Analysis of Yelp Restaurant Reviews to Recommend the Best Dishes in a Restaurant

Pown Arthi Thimiri Dayasagar, Sriharish Ranganathan, Revati Dhananjayan Lalitha, Aravindharaj Rajendran

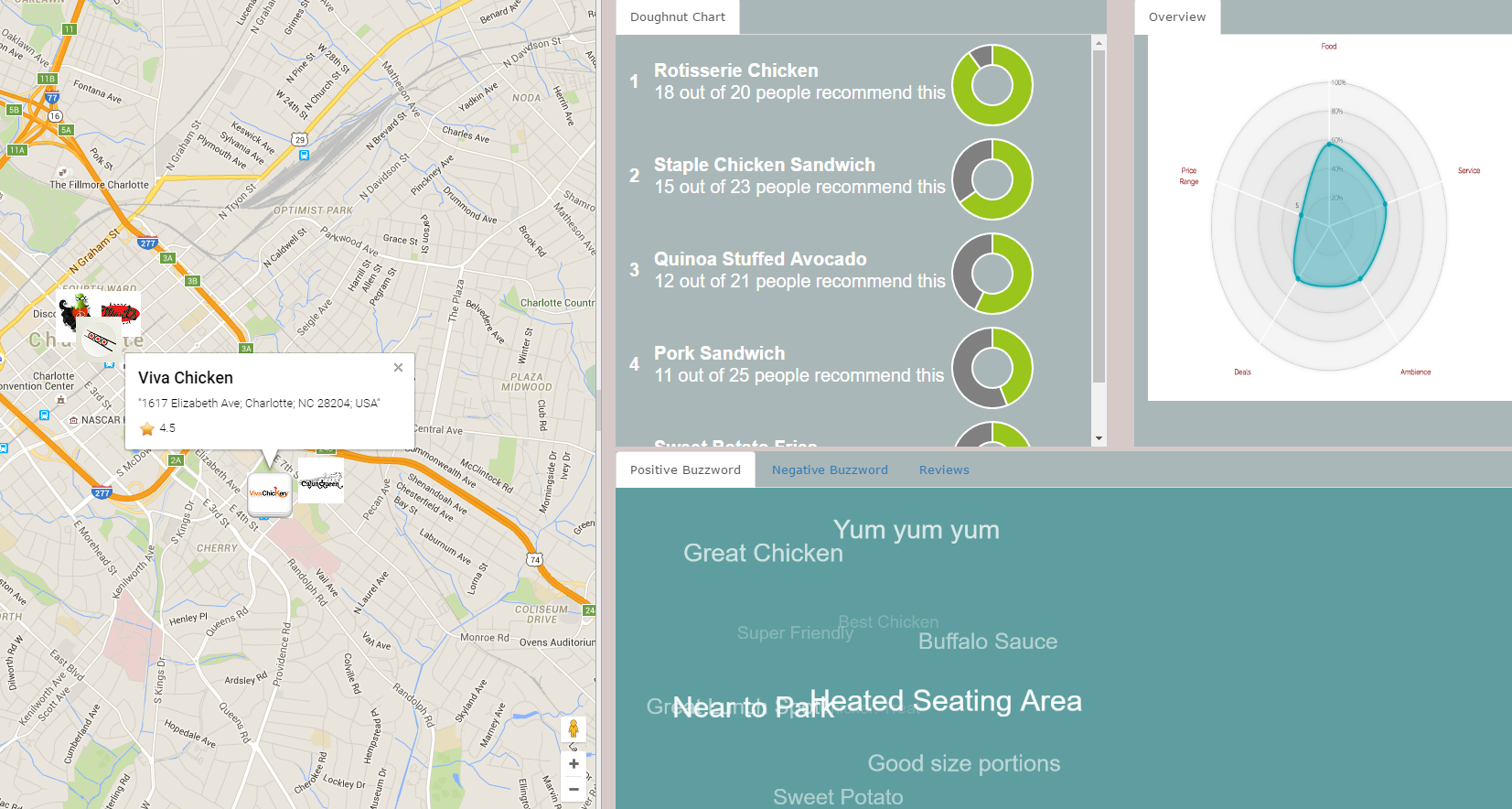


Fig. 1 – Dashboard view of the Recommended Dishes Visualization tool – Left Pane – 1. GeoMap showing the location of different Restaurants in the Charlotte area. Right Pane – From Left to Right – Top row: 2. Doughnut Chart displaying the Recommended Dishes to try out at the selected Restaurant. 3. Spider Graph showing the various business attributes of the selected Restaurant. Bottom row: 4. Tag Cloud showing positive and negative sentiment keywords from the User generated Reviews in two different tabs.

**Abstract**—Whenever we visit a restaurant for the first time, we always wonder what to order. One primary reason for this confusion is that we are not sure which food items will taste good and which ones will taste bad. In such scenario, we will most likely look online for reviews posted by other people who have already visited the restaurant. Only after reading all the reviews, we will get an idea on what to order. The Yelp.com is one of the popular websites where people from various regions of the United States are actively involved in giving their opinion / suggestion about different food items that they try out while visiting a restaurant. In this paper, we propose an interactive Visual Analytics system that analyzes the user generated reviews about different restaurants in Charlotte from the Yelp dataset and then recommends the best dishes to try out at a given restaurant. The system also performs Sentiment Analysis on the keywords extracted from the user generated reviews. Also, a graphical representation of different business attributes of the restaurant such as Price range, Ambience, Food Quality, Deals, and Service is also displayed in the system.

# Introduction

Yelp is an online business directory that lets users search for, rate, review, and discuss local businesses, both big and small. When we are new to some city or whenever we visit a restaurant for the first time, we are often apprehensive on trying out restaurants as we haven’t heard of them or we are not sure which food items will taste good and which ones will taste bad. In such situation, we either ask someone who has already visited the Restaurant to suggest some best dishes to try out or we even tend to ask the same to someone who is working at the restaurant; but here the user generated reviews, like those found on Yelp, can become important reference material in casual decision making. Our objective is to recommend the best dishes in a restaurant to the user based on the restaurant reviews in the Yelp dataset [1]. This could make it much simpler for the user to make a decision on which restaurant he/she must go and which food they could order. The system would be an interactive one with visualizations that would be easier for the user to understand and make decisions.

In recent years, user generated reviews have become a popular resource for decision making. Yelp provide a large quantity of user generated reviews for user decision making or reference. However, in most cases, reading large quantities of reviews is a difficult and time-consuming task. Also, a person does not want to spend a large amount of time making the non-critical decision of where to dine or what food to order. In this situation, a visualization that summarizes the user generated reviews is needed for perusing reviews. Our Visual Analytics system can effectively assist users by recommending the best dishes in a restaurant based on the restaurant reviews while providing high level categorization of reviews into relevant categories such as food, service, ambience, deals to help user understand why the restaurant is rated as “high” or “low”; alongside a clustered tag cloud with positive and negative words from review texts.

In a Yelp search, a star rating is arguably the first influence on a user’s judgment. Located directly beneath business’ names, the 5-star meter is a critical determinant of whether the user will click to find out more, or scroll on. In fact, economic research has shown that star ratings are so central to the Yelp experience that an extra half-star allows restaurants to sell out 19% more frequently. Currently, a Yelp’s star rating support the ability to visualize quantitative information of businesses. However, this quantitative information does not always account for the descriptions given in the user review, which can often be detailed, providing more context about the business that is being reviewed. So our system employs efficient techniques that have the potential to process user generated reviews and provide more context than only quantitative information.

# Related work

For this project, our group focused on the existing systems and came up with a model that takes the benefits of the existing systems. This section summarizes our research.

## Clustered Layout Word Cloud for User Generated Review

This paper [2] basically deals with how effectively we can represent the user generated Reviews in the form of Text Visualization so that it is easier and less time-consuming for people to make non-critical decisions. One such text visualization is the Clustered Layout Word Cloud where the most frequent decision-making words are gathered and their dependency among other decision-making words are calculated. Based upon the dependency, the words are placed at different clusters in the Word Cloud. We have planned to use this concept to gather the decision-making words from the Yelp user generated Reviews and cluster them based upon their sentiment. Ultimately, there will be two clusters in the Word Cloud where one corresponds to the positive sentiment (colored in Green) while the other corresponds to the negative sentiment (colored in Red).

## Online Spatial Data Analysis and Visualization System

This paper [3] develops an online spatial data analysis and visualization system, which facilitates end users to visualize and analyze spatial data, and to share the analysis results. Built on the spatial database, the GeoCloud which is an extra layer running upon the map can efficiently support many different visualization functions and spatial data analysis models. It also enables the MapQL technology to customize map visualization using SQL-like statements. Based on this concept, we plan to implement the visualization of restaurant’s location in Charlotte using the spatial data. The restaurants having more reviews can be displayed with a GeoCloud in order to show difference with the other neighboring restaurants.

## Multi-label Classification of Yelp Reviews into Relevant Categories

Yelp users give ratings and write reviews about businesses and services on Yelp. While these reviews and ratings help other Yelp users to evaluate the overall experience of a business; they do not convey the context which led a reviewer to that experience. Thus a high level categorization of reviews into relevant categories such as “Food”, “Service”, “Ambience”, “Deals”, and “Price Range” can help user to understand why the reviewer rated the restaurant as “high” or “low” and make a personalized choice, especially when one does not have much time to spend on reading the reviews [4]. Moreover, such categorization can also be used to rank restaurants according to these categories.

## Personalizing Yelp Star Ratings: A Semantic Topic Modeling Approach

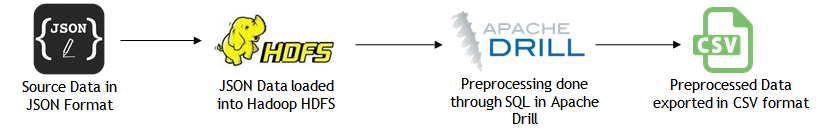
The objective of this paper [5] is to personalize the star based rating of Yelp search results by learning the expectations and preference of the user. This is particularly useful in our visualization system as we have a frame that displays the attributes of the restaurant such as cost, ambience, choice and variety of dish etc. This frame can be personalized for a particular user based on learnings from their previous reviews so as to increase the user experience. For example, if a user values quality and nutritional value of food over cost and ambience, we can highlight these attributes over others. These factors can be identified in two ways. One simple method is for the user to input these values when choosing the restaurant. The second method is to perform sentiment analysis over specific user’s review.

# System Design

The design of our system is divided into three sections. First we had to obtain and cleanse the relevant data. Next we had to build a predictive model for the system by performing analysis on the dataset. Finally, we had to implement visualizations for multiple attributes and connect them to the model.

## Data Collection and Cleansing

The dataset gathered from the Yelp website, contains almost 2.2 Million reviews, 566 thousand Business attributes and over 552 thousand business users. This dataset is in JSON format and contains information about local businesses in 10 cities across 4 countries. In order to retrieve the Restaurant information pertaining to Charlotte, the dataset was loaded into the Hadoop Distributed File System (HDFS). Then using the Apache Drill data exploration tool, the dataset is queried through SQL to retrieve the reviews and other information only for the Charlotte based Restaurants. After retrieving, the result is exported as a CSV file. The CSV is then used for analysis and creating visualizations.



## Fig. 2 – Data Preprocessing Process Flow

## Data Analysis

### Sentiment Analysis

Once the dataset is gathered, we performed Sentiment Analysis on the user generated reviews using AlchemyAPI. The analysis is done on two levels: Document-level and Keyword-level. Document-level analysis helps in finding out the overall feeling of the user for a given review. Keyword-level analysis identifies the keywords from a given review and performs Sentiment analysis on them.

### Extracting Popular Dishes in Restaurant from Yelp Reviews

Generating the recommend dishes from the user reviews is a four step process. First we gather all the positive sentiment keywords from User reviews using Sentiment Analysis. Next we collect the Restaurant “Menu” which contains the food items offered by that Restaurant. Then we compare food items from the Menu with the positive sentiment keywords and whenever there is a match, store the food items separately. And finally, count the occurrence of each unique stored food item and sort the result by count in descending order. This would give the most recommended dishes at the top of the list.

### Rating Business Attributes based on Yelp Reviews

Initially, the star rating of a restaurant is considered to set a default percentage value (in scale of 100) for its different business attributes. Then we gather all the positive sentiment keywords from User reviews using Sentiment Analysis. Following this, we compare positive sentiment keywords with word clusters containing different verbal representations of business attributes such as food, service, ambience and deals. Whenever there is a match, the default percentage value of corresponding business category is incremented. The computed result is then inputted as an array to a D3 program for generating spider chart visualization.

## Visualizations & Interface

Our system is divided into four resizable views, which are coordinated with interactions as shown in the Fig. 1. On the left panel is the GeoMap which is build using Google Maps API. The top right view shows the doughnut chart of the recommended dishes built using the JavaScript and in the next tab is the spider graph developed using the D3 framework that represents the overview of the selected restaurant. The bottom right view shows the interactive JavaScript Tag cloud of positive and negative sentiments and the corresponding reviews of the restaurant.

### GeoMap

As the dataset has the location coordinates of the different restaurants in Charlotte, a GeoMap is built using the Google Maps API. On clicking any of the restaurant, we would get a short description about the restaurant and its rating. Also, the right panel views will become visible on clicking the restaurant, which would display the overview of the restaurant, the recommended dishes and the positive and negative sentiments of the reviews for the restaurant.

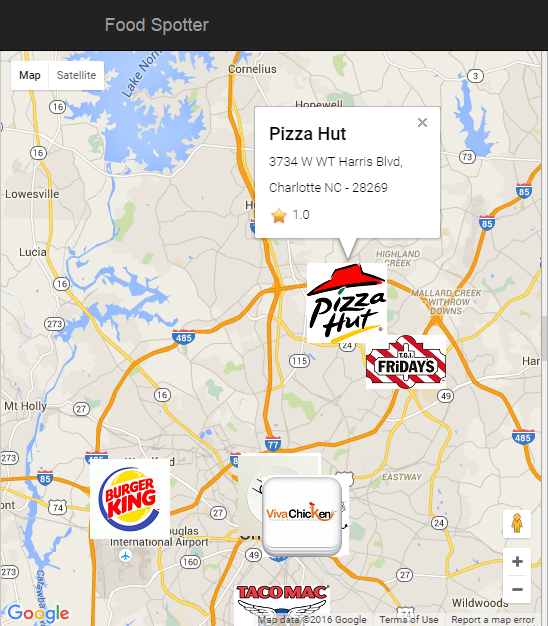


Fig. 3 – GeoMap visualizing the location of different restaurants in Charlotte

### Recommendations

A doughnut chart is generated using JavaScript based on the sentiment analysis of the user generated reviews. The green portion of the doughnut shows how many people have recommended the dish and the gray shows the remaining people in the total count of users. Suppose a customer is interested in ordering Pizza, it can be seen how many customers who had visited that restaurant have recommended for Pizza. Based on that the customer make his/her decision to order Pizza. The chart is very simple and easy to understand for the user to make a decision.



Fig. 4 – Doughnut Chart displaying the proportion of people recommending different dishes at the selected Restaurant

### Restaurant Overview

Yelp reviews and ratings are important source of information to make informed decisions about a venue. We conjecture that further classification of yelp reviews into relevant categories can help users to make an informed decision based on their personal preferences for categories. Moreover, this aspect is especially useful when users do not have time to read many reviews to infer the popularity of venues across these compelling categories such as price, service, food, ambience and deals. A spider graph is drawn using D3 library which shows the rating of the restaurant for the different business attributes. This graph could give an overview of the restaurant that makes it simple for the user to choose the restaurant based on his preferences of the attributes. Spider chart visualization is chosen for the reason that it is visually striking, and can add interest to what would otherwise be a dry data presentation.

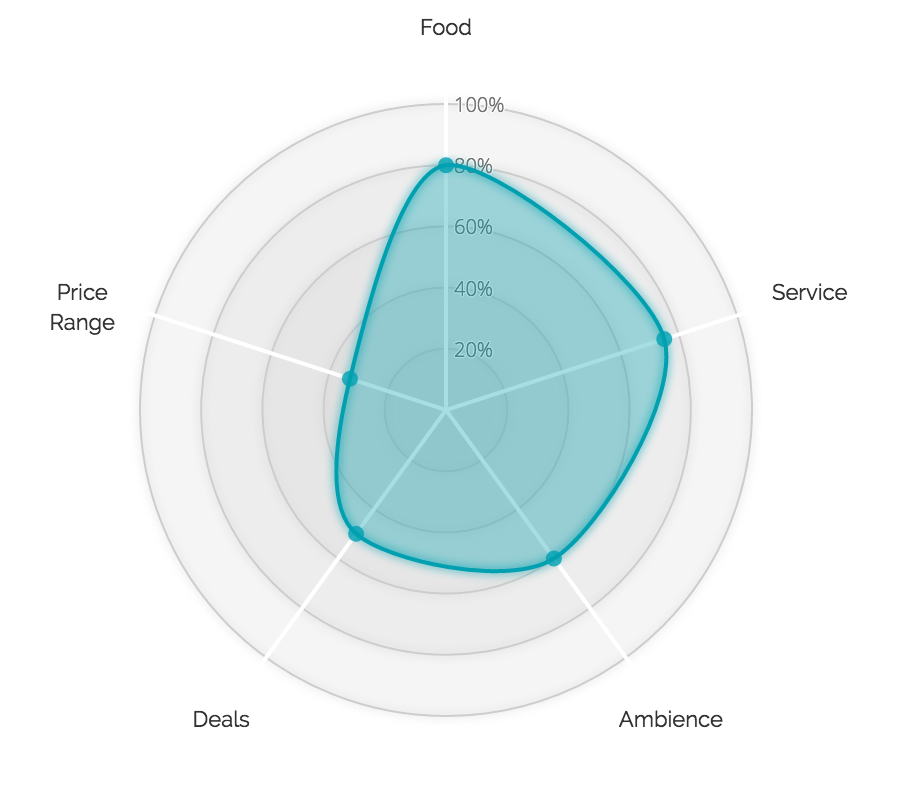


Fig. 5 – Spider Graph showing the different Business Attributes of the selected Restaurant

### Clustered Tag Cloud

The positive and negative sentiment keywords have been extracted from the analysis phase through the AlchemyAPI. Using that, Tag clouds are generated in JavaScript. The positive sentiments are shown in green and the negative sentiments in red. On clicking any of the keywords, the corresponding review gets displayed. This makes the system even more interactive for the user.



Fig. 6 – Tag Cloud showing the positive sentiment keywords colored in green



Fig. 7 – Tag Cloud showing the negative sentiment keywords colored in red

# Conclusion

In this paper, we have presented an interactive Visual Analytics system to recommend the best dishes in a restaurant based on the user generated reviews from the Yelp dataset. The system makes it simple and easy for the user to makes a decision on choosing the restaurant and placing an order. The possible future enhancements could be to provide a Map Search functionality where the user can directly enter the restaurant he/she is looking for. Also, the user’s preferences of business attributes could be taken into consideration which suggesting the restaurants. In conclusion, our Visual Analytics system can provide users with an efficient and more satisfactory experience during their review reading process.

References

1. Yelp Academic Dataset. <https://www.yelp.com/dataset_challenge>
2. Ji Wang, Jian Zhao, Sheng Guo, Chris North. *Clustered Layout Word Cloud for User Generated Review*.

<https://www.yelp.com/html/pdf/YelpDatasetChallengeWinner_WordCloud.pdf>

1. Julian McAuley, Jure Leskovec. Hidden Factors and Hidden Topics: *Understanding Rating Dimensions with Review Text*.

<http://i.stanford.edu/~julian/pdfs/recsys13.pdf>

1. Hitesh Sajnani, Vaibhav Saini, Kusum Kumar, Eugenia Gabrielova , Pramit Choudary, Cristina Lopes. *Multi-label Classification of Yelp Reviews into Relevant Categories.*

<http://www.ics.uci.edu/~vpsaini/files/technical_report.pdf>

1. Jack Linshi. *Personalizing Yelp Star Ratings: a Semantic Topic Modeling Approach*.

<https://www.yelp.com/html/pdf/YelpDatasetChallengeWinner_PersonalizingRatings.pdf>

1. Apache Drill. <https://drill.apache.org/docs/>
2. Alchemy API. <http://www.alchemyapi.com/api>
3. Java API for JSON Parsing. <http://www.oracle.com/technetwork/articles/java/json-1973242.html>