Yelp Challenge 2016

Narain Yegneswara Sharma
CSCI – 5502

Pravin Venkatesh Venkataraman
CSCI - 5502

Vaishnavi Viswanathan
CSCI – 5502

ABSTRACT
The yelp dataset has been analyzed for a while by people interested in drawing out information on how well the local businesses in different cities of countries perform. This challenge has been open to all worldwide and has gone through 6 rounds of iteration now. The way people coming to a conclusion just with the number of stars and reviews ignoring other perspectives could not be an acceptable practice. Showing accurate results of how good a business by finding strong correlations between the ratings and the reviews, recommending the local businesses with respect to the changing trend by considering all the perspectives involved within the database is the main motivation behind this project.

Keywords
Natural Language Processing; Data Mining; Graph Analysis; Trend Analysis.

1. INTRODUCTION
Yelp is a multi-national organization which connects people with local businesses via crowd-sourced reviews and ratings. Yelp hosts an annual challenge where it provides dataset with information on users, businesses and reviews. It invites data analysts worldwide to leverage the data provided to mine insights and encourages to pursue varied tracks. This challenge has gone through 6 iterations and is currently in the seventh iteration. We have used the given dataset to get knowledge and insights through three different tracks namely, Natural Language Processing, Graph and Trend analysis. By using Natural Language Processing we were able to get interesting insights into reviews made by users and how it affects businesses. We found interesting trends followed in different states and also trend similarities and dissimilarities between these states. The given dataset is also tested in social network analysis, part of this work is built into a tool which is completely data driven. In overall, the results are published in form of tables, graphs and maps. Evaluations are done using standard statistical tool like Precision and Recall, Normalization and so on.

2. RELATED WORK

1.1 Reviews and Feedback:
Yelp is an essential online recommendation tool for customers as well as businesses. Originally, the success of a business got spread through word-of-mouth. With the advent of technology, online reviews affect the success of a business. Hence it is imperative to understand what drives users to post positive or negative reviews. Prior work has shown how users perceive a particular review. Bakshi et al [3] use an inference tree of social signals from Yelp reviews to study the social signals. Jindal [5] uses a random forests to find out if a particular user is an expert reviewer. She uses the user's topical knowledge and local knowledge to train the system.

1.2 Attributes and Graph
Nodes characteristics can be defined by its attributes. In a paper published by Akshaya et al [2] just when Twitter was getting started, worked on Twitter's graph dataset with particular emphasis on the geographical location of users or nodes.

1.3 Homophily
Users tend to club with other users who share common interests, this can be seen in social networks like Facebook and Twitter as observed in the paper by Matt Lamm et al [4]. This phenomenon that is observed gives rise to formation of structures within social networks. Formation of structures can be observed in a temporal way to understand influencers in a network.

1.4. Trends
When analyzing data related to trends there can be many factors affecting them. For example, in paper [6] they have found that when a user visits a business there is a good chance that it is in his/her neighborhood. Through analysis it seems that there is a weak positive correlation between the business ratings and its neighbor ratings. Thus by using geographical neighborhood influences lower prediction error has been achieved which was not possible even when using state-of-the-art models including Biased MF, SVD++, and Social MF. In paper [7] Levent Bolelli et al they have proposed a generative model based on Dirichlet allocation for mining distinct topics in document collections by integrating temporal ordering into generative process. They have adopted to separate the documents into segments and these separate segments are propagated to influence the topic discovery. The experiment results show that segmented topic model can effectively detect distinct topics and their evolution over time.

3. MAIN TECHNIQUE

3.1 Setting up the Google Cloud Infrastructure.
Since this project is going to be computationally intensive, an infrastructure was scaffolded inside Google Cloud with the following features,

1. MongoDB service and
2. Virtual instance (2 vCPU, 7.5 GB RAM) with IPython (Anaconda) installed.

3.2 Data Sampling:
Each of the three members were working on different tracks, and we required different samples of data to work with. Hence each of us worked with small samples of data regarding which we will be talking in the individual section.

3.3 Data Collection:
Yelp has provided an archive through a one click download consisting of four flat files filled with JSON objects. The collection could be broken down into the following,
1. Users,
2. Businesses,
3. Reviews and
4. Tips.

3.4 Subtasks:
Being a group project, the plan is to try the challenge from three different subject areas which are,
1. Natural Language Processing,
2. Trend Analysis and
3. Graph Analysis.

The above three areas are taken care of individually by team members The approach taken here with this project is to see how much can Yelp dataset answer to certain hypothesis. Following paragraphs will talk on how they fit into the above mentioned plan.

3.4.1 Natural Language Processing:
Yelp is primarily driven by user-written reviews. Reviews play a major role in determining a businesses success and in prompting new users to go check out a new place. Yelp users have the opportunity to write a review for a business and then rate the business from on a scale of 1 to 5 stars.

We were primarily interested in figuring out the correlation between the review and the stars that a user gives for a particular business.

Subtasks:

1. Data Sampling: Since there were 2 million reviews, the hypothesis was tested with a smaller sample of data. Random testing was done to select 5000 reviews each time.
2. MongoDB Joins: MongoDB joins to integrate the business and reviews collections were performed in the application since MongoDB does not support joins naturally.
3. Processing of reviews: NLTK Library was used to perform text processing, such as removing punctuations and stops and generating bigrams, trigrams.

Data:
The reviews and businesses collection of Yelp was joined using Map Reduce to improve the efficiency of performing a join on a NoSQL data store.

Methodology:
As an initial step, to test the hypothesis, bigrams and trigrams were generated from the reviews. 1 star and 5 star reviews were used to generate bigrams and trigrams from which word clouds were generated using using http://worditout.com/word-cloud/ and the word cloud library in python. The word clouds obtained were encouraging, with 1 star reviews generally containing complaints from the users, such as “poor service” or “waiting time” and 5 star reviews containing positive words and praises from the customers such as “great service”.

This prompted further analysis on this hypothesis.

Figure 1. 5-Star reviews

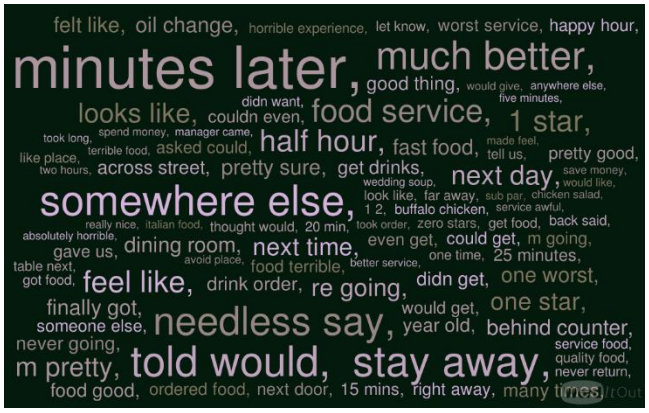

Figure 2. 1-Star Reviews

Regression Analysis:
The reviews were broken up into unigrams and then compared against a lexicon of positive and negative words which were compiled by Professor Bing Liu at UIC. Once number of positive, negative words in each review was computed along with the number of words in the review, linear regression of the number of stars on each of the attribute i.e. number of positive words, number of negative words, total number of words in a review was performed.

The results of the regression analysis are represented in Table 1, 2 and 3.

Slope	R value	P value	Std Error
0.43	0.14	4.52e-87	0.021

Table 1: Linear Regression between Stars and Number of positive words.

Slope	R value	P value	Std Error
-0.57	- 0.32	0.0	0.012

Table 2: Linear Regression between Stars and Number of negative words.

Slope	R value	P value	Std Error
-13.46	-0.17	6.8e-127	0.55

Table 3: Linear Regression between Stars and Length of Review.

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. These are expected results. 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.

Yelp ratings over the years:
In order to look at how yelp reviews have wavered over the years, a few graphs were generated to indicate how yelp ratings have changed over the years.

The graph of the number of yelp ratings over the years is very interesting. It shows that there has been a steady increase in the number of ratings over the years from 2005 till 2015. The ratings were pretty low in 2005 – 2010 probably because Yelp hadn't been very popular with the users then. It would be interesting to see the proportion of various ratings over the years.

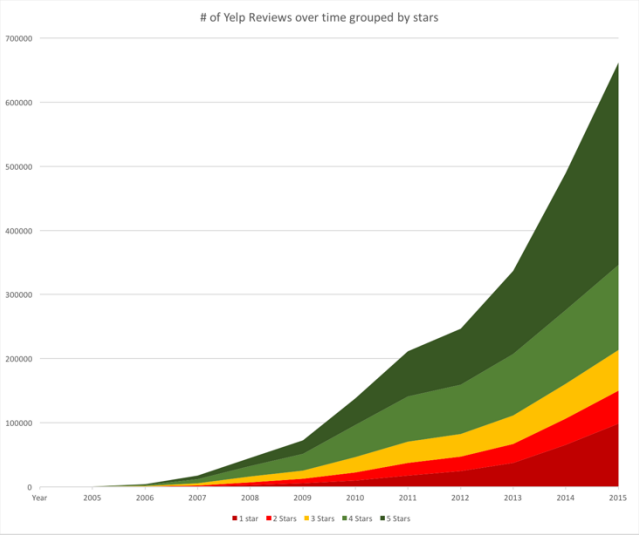

Figure 3. Yelp ratings from 2005-2015

Figures 6, 7 show the relative proportion of 1 star and 5 star reviews in 2005 and 2015.

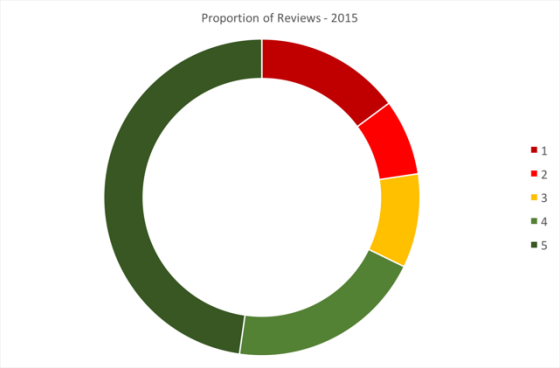

Figure 4. Rating proportions - 2015

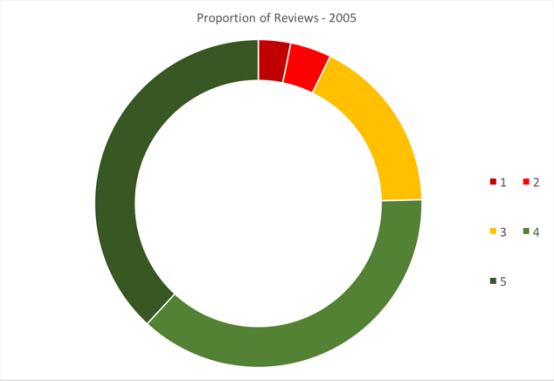

Figure 5. Rating Proportions - 2005

Classification:

The regression analysis on the reviews was converted into a classification problem, where 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.

3.4.2 Trend analysis:

We were interested in finding out the following trends.

1)Finding if people in the states considered - **Pennsylvania, Nevada, Arizona and Wisconsin** are going out more during the working days or in the weekends?
2)Finding out if there are any similarities in the trend pattern the states follow.

Procedure:
1)Data for each state was collected and analyzed 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 total businesses visited for a day was computed using these values.

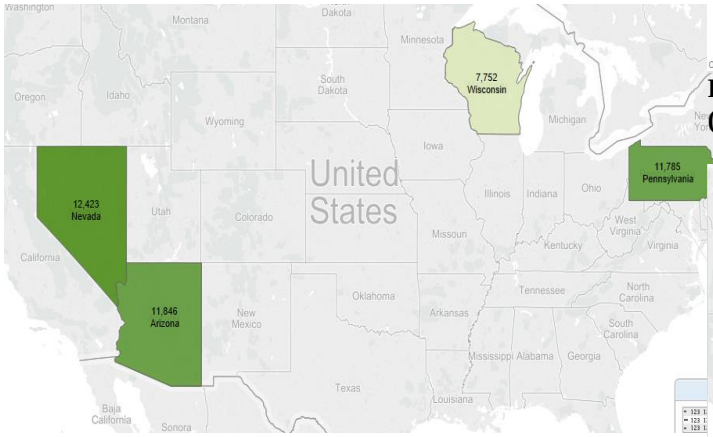

Figure 6. Total businesses visited in the morning(Monday-Sunday). (The dark green color represents the state that has the highest total and as the number decreases the color in which the map is represented, decreases)

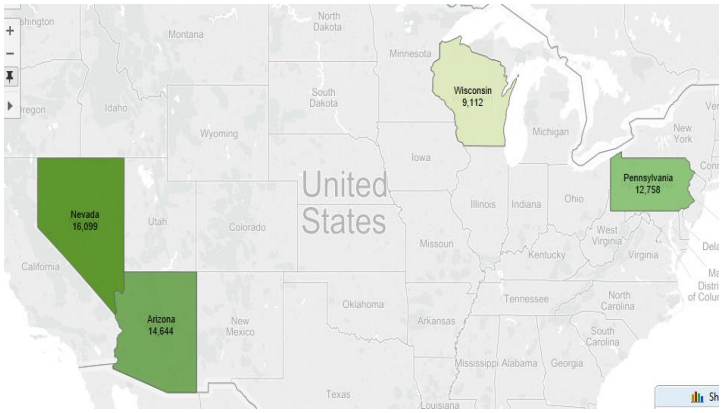

Figure 7. Total businesses visited in the afternoon (Monday-Sunday).

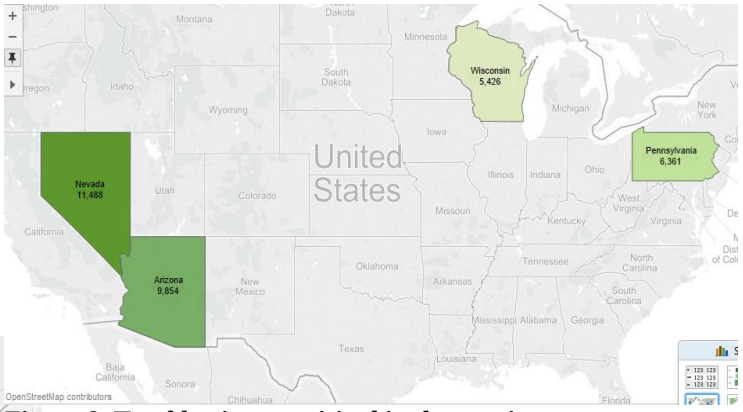

Figure 8. Total businesses visited in the evening (Monday-Sunday).

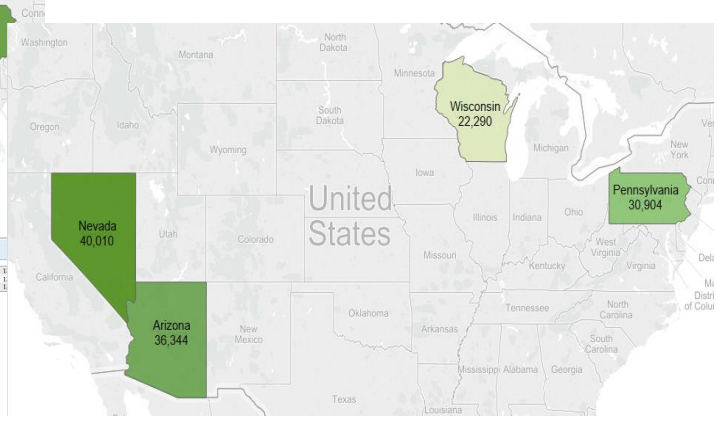

Figure 9. Total businesses visited in each state (Monday-Sunday). Here, Nevada has the highest number and Wisconsin has the lowest.

Days	Pennsylva	Arizona	Nevada	Wisconsin
Monday	4209	5231	5809	3061
Tuesday	4360	5302	5798	3118
Wednesd	4471	5336	5839	3182
Thursday	4880	5483	6045	3435
Friday	5114	5431	5863	3615
Saturday	3889	4408	4839	2964
Sunday	3981	5153	5817	2915

Table 4: Total businesses visited during each day of the week (Monday through Sunday) in each state.

A line chart was generated using the above data from table to see the trend pattern.

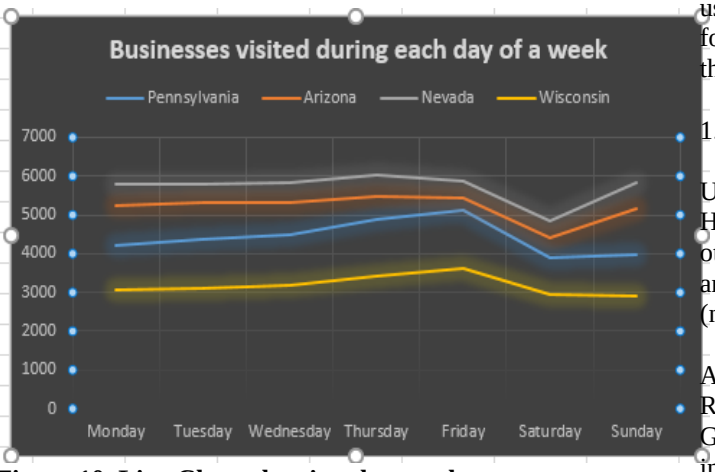

Figure 10. Line Chart showing the trend pattern followed in different states.

The results were matching with the predictions (People tend to go out more during the weekdays than weekends).

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.

Visualizations: We used Tableau to generate the maps.

3.4.3 Graph Analysis

Graph is usually a representation of a collection of nodes or vertices where a subset is connected through links or edges. Edges would usually have what is called weights which talks about the significance of the relation that exists between the two nodes.

The work that was done are the following,
1. Finding top influential users in Yelp network,
2. Study structure of Yelp network,
3. Reducing nodes and edges,
4. Building a multi-dimensional tool enabling exploratory analysis on graph.

About dataset and modelling done: Yelp's user dataset is used. A graph was formed with users as nodes and followership between users as edges. It is also understood that this is a directed graph.

1. Finding top influential users in Yelp network:

Users can influence in many different ways in a network. Here, it is taken that a user creates more influence than others if he/she is more connected (number of followers) and if he/she has connections with influential followers (number of influential followers).

An algorithm that suited this purpose was Google's Page Rank algorithm [14]. This is the initial algorithm used in Google's search engine. This algorithm in the context of internet ranks sites based on the number of visits.

Page Rank is adapted to count of number of followers and weight based on influential followers per user. This algorithm was implemented in Spark (Scala language) with the GraphX extension managed through Google's 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

Table 5: Ranked influential users generated based on implementing Page Rank algorithm on Yelp' user network.

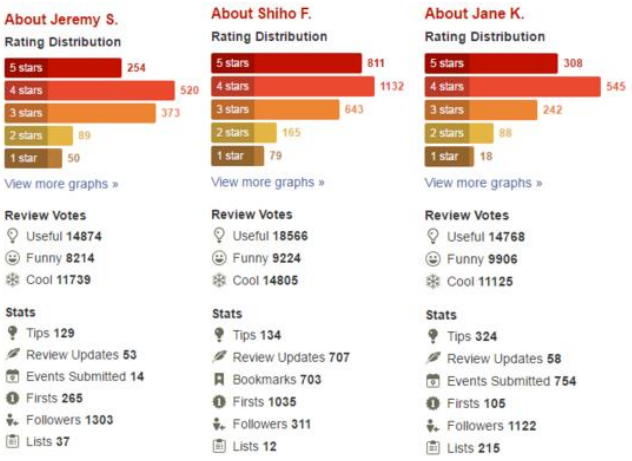

Figure 10: Statistics about top three users in Table 5 from Yelp.

Figure 11: Statistics about Jeremy S and Shiho F in Table 5 from Yelp.

Following are the observations made from this analysis,
a. Though Shiho has more reviews than Jeremy, she was ranked second. This is because Page Rank was made to count followers and not reviews.
b. Shiho has more followers than Jeremy but still she got ranked second. This could be because Shiho stopped being an elite user at 2011 (Figure 11) which in turn could have affected her influential followers' growth.

It is worth to know that Page Rank algorithm could be adapted in many different ways to suit requirements at stake.

2. Study structure of Yelp network:

When it comes to studying real networks, degree distribution stands out in being very important in understanding the overall structure of graphs.

In a graph, degree distribution is the probability distribution of degrees over the whole network. Degree of a node is the number of edges or links it has with other nodes. Since we are dealing with a directed graph here, we generate two degree distributions. One for in-degree and other for out-degree. In-degree of a node is the number of edges coming in and vice versa is out-degree.

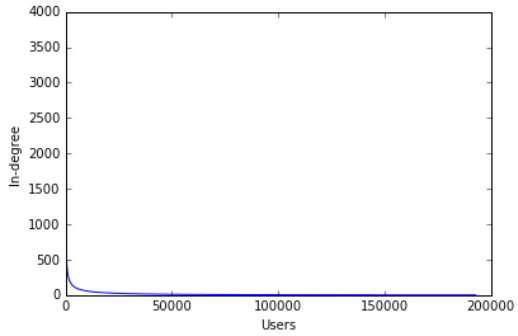

Figure 11: Users vs In-degree

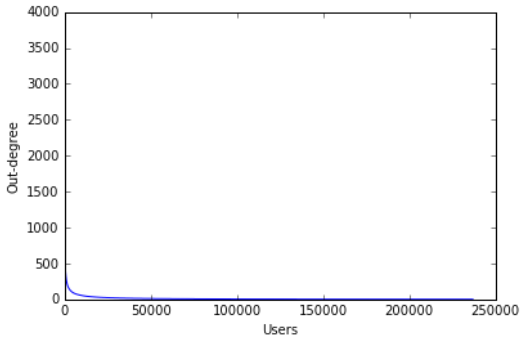

Figure 12: Users vs Out-degree

From the above two plots,
a. The degree distribution generated strongly represented a typical social graph through precedence of a Power law distribution [].
b. There are only few users with many followers and many users with few or no followers.
c. Yelp's user graph network is extremely sparse.
d. From the above points, it can also be said that Yelp's graph network is very similar to Facebook, Twitter or the Internet itself.

3. Reducing number of nodes and edges

The graph being used here has more than 3 million edges which can affect feasibility of any kind of analysis. Hence reducing the number of edges and nodes becomes a necessity.

The graph is sparse and hence the processing technique used should retain the graph structure (meaning, should not move the mean). Min-Max algorithm is used to reduce the number of edges through-out and the reduction leaves top ranked followers alone.

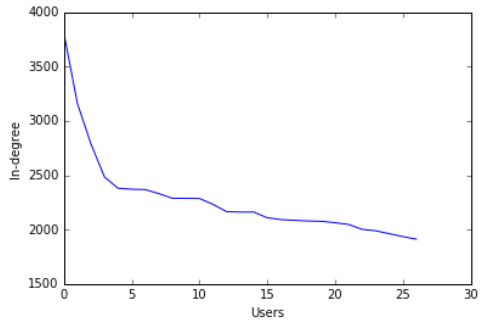

Figure 13: Users vs In-degree (after reduction)

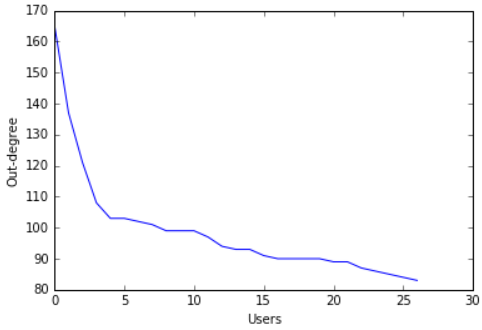

Figure 14: Users vs Out-degree (after reduction)

Figure 13 & 14 validates that the sparse structure of the graph was retained.

3. Building a multi-dimensional tool enabling exploratory analysis on graph:

Information dashboards are frequently used these days to do exploratory and descriptive data analysis. Visualizing graph in a way to do such analysis is tedious as the visualization work itself is a combination of Computer Science and Mathematics pulling concepts from even disciplines like Physics.

The reduced graph got from previous section was used here to do visualization.

The idea here is to project different attributes through different visual dimensions. Here the visual dimensions are the following,
a. size,
b. color,
c. link strength and
d. foci.

The attributes could be those related to each node or edge. Here, the weight for edge is computed on the fly.

The dashboard being built here comprises of the following,
a. dimension selector,
b. attribute selector for visual dimension selected,
c. the visualization itself.

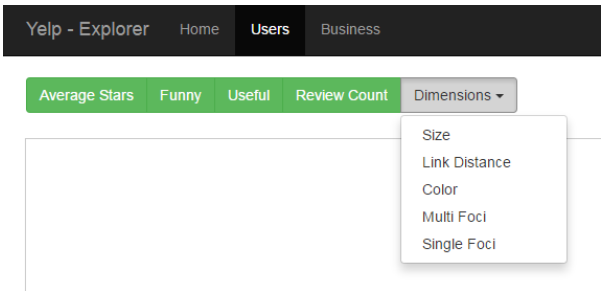

Figure 15: Assemblage of selectors for visual dimension and attribute

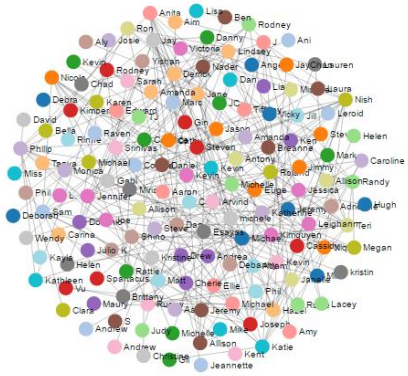

Figure 16: Force-directed layout applied to reduced form of Yelp's user network

This graph visualization uses Force-directed layout algorithm for drawing the graph in which the nodes have forces being exerted in-between.

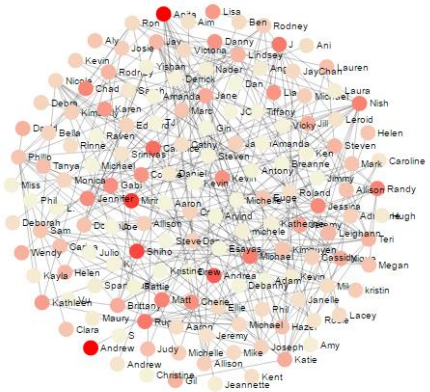

Figure 17: Color based on a selected attribute

The selected attribute is rescaled to a range of 1 to 10 and mapped to gradient scale with a range from 'beige' to 'red'.

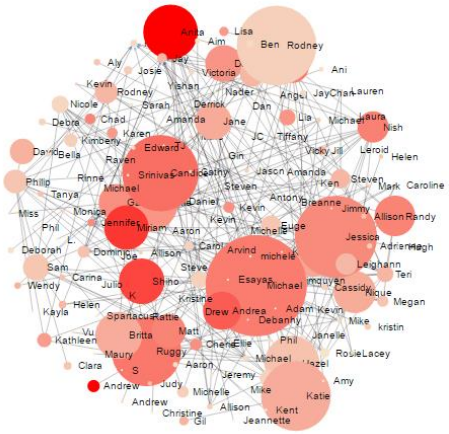

Figure 18: Size based on a selected attribute

In order to avoid over sizing a node or circle, the selected attribute is rescaled to a range of 1 to 100 and used for the circle size. Also to avoid overshadowing (one big circle over a small circle), the nodes are sorted and inserted.

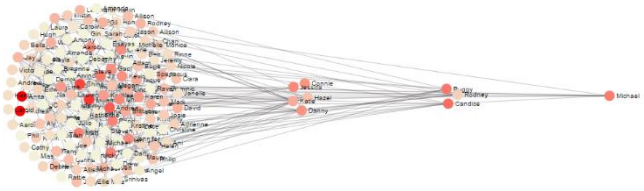

Figure 19: Foci based on a selected attribute

With Foci as a visual dimension, selected attribute is sorted by value, split into quartile and is positioned at four different points one after another.

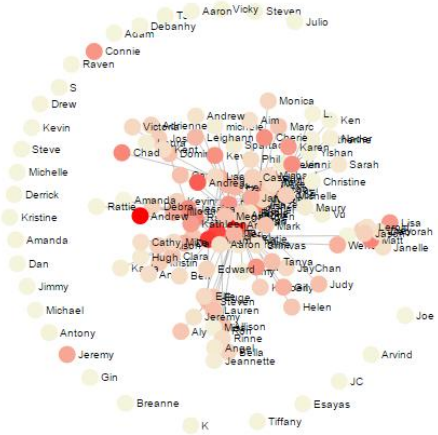

Figure 20: Link strength based on a selected attribute

In the above graph, attribute is rescaled to 1 to 10 and set as link strength.

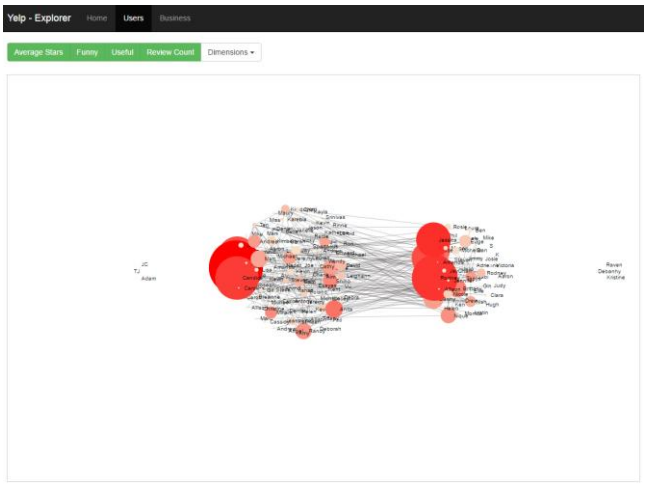

Figure 21: Dashboard with all the components

Steps involved in using the dashboard,
1. Navigate to 'Users' tab.
2. Click the dropdown 'Dimensions' and select one visual dimension you want to project an attribute through.
3. Select an attribute from the green colored button group.
4. The selected should get applied to the visualization. Repeat 2 and 3 to add more combinations.
5. 'Users' tab could be clicked to return to a neutral state.

Note that more than one combination of a visual dimension and an attribute could be included inside the visualization. They are almost like layers one above other.

This dashboard is hosted at,
http://v-pravin.github.io/yelpExplorer/users.html

The power of information visualization lies in knowing about the dataset in the shortest time possible.

Following are some of the insights got from the dashboard in just few minutes from selecting size for 'funny' attribute, color for 'review_count' attribute and foci for 'useful' attribute,
a. Matt's reviews a lot but his reviews are less useful and less funny.
b. Rodney's reviews are useful and funny even though he doesn't review much.
c. Michael actively reviews about businesses, his reviews are funniest and also most useful.

4. EVALUATION

4.1 Natural Language Processing:

Accuracy score:

The accuracy_score function of the Scikit-learn library computes the fraction or count of correct predictions made by the classifier. 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 precision is the ability of the classifier to minimize the false positives whereas Recall is the ability of the classifier to identify all the positive samples. As it can be observed from Figure 1, the average precision is about 0.70 and recall about 0.67.

	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

Figure 1. Precision and Recall

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 of 2 million reviews, it still is good enough to predict a lot of variations.

4.2 Trends Analysis

When looking at the line charts we can see that some states are similar in the way they follow the trend pattern. But we cannot come to a conclusion by looking at the map since there are many factors to be considered like the total number of population in a state before generalizing that some states are similar. So, normalization (min-max normalization) was performed by considering several factors and then a conclusion was drawn by finding Euclidean distances between the records(states) considered.

States	total businesses visited in a week - Monday - Sunday	Total population in 2015
Pennsylvania	30904	12802503
Wisconsin	22290	5771337
Nevada	40010	2890845
Arizona	36344	6828065

Table 1. Data to be normalized (Normalization of the total businesses visited in each state and the total population of those states was done).

normalization

	Total businesses visited in a week (Monday - Sunday)	Total population
Pennsylvania	0.4861	1
Wisconsin	0	0.2906
Nevada	1	0
Arizona	0.7931	0.3972

Table 2: Normalized table

Then, **Euclidean distances** were computed between records to find out the most similar records (states) using the normalized values.

$$d(\mathbf{p}, \mathbf{q}) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2}.$$

	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

Table 3: Finding similar states by computing Euclidean distances.

Thus, the most similar states from this given data is evident.
1)The entry **0.4478** shows that **Nevada** and **Arizona** are the most similar states that has the same trend pattern (During the weekends alone, there is a drastic drop on Saturday and sudden increase on Sunday).
2) Whereas our prediction of **Pennsylvania-Wisconsin** being similar states does not match the results of normalization since Pennsylvania has more population than Wisconsin and considering the total population and normalizing the data has given us accurate results.

4.3 Graph Analysis

1. Ranking

Accuracy here is measured by quantifying the loss wherein those incorrectly ordered users are accounted as loss. This way of finding the accuracy is also called as *Label Ranking* [16]. For the ranking generated through Page Rank the average precision score obtained is 1.3%. This gives us a ratio of correctly versus incorrectly ordered ranks.

2. IQR (Inter Quartile Range)

Interquartile Range (IQR) is a measures variability by dividing a given attribute into quartiles. This is used in validating split made when Foci is selected as a dimension.

3. User dataset is loaded into MongoDB and insights got from the visualization is evaluated by querying the dataset.

5. CONCLUSIONS

Natural Language Processing:
1) Correlation between Reviews and the stars given to the reviews.
2) 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:
1) People in all the 4 states considered, go out more during the weekdays than during the weekends.
2) The pattern of trend that the states Nevada and Arizona follow are similar.
2.1) Even though in all these states the population go out less in the weekends, the pattern for states **Arizona** and **Nevada** is similar – it decreses drastically on Saturday and increases on Sunday. The results were evaluated using normalization and the results were matching the predictions.
2.2) Whereas even though the line chart shows that states **Pennsylvania** and **Wisconsin** follow similar patterns, the pattern being - people going out on Saturdays and Sundays decrease constantly, the normalization results for it did not match since Pennsylvania had more population than Wisconsin and normalizing the data showed not much similarity between them).

Graph Analysis:
Though Page Rank through years becoming more complex and accurate, the algorithm still does an excellent job in helping ranking users and ultimately finding the top influencers. Just by generating probability degree distribution, we were able to understand the characteristics of the user graph regardless of size. It could also observed information visualization has helped a lot in pulling out insights from the dataset in shortest amount of time on Yelp's user network. As an ensemble of the work done could lead to better understanding about networks in general regardless of the context and further adaptation of new techniques.

6. FUTURE WORK

Natural Language Processing:
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:
1)The total number of businesses visited in the morning, afternoon and in the evening were different in each state. Finding out what reason may it be for the people in some states go out more in the afternoon, mornings or in the evenings?
2)Finding the reason behind a business 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? Using the results of it one can judge the working and efficiency of a business and recommend it to friends of friends if it is really doing well in what they do.

Graph Analysis:
The analysis done in regard to this part of the project is related to User dataset. Future work could involve expanding to Business and Reviews dataset. Working with Business and Reviews dataset could involve modelling relations between nodes.

Cliques are a subset of vertices such that any two vertices are adjacent. Study of cliques in a graph is a potential future course of this project as it has been widely accepted that cliques facilitate inter-group communications and give opportunities to model clusters.

The interactive dashboard that is built could further be improved by expanding to new ways of doing exploratory and descriptive analysis.

7. 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] http://aimotion.blogspot.com/2009/09/data-mining-in-practice.html
[12] http://kb.tableau.com/articles/knowledgebase/filled-map-custom-groups
[13] https://en.wikipedia.org/wiki/List_of_U.S._states_and_territories_by_population
[14] "The PageRank Citation Ranking: Bringing Order to the Web", L Page, S Brin, R Motwani, T Winograd, 1999, ilpubs.stanford.edu.
[15] "Power-law distributions in empirical data", Aaron Clauset, Cosma Rohilla Shalizi, M. E. J. Newman, 2009, SIAM Review.
[16] "Mining multi-label data. In Data mining and knowledge discovery handbook", Tsoumakas, G., Katakis, I., & Vlahavas, I, 2010, Springer US.