Yelp Challenge - Final Presentation

Team:
Narain Yegneswara Sharma,
Vaishnavi Viswanathan and
Pravin Venkatesh Venkataraman

Course: Data Mining (CSCI 5502)
Section: In class

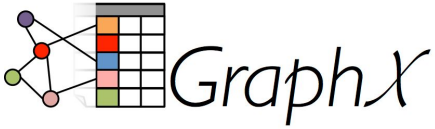
Motivation

- Yelp has amassed massive amounts of information through years of being a well visited website for reviews about businesses.
- Yelp hosts an annual challenge at https://www.yelp.com/dataset_challenge to let their competitors prove hypothesises.
- Dataset available for this challenge is a composition of exhaustive information about
    - 142 million users,
    - 100 million reviews and
    - 2.1 claimed businesses.
- Knowledge and insights into users and business indexed under Yelp.
- Advent of non traditional ways to visualize data.

Literature Survey

- Levi, Asher, and Osnat Mokryn. "The Social Aspect of Voting for Useful Reviews." Social Computing, Behavioral-Cultural Modeling and Prediction. *Springer International Publishing*, 2014. 293-300.
- Rahman, Mahmudur, et al. "To catch a fake: Curbing deceptive Yelp ratings and venues." Statistical Analysis and Data Mining: *The ASA Data Science Journal* 8.3 (2015): 147-161.
- Hu, Longke, Aixin Sun, and Yong Liu. "Your neighbors affect your ratings: on geographical neighborhood influence to rating prediction." *Proceedings of the 37th international ACM SIGIR conference on Research & development in information retrieval*. ACM, 2014.
- Bakhshi, Saeideh, Partha Kanuparthy, and David A. Shamma. "If it is funny, it is mean: Understanding social perceptions of yelp online reviews." *Proceedings of the 18th International Conference on Supporting Group Work*. *ACM*, 2014.

Tools



Google Cloud Platform

Google Data Proc

Our Work

- Tested the following hypotheses againsts Yelp dataset.
  - Natural Language Processing
    - Performed correlation analysis between reviews and the ratings given by the user.
  - Cultural Trends
    - Found the states that has similar trends by applying normalization and there by computing euclidean distances between them.
  - Social Graph Analysis
    - Found the most influential users in Yelp.
    - Built a generic multi dimensional data driven tool or dashboard to study and get insights from Yelp’s user network.

Natural Language Processing

- Reviews play a major role in determining a business's success and in prompting new users to go check out a new place.
- Figuring out the Correlation between the number of stars and the review sentiment.
- Initial Step:
  - Generated word clouds using bigram and trigram frequencies.
  - Results were encouraging.
  - 1 Star reviews had complaints about the customer service whereas 5 Star reviews had compliments about the customer service.

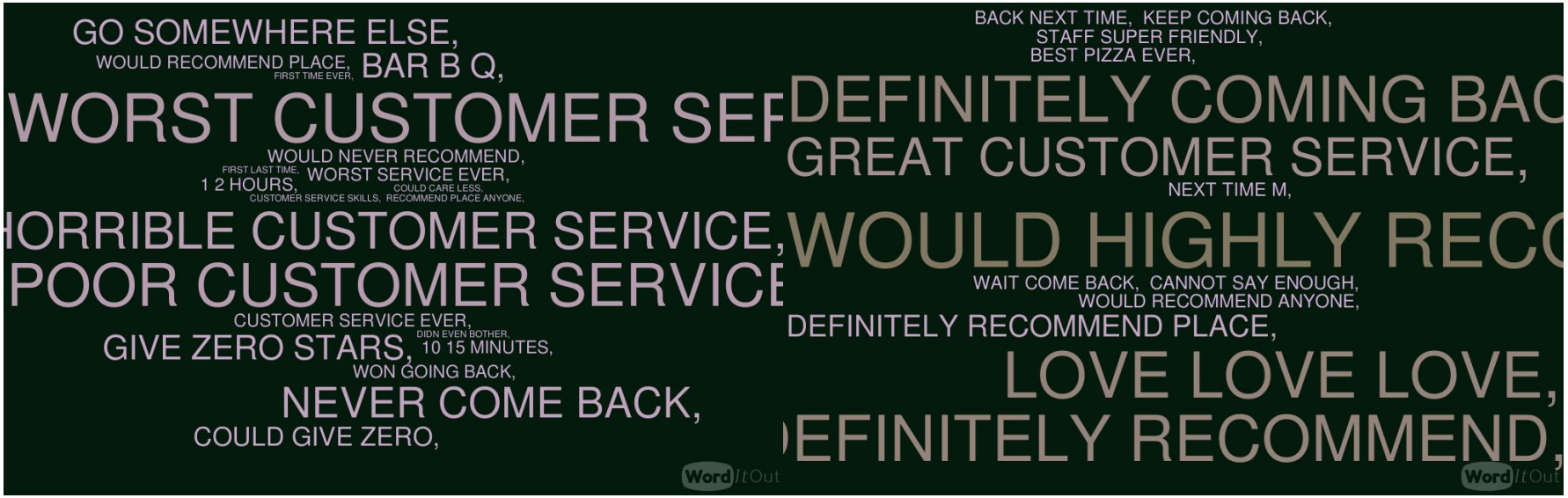
Regression Analysis

- Reviews were broken up into unigrams and then compared against a lexicon of positive and negative words.
- Linear regression of the number of stars on each of the attribute ie number of positive words, number of negative words, total number of words in a review was performed.
- The rating of a review increases by 0.14 for every positive word in the review and decreases by -0.32 for every negative word in the review.
- A little surprising result was that the rating of a review decreases by 0.17 for every extra word in the review. A reason for this could be because users on an average tend to write lengthier reviews complaining, whereas the positive reviews tend to be shorter and to the point.
- The regression models explains upto 10% of the variation in the ratings of the stars.

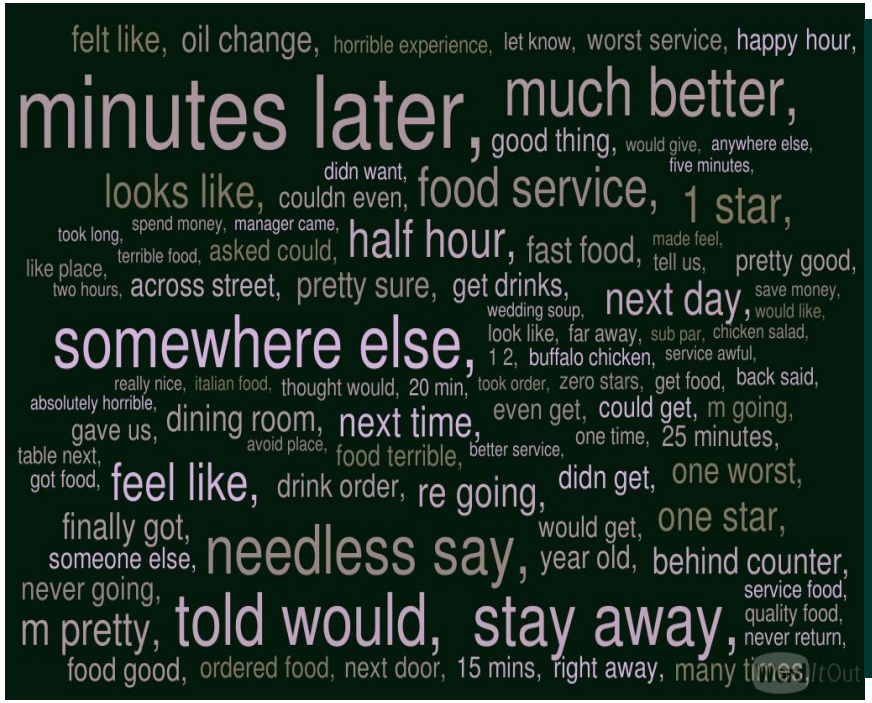
Classification Problem

- The regression analysis on the reviews was converted into a classification problem.
- 1, 2 and 3 star rated reviews were considered as negative and 4, 5 star rated reviews were considered as positive.
- Using Support Vector Machine to classify the review using the Scikit-learn library yielded an accuracy of 70% on the test data.
- If 3 star reviews were considered to be neutral, then an accuracy of 67% was achieved by counting the number of positive and negative lexicons and determining the sentiment of the review.

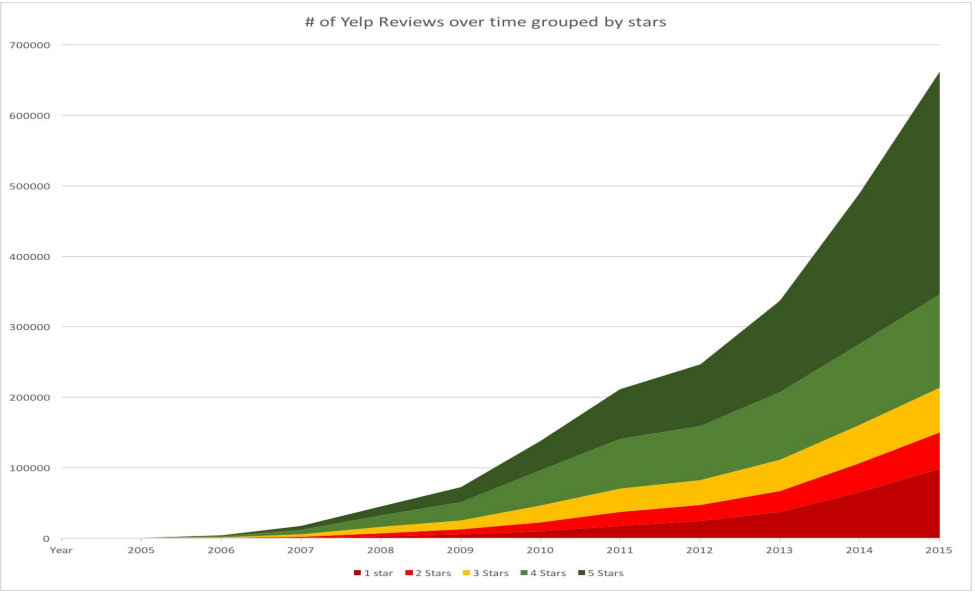
Word clouds - Trigrams

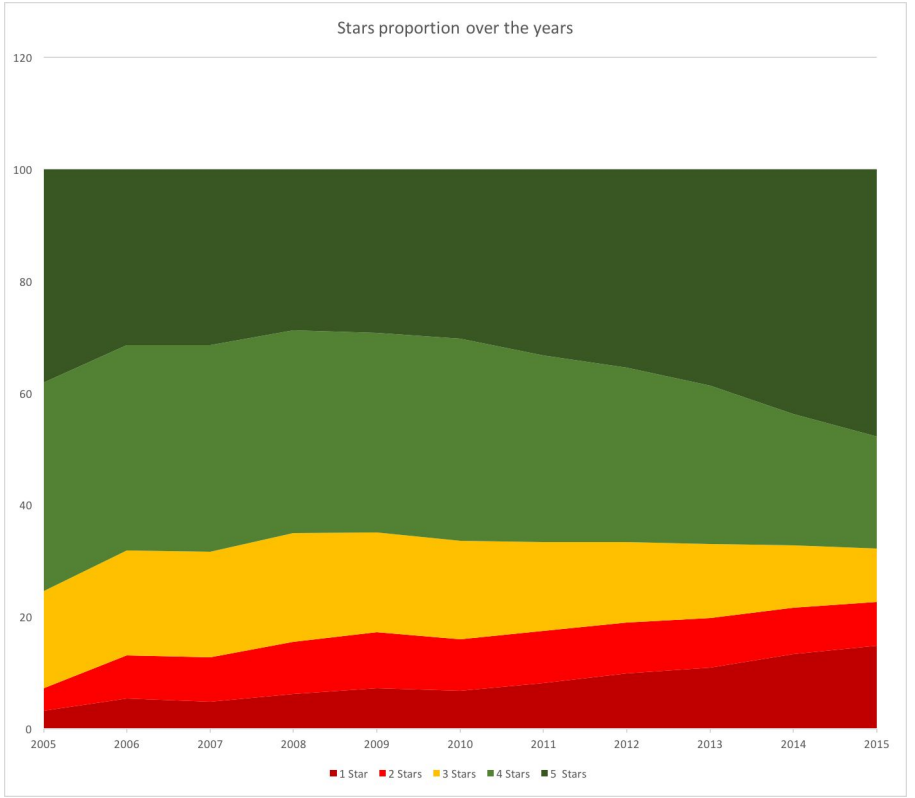


Word clouds - Bigrams

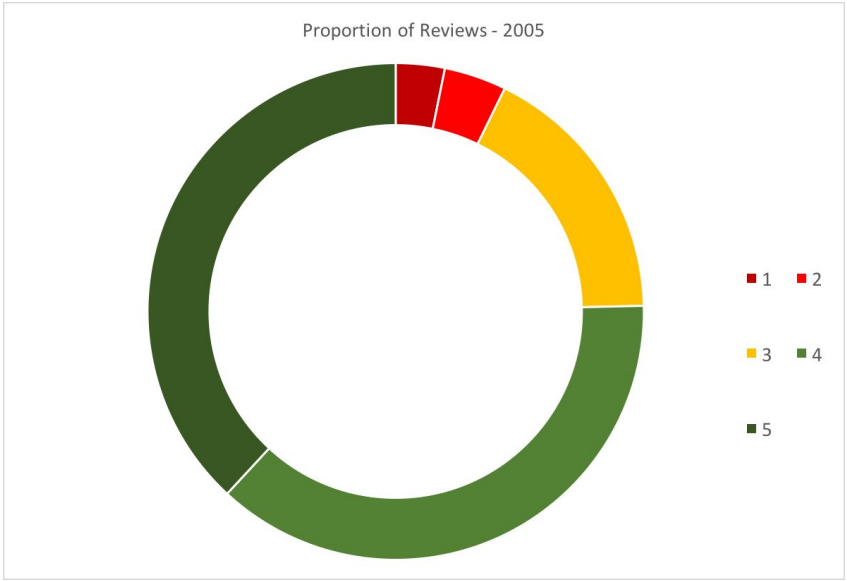


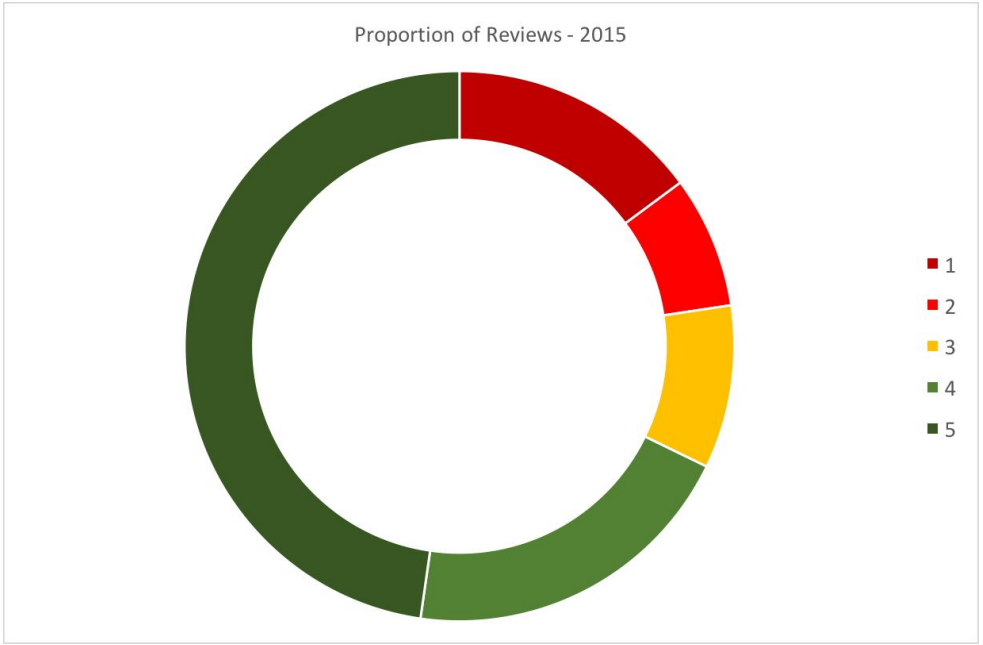
Yelp Ratings over the years.



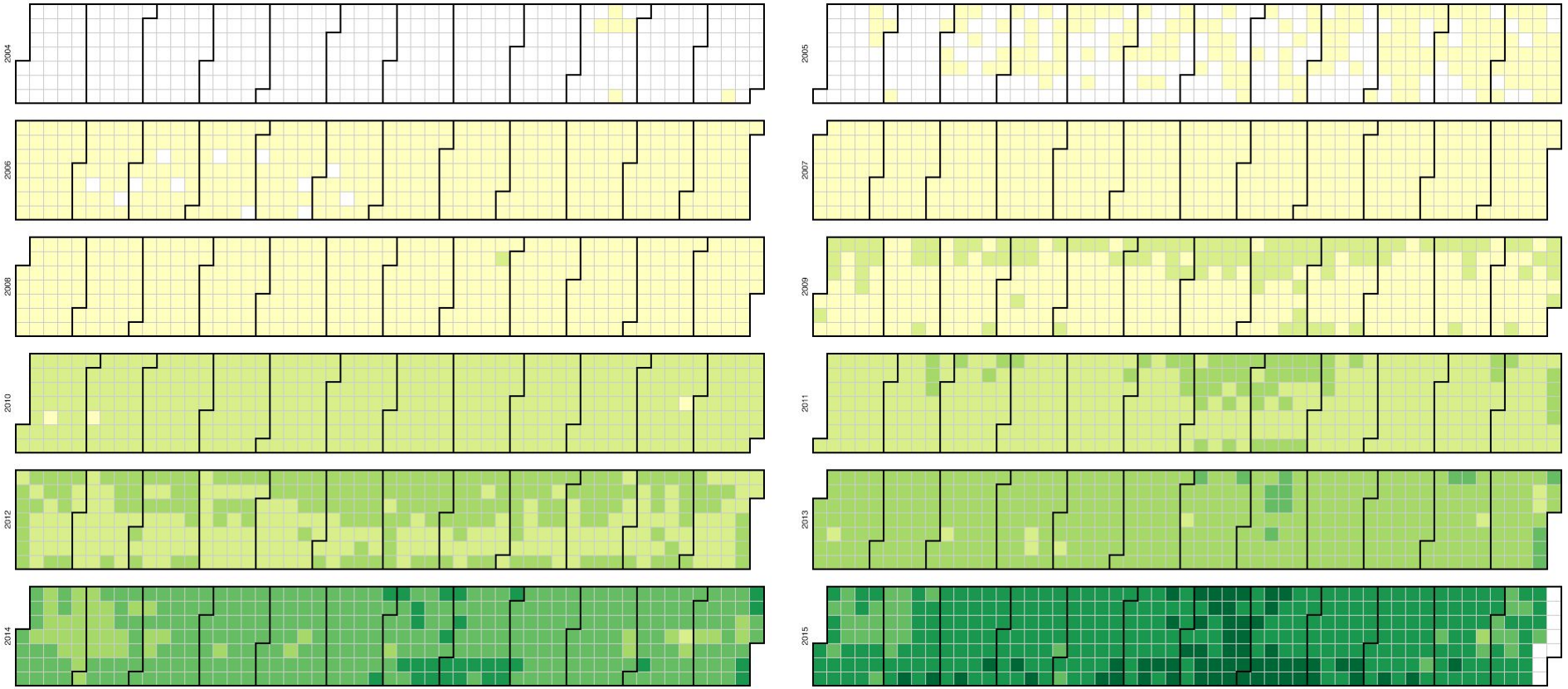


Yelp - Proportion of Reviews 2005 vs 2015





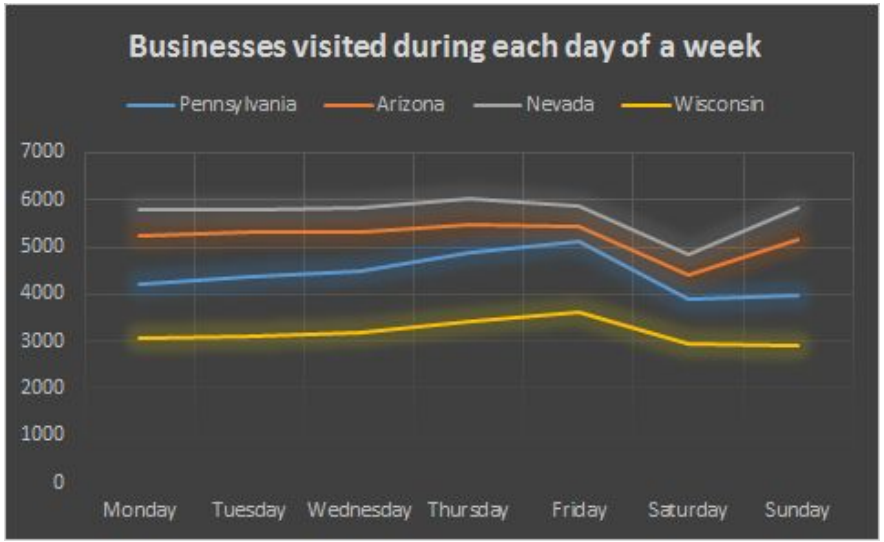
Calendar Heat Map



Trends Analysis

- Finding if people in the states considered - Pennsylvania, Nevada, Arizona and Wisconsin are going out more during the working days or in the weekends?
- Finding out if there are any similarities between the states considered?
- Procedure:
  - Data for each state was collected and analysed i.e the total number of businesses visited in each day of the week with time slots (Morning -6:00 AM-11:59 AM,Afternoon - 12:01 PM-6:00 PM and Evening-6:00 PM-12:00 PM) was observed.
  - The results were matching with the predictions (People tend to go out more during the weekdays than weekends).

Trends Analysis



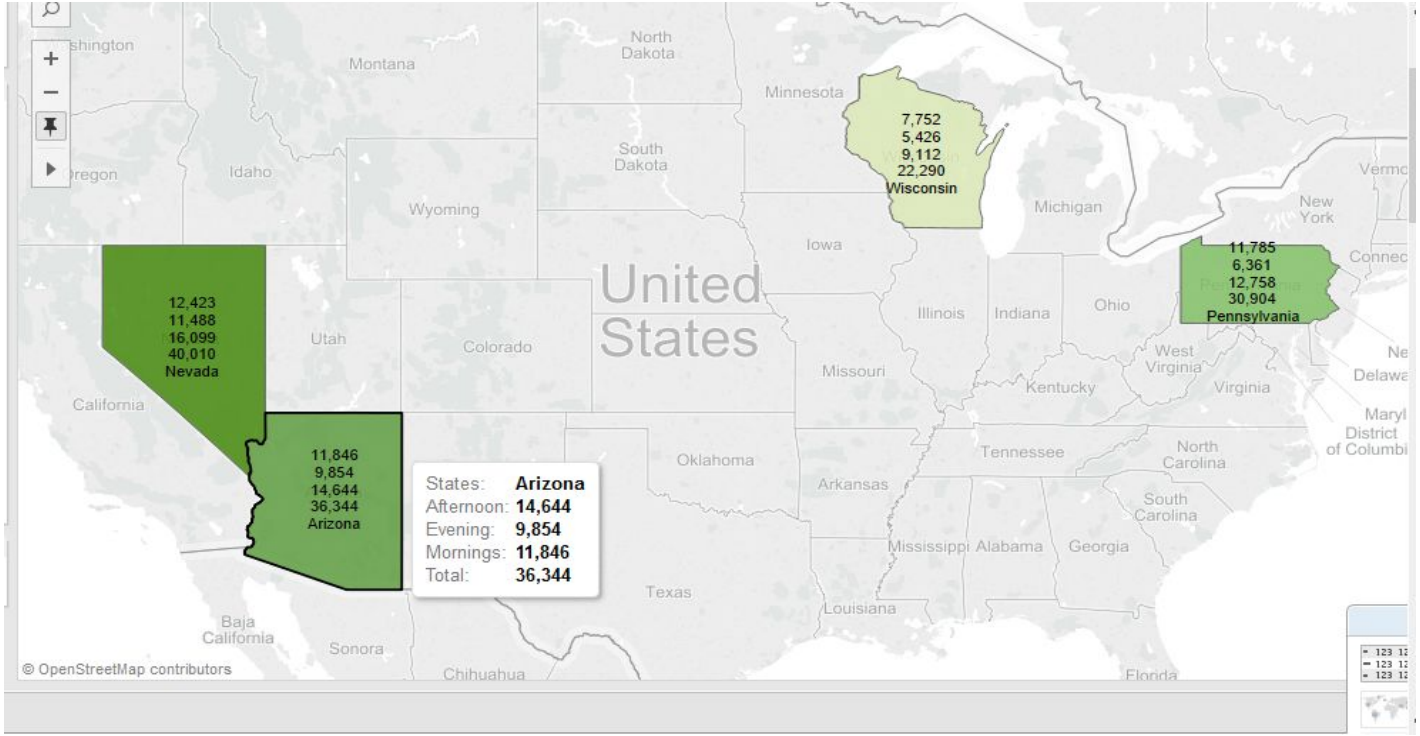
- Here, during the weekdays, there is a steady growth in all the four states and it drastically drops during the weekends.

Trends Analysis

In all the states considered there is always a gradual increase in the businesses visited during the weekdays whereas,

- If we observe the line graph closely, we can see that states Nevada and Arizona are similar i.e During the weekends alone, there is a drastic drop on Saturday and sudden increase on Sunday.
- Also, states Pennsylvania and Wisconsin appears to be similar i.e. During the weekends there is a gradual decrease in the businesses visited.
- So, to confirm this similarity Normalization technique was used.
- The results of the normalization showed that Nevada-Arizona were similar. Whereas our prediction of Pennsylvania-Wisconsin being similar did not match.

Trends Analysis



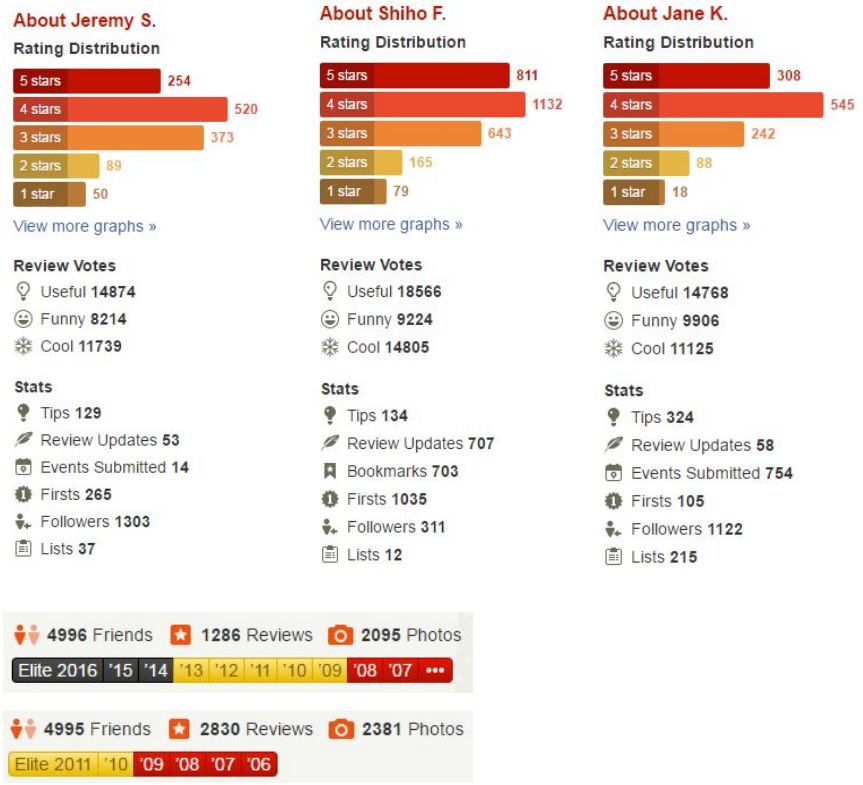
Graph Analysis: Influential Users

- Form a graph from Yelp's user dataset.
- Graph representations,
  - Directed Graph
  - Node - User
  - Edges - Followership
- Ranking algorithm used: Page Rank
- Here the algorithm is implemented so that each user is ranked based number of in-degrees and quality of those in-degrees.
- Infrastructure: Graphx on Spark deployed in a cluster built with the help of Google Data Proc.

user_id	name	since
nkN_do3fJ9xekchVC-v68A	Jeremy	2005
2l0O1EI1m0yWjFo2zSt71w	Shiho	2006
uguXfIEpI65jSCH5MgUDgA	Jane	2005
7uxXuCcpw9-mUS3OJVw8aQ	Jessica	2005
1kpMAKRZuAz3OzxBav3XTg	Ligaya	2005
8J4IIYcqBlFch8T90N923A	Joan	2004
i63u3SdbrLsP4FxiSKP0Zw	Nish	2005
Rir-YRPPClKXDQbc3BsVw	Megan	2006
zTWH9b_ItSdLOK9ypeFOIw	Teri	2006

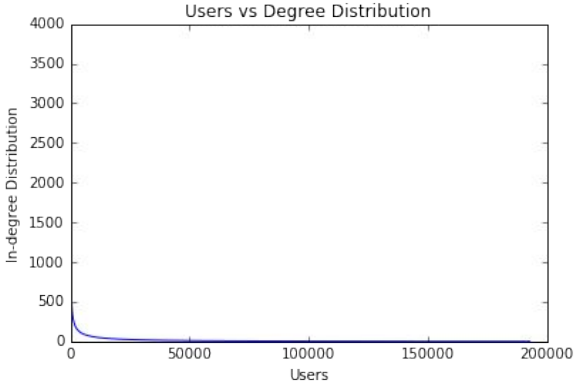
Graph Analysis: Influential Users

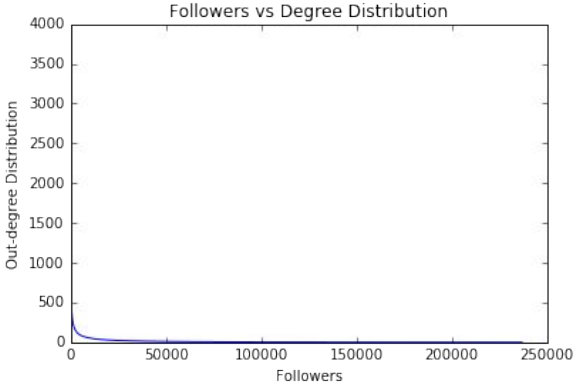
user_id	name	since
nkN_do3fJ9xekchVC-v68A	Jeremy	2005
2l0O1EI1m0yWjFo2zSt71w	Shiho	2006
uguXfIEpI65jSCH5MgUDgA	Jane	2005
7uxXuCcpw9-mUS3OJVw8aQ	Jessica	2005
1kpMAKRZuAz3OzxBav3XTg	Ligaya	2005
8J4IIYcqBlFch8T90N923A	Joan	2004
i63u3SdbrLsP4FxiSKP0Zw	Nish	2005
Rir-YRPPClKXDQbc3BsVw	Megan	2006
zTWH9b_ItSdLOK9ypeFOIw	Teri	2006



Graph Analysis: Degree Distribution

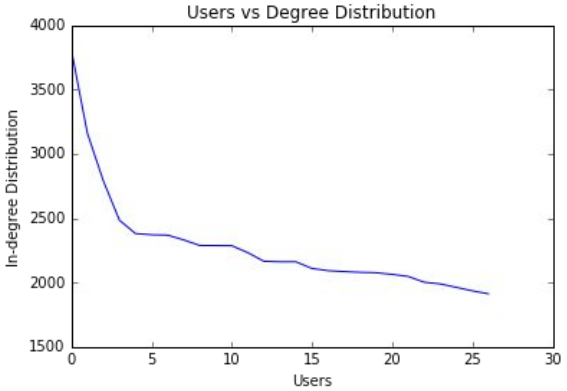
- Study of degree distribution i.e a plot of Users vs In-degree Distribution and Followers vs In-degree Distribution.
- The degree of distribution of observed strongly represented a typical social graph through precedence of a Power law distribution.
- Therefore, Yelp has the following,
  - Many users only with few followers.
  - Few users with many followers.
  - Graph is extremely sparse.
  - This graph is similar to Facebook, Twitter or any other social graph.

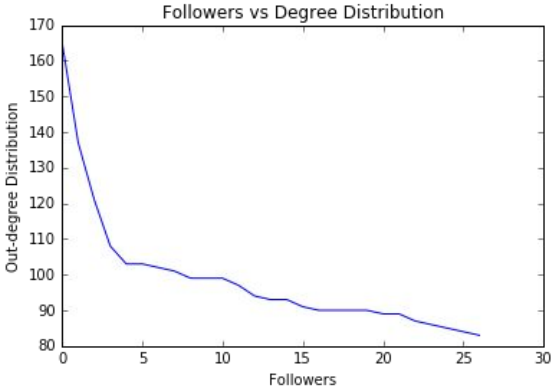




Graph Analysis: Build Multi-dimensional Dashboard

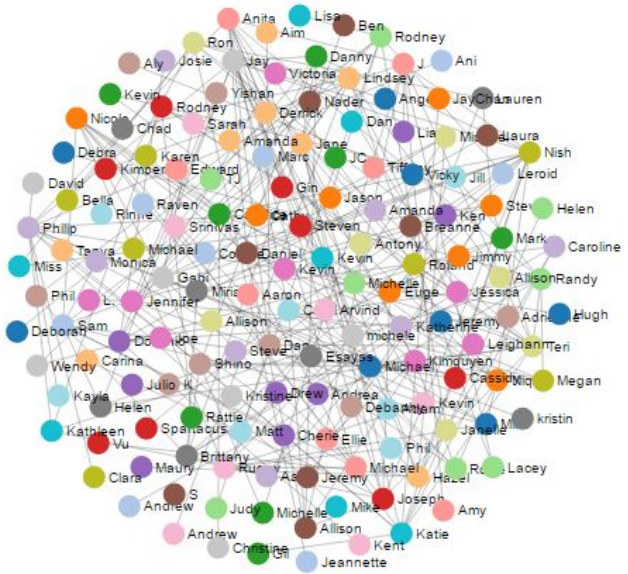
- Graph processing is relatively expensive.
- A dashboard that allows attributes to be correlated with each other inside the context of graph.
- Sampling, a necessity with the scale (3 million edges; Gephi could handle only up to 1 million) of data provided by Yelp.
- Sampling should retain the sparsity of the graph.
- The number of edges is ranged equally and the followers included in this range are top ranked among others.



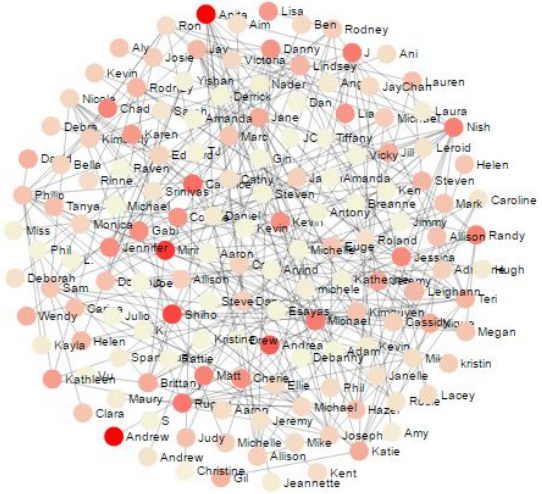


Graph Analysis: Build Multi-dimensional Dashboard

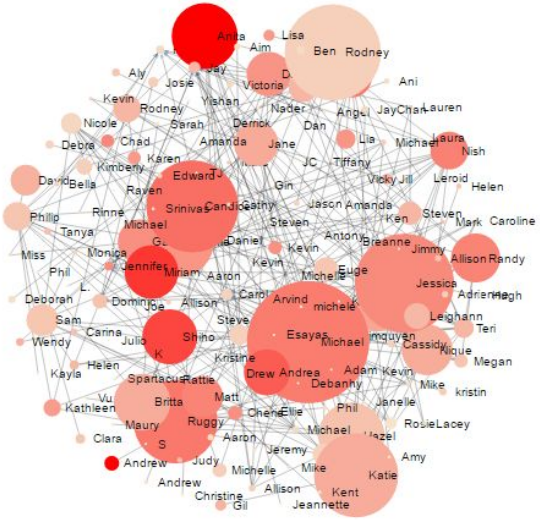
- Yelp user graph is reduced to manageable size with the same sparsity.
- Graph is built using force-directed layout (attractive and repulsive forces between nodes)
- Following ways are used to represent different attributes from either node or edge dataset,
  - Link strength,
  - Color of nodes,
  - Size of nodes and
  - Having multiple foci (foci being center of the graph).



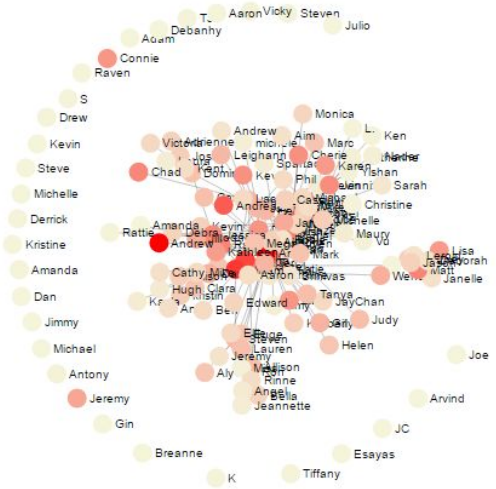
Graph Analysis: Build Multi-dimensional Dashboard



Color

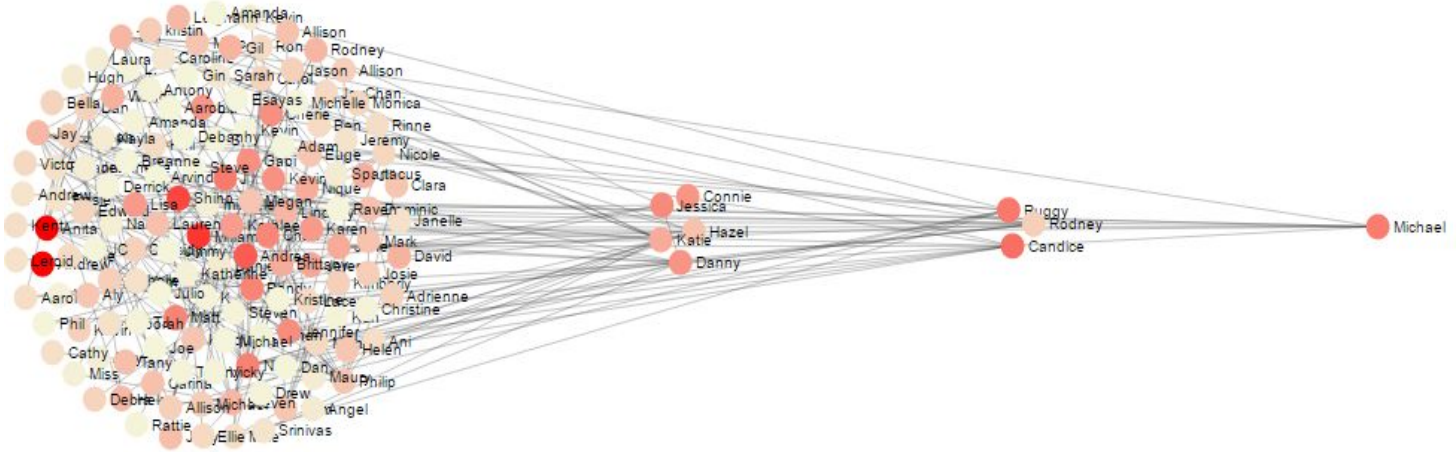


Size



Link strength

Graph Analysis: Build Multi-dimensional Dashboard

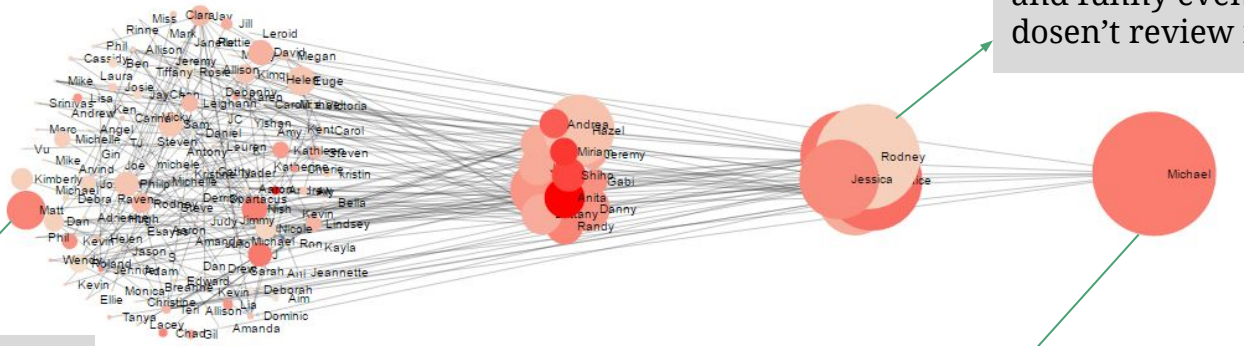


Multi Foci

Graph Analysis: Build Multi-dimensional Dashboard

Link to dashboard → v-pravin.github.io/yelpExplorer

Graph Analysis: Build Multi-dimensional Dashboard



Rodney's reviews are useful and funny even though he dosen't review much.

Matt's reviews a lot but his reviews are less useful and less funny.

Settings Selected:
Size: Funny
Color: Review Count
Foci: Useful

Michael actively reviews about businesses, his reviews are funniest and also most useful.

Evaluation

- Natural Language Processing
  - **Accuracy score:**
    - The Accuracy score achieved by the classifier was about 68% on an average, on a training/test set split of 80-20.
  - **Precision and Recall:**
    - The average precision is about 0.70 and recall about 0.67.
  - **R^2 Score:**
    - R^2 is the coefficient of determination, which determines the level of variance that can be predicted by the model.
    - The regression model constructed has an R^2 score of 0.20, which is not very much but compared to the huge dataset, it still is good enough to predict a lot of variations.

	precision	recall	f1-score	support
positive	0.53	0.71	0.61	123724
negative	0.80	0.65	0.71	217472
avg / total	0.70	0.67	0.67	341196

Evaluation

- Trends Analysis

Normalization - Min-Max Normalization

- Normalization of the total businesses visited in each state was done by using the total population of those states in year 2015.
- Euclidean distances was computed to find out the most similar records(states) using the normalized values.

	Euclidean distances			
	PA	WI	NV	AZ
PA	0	0.8599	1.1243	0.6764
WI	0.8599	0	1.0413	0.8002
NV	1.1243	1.0413	0	0.4478
AZ	0.6764	0.8002	0.4478	0

Evaluation

- Thus the most similar states from this given data is evident.
- 0.4478 shows that Nevada and Arizona are most similar.

Graph Analysis

- Degree distribution helped in sampling large graph and retain sparsity of graph.
- IQR was used to evaluate multi foci arrangement.
- NoSQL queries were used to evaluate the validity of the visualizations.

Conclusions

- Natural Language Processing
  - Correlation between Reviews and the stars given to the reviews.
  - Positive correlation between positive words and Stars, negative correlation between negative words and stars.
  - Review Length negatively correlated to stars. Unhappy customers give more lengthier reviews.
- Trend Analysis
  - People in all the 4 states considered go out more during the weekdays than during the weekends.
  - Trends in states Nevada and Arizona are more similar.

Future Work

- Detecting fake reviews in Yelp Reviews.
  - Crowd sourced reviews often have the issue of having fake reviews posted by people wanting to sabotage a business. Detecting those reviews could be of great help to local businesses as well as Yelp.
- Trend Analysis:
  - Finding reason behind a businesses being visited often? Is it because of these businesses being really good in what they do? Or whether when a user visits a business there is a good chance that it is in his/her neighborhood?
- Graph Analysis
  - Current analysis to be expanded to Business dataset.
  - Predicting business trends and finding the cause for events.

References

[1] “Community detection in graphs”, Physics Reports, Santo Fortunato, February 2010.

[2] “Why we twitter: understanding microblogging usage and communities”, Akshay Java, Xiaodan Song, Tim Finin, Baltimore and Belle Tseng, 9th WebKDD and 1st SNA-KDD, 2007.

[3] "If it is funny, it is mean: Understanding social perceptions of yelp online reviews.", Bakhshi, Saeideh, Partha Kanuparthy, and David A. Shamma,Proceedings of the 18th International Conference on Supporting Group Work. ACM, 2014

[4] “Network structure in matters of taste: can social networks predict cuisine taste?“, Matt Lamm, Imanol Arrieta Ibarra and Camelia Simoiu.

[5] "Finding local experts from Yelp dataset.", Jindal, Tanvi, University of Illinois at Urbana-Champaign, 2015

[6] “Your Neighbors Affect Your Ratings: On Geographical Neighborhood Influence to Rating Prediction”, Longke Hu,Aixin Sun,Yong Liu, School of Computer Engineering, Nanyang Technological University, Singapore,2014.

[7] “Finding Topic Trends in Digital Libraries”, Levent Bolelli, Seyda Ertekin, Ding Zhou, C. Lee Giles, Google Inc., The Pennsylvania State University University Park, Facebook Inc.,2009.

[8] http://minimaxir.com/2014/09/one-star-five-stars/

[9] Scikit-learn: Machine Learning in Python, Pedregosa *et al.*, JMLR 12, pp. 2825-2830, 2011

[10] http://worditout.com/word-cloud/

[11] “Finding Topic Trends in Digital Libraries”, Levent Bolelli, Seyda Ertekin, Ding Zhou, C. Lee Giles, Google Inc., The Pennsylvania State University University Park, Facebook Inc.,2009.

[12] http://aimotion.blogspot.com/2009/09/data-mining-in-practice.html

[13] http://kb.tableau.com/articles/knowledgebase/filled-map-custom-groups

[14] https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_population