viewer about the data as well as cultural trends across the country.

2 LITERATURE REVIEW

This is the ninth round of Yelp's data challenge. Previous challenges have resulted in some novel research work on social networks and analysis of data. [Raju, Basnage, and Yin] talk about how visualization of business reviews can aid business owners in understanding people's perception and compare their business to other similar businesses. [Hajas, Gutierrez, and Krishnamoorthy, 2014] model the trend and behavior of reviews and ratings around the college towns for last seven years. They conclude with the observation that cumulative ratings of the restaurants converge while the user reviews fluctuate. The authors make some interesting use of visualizations to put their point forward, especially heat maps. [Huang, Rogers, and Joo, 2014] talk about how topic modeling on user reviews can help in getting insights on user needs and in turn enabling restaurants to better serve their customers and grow economically.

[Chen, Chen, Chen, and Joshi, 2016] describe how social media and images uploaded on them could be used for marketing and recommendation of local business. Authors propose detecting finer grained details about business concept from the images using Convolution Neural Networks. [Heer and Boyd, 2005] propose a new system of visualization to explore and navigate large-scale online social networks.

[Van Wijk and Van Selow, 1999] put forward a novel approach of visualizing time series data and discovering patterns and trends on different scales like Day, Week and Month in parallel. They propose a cluster and calendar based visualization. [Gove, Gramsky, Kirby, Sefer, Sopan, Dunne, Shneiderman, and Taieb-Maimon, 2011] talk about how a heat map and matrix can be used to visualize social network data over time and help users in exploring temporal trends in the network. [Fisher, 2007] propose the idea of imposing user information over the map to better interpret user's behavior.

3 BUILDING THE DASHBOARD

The primary purpose of a dashboard is to serve as a web page which collates information about a business. The Yelp dataset contains a plethora of information about various businesses across the country, and so it was a natural choice for us to develop a dashboard to display the business information that we intended to visualize.

4 VISUALIZATIONS

4.1 User Connections

In our first visualization, we look at the social network [Heer and Boyd, 2005] between the top reviewers of the Yelp DataSet. The motivation behind this visualization was to find patterns between users based on their review count and average ratings. Our first and intuitive attempt was to visualize this dataset as a network graph. But given the large number of nodes we had, this turned out be a futile approach, and the visualization looked like a hairball and was less visually appealing.

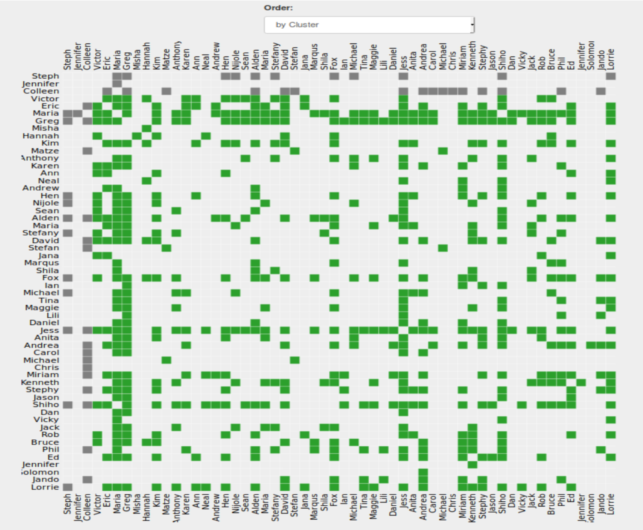
For our second attempt, we analyzed the user data as a matrix [Gove et al., 2011] with as many rows as there are users in our dataset. The color of each cell (i,j) of the matrix indicates if the users are friends. We can explore the visualization and reorder the matrix based on review counts, average user rating, and alphabetically. The visualization is built using D3 and inspired by by Mike Bostock's visualization of character co-occurrence in the different chapters of Les Miserables.

4.2 Tracking Number of Check-Ins

Our next visualization is a spatial view of ”check-in” culture across the US and Europe for each of the 7 days of the week. The question that the visualization tries to answer is ”What restaurants have the most guests at any given day of the week?”. To begin with, we plotted the check-ins for all types of businesses on the map [Fisher, 2007]. But alas, the data was skewed: there were far more check-ins for airports than for other types of businesses. The map displayed concentrated dark-purple circles at every airport in the dataset. this led us to filter the data to get only those entries that were of type

[*] e-mail: nehal.kamat@colorado.edu
[†] e-mail:aadish.gupta@colorado.edu
[‡] e-mail:sachin.muralidhara@colorado.edu
[§] e-mail:keerthi.chikalbettupai@colorado.edu
[¶] e-mail:naif.alharthi@colorado.edu

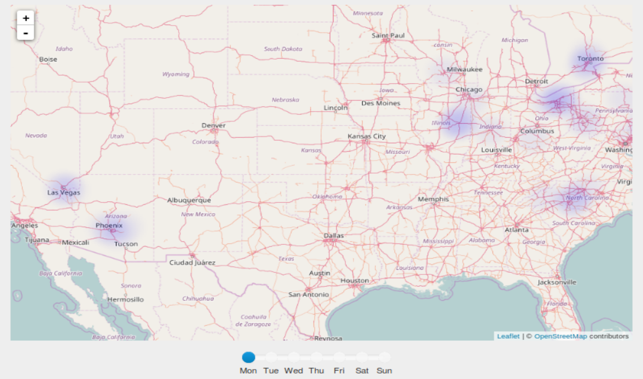
Figure 1: Visualization Depicting User Network



”Restaurant”. Dealing with only one type of data would do away with the issue of data imbalance and would still be effective in getting the point across.

The visualization works by showing a spatial heatmap of the number of user check-ins across 2009-2017 aggregated on the 7 days of the week. A linked bar at the bottom allows users to select any day of the week, which then changes the heatmap view to display the heatmap for that particular day. This particular way of visualizing check-ins allows the users to see the trend of check-ins across different regions of the USA and Europe as the the week progresses. Our findings follow our intuitive understanding of social behaviour: most check-ins occur on Fridays, Saturdays and Sundays (weekends). One can also notice how the heatmap begins to spreads after Wednesday, probably telling us that the most people are eagerly anticipating the end of the work-week.

Figure 2: Visualization Depicting Heat-Map for Check-ins



4.3 Most Popular Word Clusters

The following visualization has been created by fitting a 50-topic model using the gensim [Rehurek and Sojka, 2010] implementation of LDA [Blei, Ng, and Jordan, 2003] and the pyLDAvis [Sievert and Shirley, 2014] library. [Huang et al., 2014] describe a method to improve a restaurant’s revenue by understanding the products the customers care about. In their method, they make use of all the reviews provided in the dataset. However, our criteria to select the reviews is two fold:

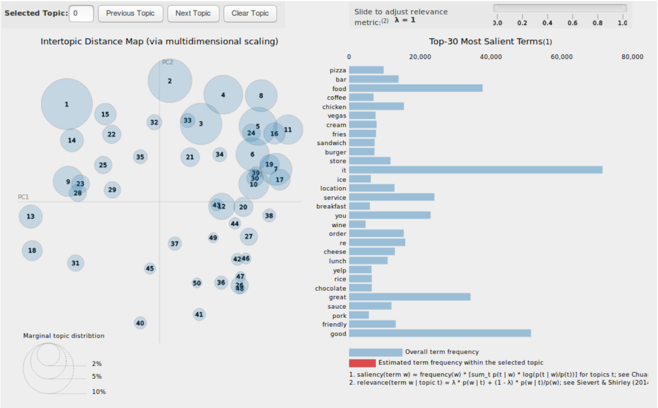
- Select the id’s of users who have contributed to at least one five star review
- From this pool of users who have contributed to at least one five star review, determine the users who have written at least 100 reviews and gather all their reviews

We also filter out tokens that appear in less than 20 reviews and more than fifty percent of the reviews. Our 50-topic model is fit to a corpus containing 73737 documents and 14078 tokens.

The two facets to the visualization are the prevalence of the topics across the documents and the frequency of words across the topic and the documents. The prevalence of the topic is represented by the size of the circle and the frequency of words within the topic is represented by the horizontal bar chart. The parameter lambda is used to understand the ratio of the word frequency given the topic to the overall frequency of the word.

Hovering over a topic gives us the most relevant terms and hovering over a word gives us the frequency of a word over the topics. It helps us understand how the words are used in a variety of contexts. The distance between the topics determines the semantic relationship between topics. In our visualization, we can see that topics 1 and 14 are semantically related. Topic 1 is comprised of dishes that fall under the Mediterranean and Mexican cuisine whereas topic 14 can be loosely labeled as finger food.

Figure 3: Visualization Depicting Common Topics



4.4 Ratings Trend

The fourth visualization on the dashboard provides a time-series view of the average ratings of the top-10 most reviewed businesses starting from 2007. The plot shows us how only a couple of these businesses have been popular right from the time Yelp was started, whereas a few other either started late or gained user-traction only a few years later. What’s more interesting to note is that these businesses average a 3.5 rating as of today when taken together, even if a few of them started off with pretty high average ratings. To understand how these ratings average out, we added another layer of analysis to the mix. Clicking on any of the trend lines (for any business) displays a calendar view with each cell representing a given day during the year and having a different shade of blue if there were ratings provided on that day. The brightness of the blue colour depends on the average of the total ratings provided on the given day, so the darker the blue, the better the average rating on that day.

Figure 4: Visualization Depicting Business Trend Over Years

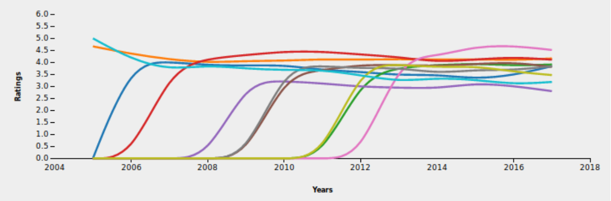
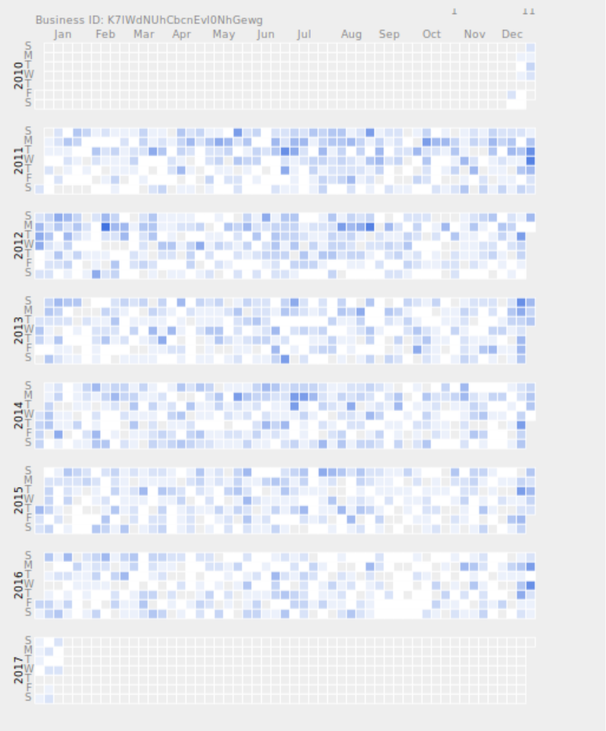


Figure 5: Visualization For a Particular Business Across Years

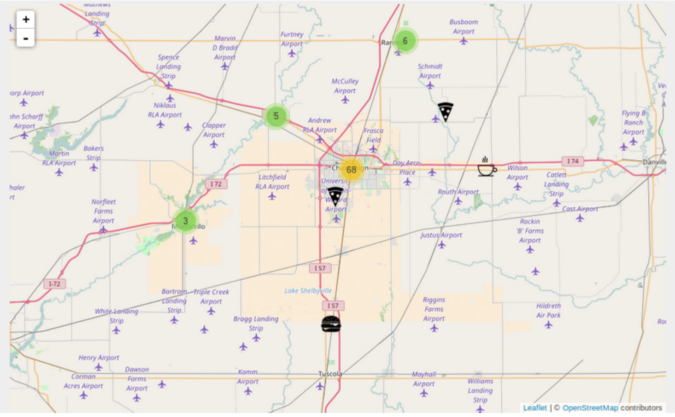


4.5 Plotting Fast-Food Joints

The focus on this visualization is to determine the most popular fast-food joints from the dataset based on a business' rating. The given dataset comprises of different categories of businesses. We select those businesses that are categorized as resturants and have at least a rating of four. We further divide them into three sub-categories namely pizza, burger and coffee. Each sub-category is represented by a specific marker.

This visualization can help new users identify areas which have a high concentration of good fast-food joints and also zero-in on the joint itself. We found that for cities such as Tempe and Champaign, some of the popular food-joints are located near the University.

Figure 6: Visualizing Fast Food Joints



5 Conclusion

Our dashboard comprises of various high-level views of different segments of the data. We've tried to diversify our approach in providing insights into the data by using different analysis techniques such as correlation matrix, line graphs, spatial plots and calendar views. Our bench-marking involved showing the dashboard to a few colleagues and friends, who suggested some minor tweaks to be incorporated into the dashboard, one of them being the calendar view. Given more time, more precise and distributed bench-marking techniques can be employed to validate the effectiveness of the visualizations on the dashboard. There's more scope for further analysis in the way the visualizations can be modified or scaled to encompass more data that Yelp might provide in the future, or even gaining deeper or more diverse insights into culinary and social trends associated with user reviews.

References

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Bor-Chun Chen, Yan-Ying Chen, Francine Chen, and Dhiraj Joshi. Business-aware visual concept discovery from social media for multimodal business venue recognition. In *AAAI*, pages 101–107, 2016.

Danyel Fisher. Hotmap: Looking at geographic attention. *IEEE transactions on visualization and computer graphics*, 13(6):1184–1191, 2007.

Robert Gove, Nick Gramsky, Rose Kirby, Emre Sefer, Awalin Sopan, Cody Dunne, Ben Shneiderman, and Meirav Taieb-Maimon. Netvisia: Heat map & matrix visualization of dynamic social network statistics & content. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on*, pages 19–26. IEEE, 2011.

Peter Hajas, Louis Gutierrez, and Mukkai S Krishnamoorthy. Analysis of yelp reviews. *arXiv preprint arXiv:1407.1443*, 2014.

Jeffrey Heer and Danah Boyd. Vizster: Visualizing online social networks. In *Information Visualization, 2005. INFOVIS 2005. IEEE Symposium on*, pages 32–39. IEEE, 2005.

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

Akhila Raju, Cecile Basnage, and Jimmy Yin. Visualizing yelp ratings: Interactive analysis and comparison of businesses.

Radim Rehurek and Petr Sojka. Software framework for topic modelling with large corpora. In *In Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. Citeseer, 2010.

Carson Sievert and Kenneth E Shirley. Ldavis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces*, pages 63–70, 2014.

Jarke J Van Wijk and Edward R Van Selow. Cluster and calendar based visualization of time series data. In *Information Visualization, 1999.(Info Vis' 99) Proceedings. 1999 IEEE Symposium on*, pages 4–9. IEEE, 1999.