CSE 6242: Data & Visual Analytics
Fall 2018
Progress Report - “Yelp me out here!”
Team #36 – The Plug-ins
November 4, 2018

Introduction:

Setting up a new business or qualitatively analyzing pain areas to identify areas of improvement are challenges every businessman is familiar with. Even now less than 30% of businesses survive past their first year of operation. We are here to “Yelp you out!”

Problem Statement:

We aim to develop a resourceful business platform that would enable existing business owners and entrepreneurs to establish and flourish their businesses at a healthier and faster pace. Merchants currently lack the tools and resources to statistically evaluate their performance and hence make abstract decisions based on that. We seek to eliminate this issue by providing:

- Conclusive analyses of performance of their business based on customer needs
- Patterns and satisfaction quotient
- Dynamic logistic requirements
- Recommendations for new outlet location.

Survey:

Vast and varied research has been carried out on the Yelp dataset which mine for interesting insights into the user trends, perceptions and business operations and the correlation between user reviews and business performance. They have been mostly specific to a particular insight, say for example, finding local geographical experts and influential users[2][3], Or they may point the impact of a user review on Businesses[4]. Some highlight the need for visualizations to improve comprehension[15][16]. And there are some more which focus on the sentiment expressed in text, and use it to predict the rating[18]
Most of the above mentioned research are scattered and disconnected. Our effort is to build an insight tool which aggregates these findings and leverages it to apply specifically to the Yelp Dataset. Below, we describe the relevant work, which provide the basis for building our tool.

A. <u>Visualization</u>

Visualisations are an integral part of our project. There have been past works which talks about effective visualisations. Lee et. al.[15], and Vliegen et al.[16] both indicate

the red flags in a visualisation that makes it difficult to extract information out from it. As mentioned in Wijk[17], effective visualisations should enable the users to generate useful and actionable insights.

B. Business impact of reviews

Reviews and ratings are critical to businesses if they want to increase their revenue. It's imperative that Businesses glean useful insights from these reviews to improve businesses and their reputation[4]. A reviewer may potentially influence the Yelp recommender system which correlates to more visitors to the businesses[9].

Interestingly, businesses recognize this importance, to the extent of faking reviews to improve businesses[6]. Businesses can strategize and set up their establishment in tune with various factors that impact customer satisfactions. The yelp data comes handy and informs businesses of the steps to take.[7]. In addition, the reviews also help understand the features to add, which positively affect businesses[8].

C. Factors influencing ratings and reviews

There is also a reverse correlation of how locations, neighboring competing business affect reviews. These reflect a hotspot where businesses appear to be clustered and are frequently reviewed. Even the rating system is dependent on how one one business compares to another[5].

Reviews are also dependent on how an elite/expert reviewer reviews an establishment[1]. Once can identify such experts from the Dataset itself[2] or through relationships established via reviews[3].

D. Sentiment analysis of textual data

Textual data can reveal People's opinions, evaluations, emotions and attitudes towards products, services, organizations and their attributes[10]. There are three levels of sentiment analysis Document-level, Sentence-level[11],[12] and Aspect-level[13]. These methods can be leveraged to successfully predict a rating from the review data[14]. Another method identified logistic regression, which used unigrams and bigrams in the review data to estimate a rating from the review[18]

Proposed Method:

We will be utilizing Yelp dataset to produce worthy, expressive analysis of current trends and patterns that could enable merchants with their decision making process.

Approach:

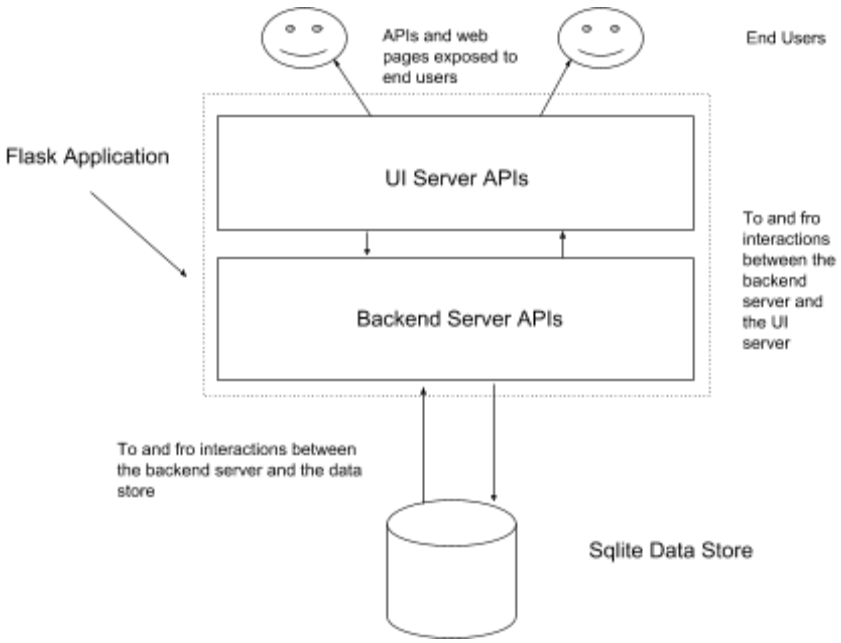
To facilitate the features in the project, we implemented the following components:

1. A flask backend server
2. Data store powered by Sqlite
3. A UI server

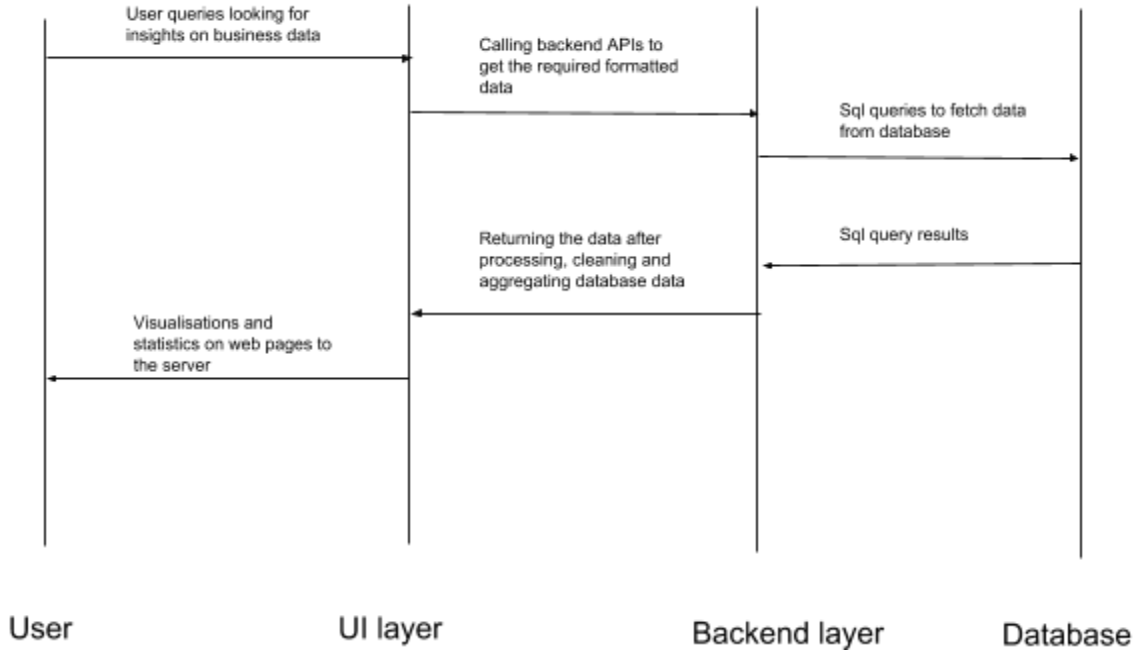
The functionalities of each of the components is as follows:

1. Sqlite data store : This is the main data store for us. We have loaded the whole yelp dataset into the mysql dataset using the same schema as in the data. All the data for analysis is churned from this data store using sql queries.
2. Flask backend server : This is the main server that powers the APIs to provide the relevant data in the required format to the UI layer. This layer interacts with the data store below it, processes the data and gives the data to the above layers.
3. UI server : This is the server that interacts with the backend APIs, get the data for each metric and displays the visualisation and stats through web pages and D3.

   The UI and backend server are logically different entities but both are hosted in the same flask application.
   The overall architecture and interaction among the components looks like this:



The interaction diagram for the components in our project is as follows:



Algorithms:

- We implement a tag cloud depicting the most commonly used phrases and the size of each word depicts the relative frequency of that word across reviews for any business. To be able to any useful data from the reviews, we need to preprocess them. For this, we use NLTK library. We will be using stemming to reduces inflected (derived) words to their root form. This helps in narrowing down to fewer word categories. For example, “fishing”, “fisher” and “fished” come from the same root word “fish”. Lemmatization is a process that’s a bit more involved. It also looks up the context and tries figure out a connection with other words. For example, “better” can be linked to “good” with lemmatization; stemming cannot do that. Bigrams are a sequence of two adjacent elements from a string of tokens, which are typically letters, syllables, or words. They can improve the understanding of the reviews. For example, “amazing service” tells us much more than just “amazing” or “service”.

- We implement a bar chart depicting the traffic volume of businesses throughout the day spread across geographies.

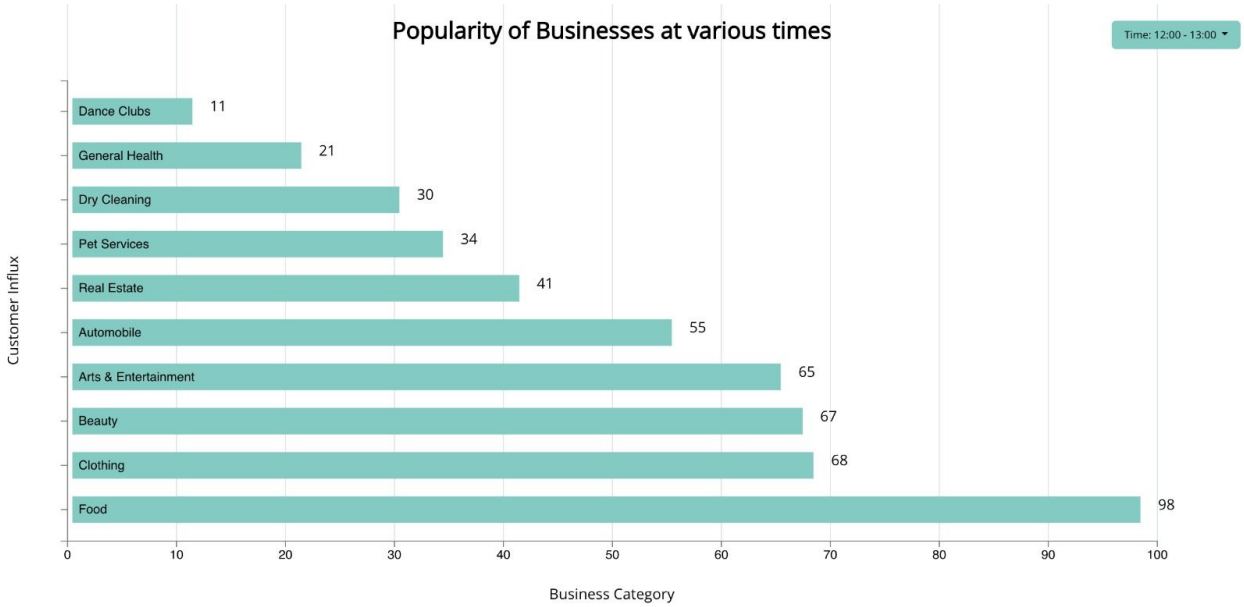
- We also implement correlation of secondary factors like parking, Wi-Fi, ambience, can also be connected to the review using machine learning on the dataset. We will provide a predictive tool with checkboxes for various secondary factors. This will enable business owners to see how introducing or withdrawing a service can affect the ratings and consequently the revenue.

- We also include a pie chart representing the distribution of these secondary factors across ratings for a genre of business. So a business owner could see directly which secondary factors are more popular with higher rated merchants in the same business.

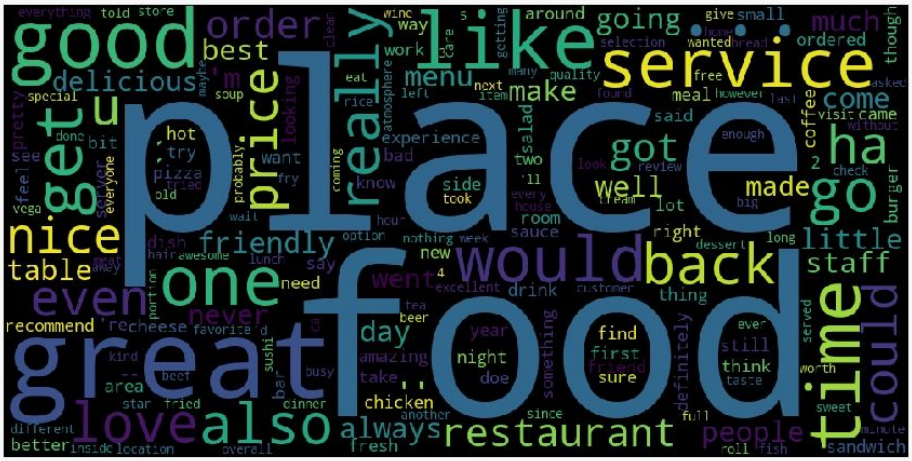
- We also plan to create a location based heatmap of the distribution of businesses so that the user can see what business can burgeon in which area.

Design of Experiments/Evaluation:

We plan to create multiple visualizations to better consume the data. Bar charts are an intuitive way of visualizing data. For example, we display below, the Peak Hour Analysis of businesses, which gives the owners an idea of how the work hours show be and how staffing and other logistics should be handled. The dropdown can be used to select the time span for which you want to see the data.



We produced word clouds for a better visualization of what the majority of the reviews say. We will be creating tag clouds for negative reviews and positive reviews separately. This will tell what is going good and what aspects need improvement.



To test the correctness, the plots will be verified against the dataset manually for a sample set. The authenticity of tag cloud and distribution of secondary factors(Wi-Fi, parking etc) across ratings (1 to 5) will also be checked for a sample set. Anomalies will also be visible intuitively incase of any disparity.

Plan of Action:

Original plan of action:

Participants	**Week 1**	**Week 2-3**	**Week 4-5 (finalpoint check)**	**Week 6**
	Data manipulation and DB & project setup	**Developing Metrics**	**Metrics cont. and UI/UX**	**Deployment and Testing**
Aastha and Anirudh	Clean data	Peak hour analysis metric	Impact of peer and expert review metric	Testing
Archit and Havish	Design Database schema and insert data	Geographical distribution of business metric	Correlation to secondary factors metric	Debugging and polishing
Manu and Sangharsh	Environment setup	Sentimental analysis metric (Phase 1 complete)	Implementing design details	Gathering feedback and improvising

Completed

Almost Complete

Progress so far and future plan of action

Current plan of action is same as original plan of action. Mid-point check-point (Week3) has been met.

All team members have contributed similar amount of effort, timely delivering individual deliverables and helping out team members whenever necessary.

- Aastha worked towards data cleaning and analysis
- Anirudh preprocessed the customer reviews, compiled literature survey and report
- Archit designed and set up the database
- Havish developed the metric for peak hour analysis (popularity of business at different points of time during the day)
- Manu developed the metric (tag cloud) for sentimental analysis of reviews to extract the most commonly used words across reviews
- Sangharsh set up the environment and the repository

All the mid term goals have almost been met and we are on schedule for final checkpoint.

References:

[1] The Influence Of Expert Reviews On Consumer Demand For Experience Goods: A Case Study Of Movie Critics
[2] Finding Local Experts From Yelp Dataset
[3] Identifying Influential Users in Social Network with Review Data
[4] Reviews, Reputation, and Revenue: The Case of Yelp.com, Michael Luca
[5] Analysis of Yelp Reviews, Peter Hajas
[6] Fake It Till You Make It: Reputation, Competition, and Yelp Review Fraud, Michael Luca
[7] Restaurant Setup Business Analysis Using Yelp Dataset
[8] Improving Restaurants by Extracting Subtopics from Yelp Reviews
[9] Measuring user's influence in the Yelp recommender system
[10] B. Liu. *Sentiment Analysis and Opinion Mining*. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 2012.
[11] W. Medhat, A. Hassan, and H. Korashy. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4):1093-1113, 2014.

[12] A. Collomb, C. Costea, D. Joyeux, O. Hasan, and L. Brunie. A Study and Comparison of Sentiment Analysis Methods for Reputation Evaluation. Technical Report RR-LIRIS-2014-002, LIRIS UMR 5205 CNRS/INSA de Lyon/Universite Claude Bernard Lyon 1/Universite Lumiere Lyon 2/Ecole Centrale de Lyon, Mar 2014.
[13] C. Rohrdantz, M. C. Hao, U. Dayal, L-E. Haug, and D. A. Keim. *Feature-Based Visual Sentiment Analysis of Text Document Streams. ACM Transactions on Intelligent Systems and Technology, Special Issue on Intelligent Visual Interfaces for Text Analysis*, 3(2):26:1–26:25, Feb 2012.
[14] Yelp Dataset Challenge: Review Rating Prediction
[15] : Sukwon Lee ; Sung-Hee Kim ; Ya-Hsin Hung ; Heidi Lam ; Youn-Ah Kang ; Ji Soo Yi - How do People Make Sense of Unfamiliar Visualizations?: A Grounded Model of Novice's Information Visualization Sensemaking
[16] : Roel Vliegen ; Jarke J. van Wijk ; Erik-Jan van der Linden - Visualizing Business Data with Generalised Treemaps
[17] : J.J. van Wijk - The value of Visualization
[18] Yelp Dataset Challenge: Review Rating Prediction, Nabiha Asghar.