# Introduction

## Background

Eating meal at the restaurant is one of the most important and enjoyable activity for most of American and people around the world. What type of food, verities of dishes, test of the dishes, and level of overall service provided by restaurant play important role of meeting individual’s satisfaction. Each restaurant, especially operated by local business owner, provides some specialties items, varies in their test, and most importantly different level of overall services. Expectation associated with these dimensions is widely varied by customer and but play key role to satisfy customer. However, every customer likes different test of food and have different expectation on how services will be provided to them e.g. table location, taking order, delivering food, and even type of greetings. Two restaurants providing same type of menus, differ in food test and/or level of services. Therefore, selecting restaurant just based on menu may lead to different level of eating experience and overall satisfaction for individuals. So, it is advantageous for food lovers to accurately receive new restaurant recommendation based on restaurant type as well as similar people opinion before trying out new restaurant.

## Problem

Data that might contribute to identify best suitable restaurants (venues) for active customer might include available restaurant in nearby areas, type of restaurant, menu, and customer rating. This project aims to identify top restaurant based on content as well as user-based collaboration filtering recommendation algorithm.

## Interest / Targeted Audience

Obviously, customer, who want to try out new restaurant and leave it with high level of satisfaction, would be most interested in result. Others, like restaurant’s owners and operators may also interested knowing result and would able use it for directly or indirectly improvement of their business performance.

# Data Acquisition and Cleaning

## Data Source

Wikipedia is used to identified major cities of Oklahoma State. The list of restaurants based on geo location is fetched using Foursquare “Explore” API. Further details about restaurant e.g. number of likes, number of tips, rating etc. as well as users who likes them are fetched using “Venue” and “Likes” APIs respectively.

## Data Cleanup

Data is downloaded using multiple APIs and captured in pandas dataframes for further manipulation and transformations.

There are several limitations, data gaps and completeness issues were observed with foursquare dataset.

First, current Foursquare “likes” API doesn’t have rating information e.g. stars rating. Just knowing given liked it or not, is insufficient to calculate recommendation rating or similarity index for any recommendation system. Hence, I have generated rating number e.g. 1-10, using python random generator and assigned value to every user who liked one or more venues. Since I am generating rating value randomly, it won’t reflect true opinion of user, but it will prove concept. Also rating re-assignment rating will lead recommendation system to produce different results.

It appears that foursquare has recently changed data policy regarding fetching friends list for a given users. Result of which I am unable to fetch friends list for a given user. My original plan, when I started working on project, was to calculate weight based on likes of user’s friend and use it for final recommendation rating to improve likely hood of user to like recommended restaurant. However, due to recent change in data policy, I had to drop my idea.

Thirdly, venue category is selected by (I assume) business owner or assign personal by business entities. Hence, it is subject to individual decision to assign what so ever category he/she feels like to assign. Result of which, same kind of restaurant may have assigned different category and different kind of restaurant may have assigned same venue category. Inappropriate category assignment reduces the accuracy of recommendation system.

Venue list contains verities of venues. For restaurant recommendation needs to focus only restaurant type food serving venues. Therefore, I have captured all venue’, related only with food, drinks and coffee section, for these five cities, performed due diligence to removed unwanted categories which are not related to restaurant. Result of this exercise, 74 distinct venue categories are selected from foursquare explore API venue list.

The following venue categories are excluded. If new venue categories are included in selected section and associated venue is not restaurant, manual analysis needs to be conducted to exclude them.

|  |
| --- |
| Smoke Shop |
| Winery |
| Hookah Bar |
| Comedy Club |
| Arcade |
| Bakery |
| Hotel |
| Bookstore |
| Bowling Alley |
| Music Venue |
| Concert Hall |
| Food Truck |
| Liquor Store |

In addition to Venue Category, I have decided to include like count, tips count, price tier related information for recommendation decision. Some of the venue do not have value populated for these attributes. Instead of removing those records from samples, I have decided to populate appropriate default value for those attributes. Some of the restaurant, venue API didn’t return any record. Hence, I have removed those venue from sample. Result of which, I have 851 restaurants in final samples.

Active User Selection

Both recommendation systems, need to gather active user’s like details and build user profile. Hence, I have selected user, Danny C (user id: 13210167), who has captured reasonable amount of like e.g. 55 needs as active user for recommendation systems to generate restaurant recommendation list. Same work can be repeated for other users. Both recommendation algorithm suffers if we do not have venue’s like history for an active user. Hence, it is important to find users who has reasonable number of likes.

# Future Selection

Content based recommendation is based on users’ taste and the content or feature set items. The following sets of attributes are used as futures. Venue Category and Price Tier Message both are categorical attributes, hence one hot coding technique is used. Result of which, 93 new futures are created based on value captured in these two attributes.

Others are numerical attributes and their values are having different ranges. Hence, I have normalized those values to produce correct result.

|  |
| --- |
| Venue Category |
| Price Tier Message |
| Rating |
| Rating Signals |
| Likes Count |
| Photo Count |
| Tips Count |

In addition to user profile, User-based collaborating filters needs other user’s likes for various venues to build similarity between active users and other users and use it to recommend restaurant to active users. Other users’ likes and rating details for same venue’s which active users likes were extracted out of likes API dataset and used to generate necessary matrixes

# Exploratory Data Analysis

## Data Statistics after performing basic cleanup and transformation

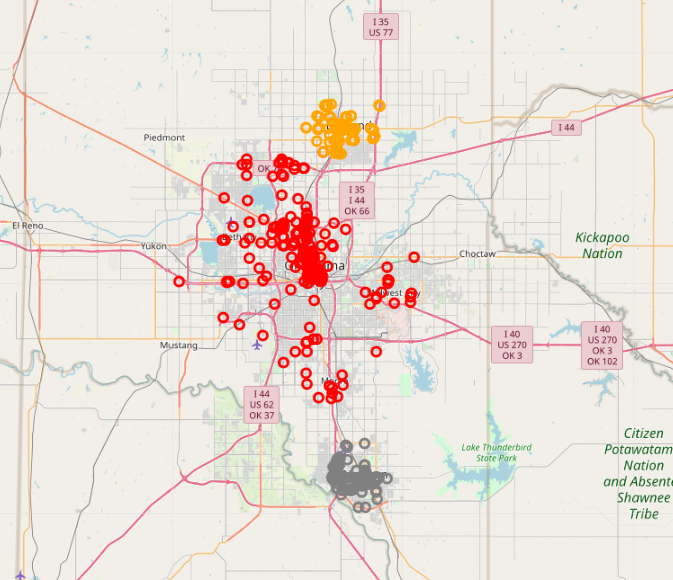
### Top 15 Venue Category by restaurant count

|  |  |  |
| --- | --- | --- |
| **No** | **Venue Category** | **Count** |
| 1 | Coffee Shop | 84 |
| 2 | Bar | 65 |
| 3 | American Restaurant | 52 |
| 4 | Donut Shop | 52 |
| 5 | Café | 46 |
| 6 | Mexican Restaurant | 49 |
| 7 | Burger Joint | 42 |
| 8 | Pizza Place | 37 |
| 9 | Brewery | 24 |
| 10 | Italian Restaurant | 21 |
| 11 | Sandwich Place | 20 |
| 12 | Fast Food Restaurant | 17 |
| 13 | Bakery | 17 |
| 14 | BBQ Joint | 17 |
| 15 | Asian Restaurant | 16 |

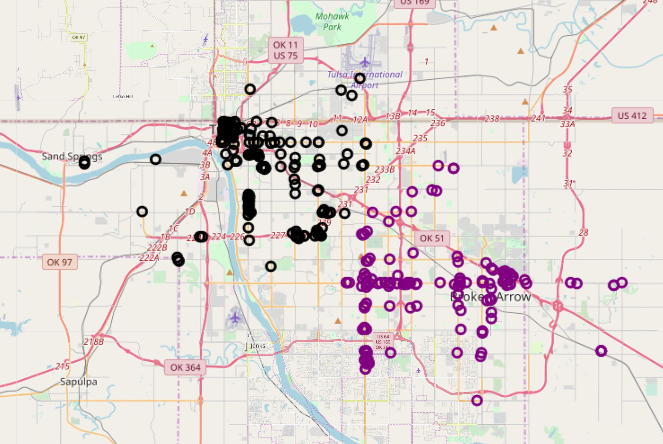
### Distribution of restaurant by city

|  |  |
| --- | --- |
| City Name | Restaurant Count |
| Oklahoma City | 279 |
| Tulsa | 187 |
| Norman | 148 |
| Broken Arrow | 134 |
| Edmond | 92 |

Oklahoma City / Edmond / Norman



Tulsa / Broken Arrow



### Restaurant distribution by Price Tier Message

|  |  |
| --- | --- |
| Price Tier | Restaurant Count |
| Cheap | 399 |
| Expensive | 59 |
| Moderate | 368 |
| Very Expensive | 14 |

### Top 15 users by likes counts

Total 23,557 records are fetched by likes API for point of interest venues (851). Same users have liked more than one restaurant. The following table shows, top 15 users who has liked restaurant(s).

|  |  |
| --- | --- |
| **User\_Id** | **Count** |
| 27816930 | 108 |
| 420364614 | 100 |
| 142383072 | 65 |
| 571110 | 62 |
| 11582716 | 61 |
| 85997014 | 60 |
| 37430283 | 60 |
| 430471 | 57 |
| 53631783 | 56 |
| 13210167 | 55 |
| 108718780 | 55 |
| 8451749 | 55 |
| 1858992 | 55 |
| 2175812 | 53 |
| 85182 | 52 |

### Target value – Calculation of overall recommendation index

For content-based recommendation system, user rating matrix is multiped with restaurant matrix (as result of hot coding generated features) to create weighed feature set for the restaurants. Aggregate the weighted restaurant and normalize them to generate active user profile. Then multiple user profile with other restaurant matrix to generate weight restaurant matrix and produce weighted rating. Aggregate these weighted rating to produce active users’ possible interest level in these restaurants.

Pearson Correction Distance formula is used to calculate distance between active user and other users and used to calculate similarity index or proximity between active user to other users. Weighted Rating Matrix (WRM) is created multiplying user rating for all restaurants (rated by active users or not) with similarity index. WRM represents the user’s neighbor’s opinion about the restaurants for recommendation. It incorporates the behavior of other users and gives more weight to the ratings of those users who are more similar to the active user. Finally aggregate all weighted rates to generate final recommendation matrix. However, different number of users may have rated each restaurant. Therefore, we need to normalize weighted rating value by dividing it by the sum of the similarity index for users. Result is potential rating that our active user will give to these restaurants if he/she has visited restaurant and it is calculated based on his/her similarity to other users in neighborhood.

### Relationship between Recommended Restaurant and Active User’s like profile

Type of restaurant what active user liked and provided rating value both play important role for both recommendation system. However, content-based recommender unable to recommend restaurant in different category if active user is never liked restaurant in that category.

Both systems heavily and positively depend on (higher) rating value. If active user has provided higher rating for a given category, content-based recommender assigns higher weightage to restaurant associated with that category and user based collaborative recommender give higher weightage to users who also highly satisfied with that restaurant.

### Relationship between Recommended Restaurant and like data availability from other users

User based collaborative recommender unable to recommend venue if there is no like data available from other user for the venue which active users liked.

I also noticed, active users have not liked venues from other cities, and most of the users in this city also have similar situation, content-based algorithm failed to suggest similar restaurant in other cities. If some of the other users have liked restaurant in other cities, but very few likes are available, recommender still unable to recommend restaurant from other city with high recommendation rank. In other word, user likes data availability, active user and other users common interest and variety and/or location of liked venues played important role for recommender to suggest restaurant from other cities.

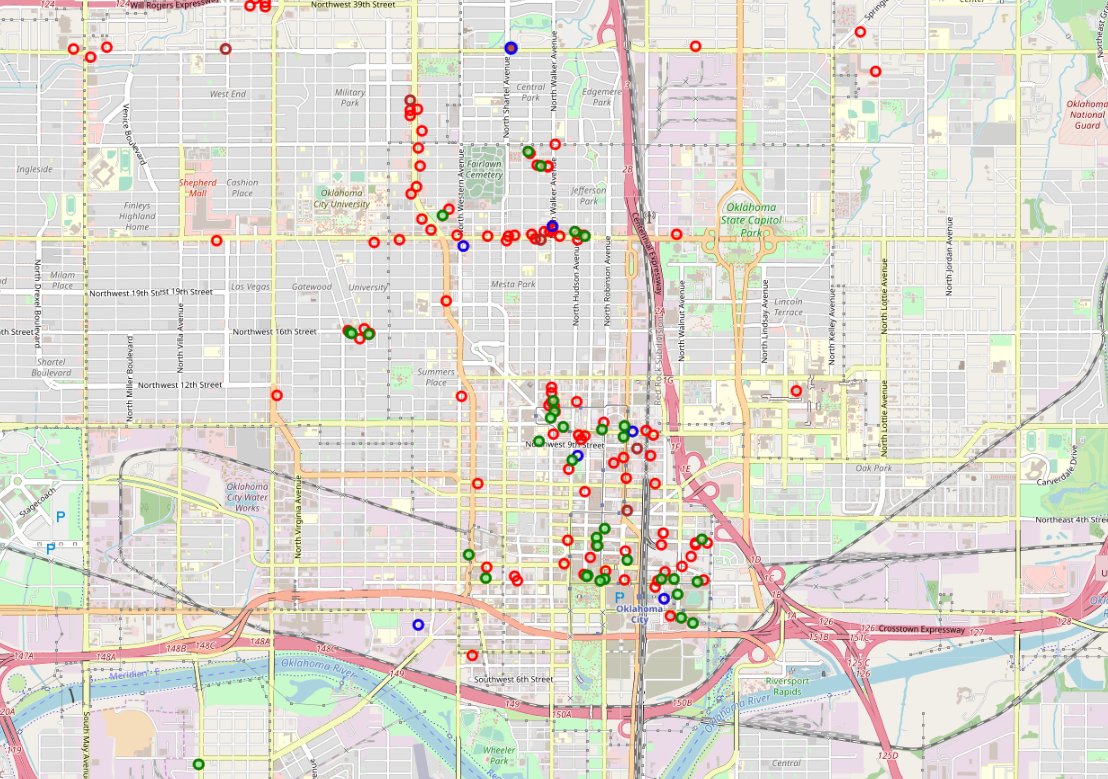
Visualization

Blue Circle with Brown fill – suggested by both recommender

Green Circle with green fee – User has rated / liked

Blue Circle with no fill – recommended by user-based collaborative filter recommender

Brown Circle with no fill - recommended by content-based collaborative filter recommender



# Recommendation Algorithms

There are two types of recommendation algorithms, content-based and collaborative filter based, are used to build restaurant recommendation system for active user. Content based algorithm provides recommendation based on user taste (e.g. rating, likes) and attribute (e.g. category) of restaurant, while user-based collaborative filter-based algorithm provides recommendation based on user of the same neighborhood with whom she or he shares common preferences.

Former algorithm tightly depends on user profile / taste as well as content’s properties e.g. restaurant category, so user who would like to try our new restaurant within same category would prefers result based on content-based system. Later algorithm heavily depends on user of the same neighborhood sharing common preferences and suggest new restaurant, irrespective of active user has every visited any restaurant in category of recommended restaurant. So collaborative recommendation system would be more suitable to those active users who likes to try out new restaurants even though he or she never visited similar type of restaurant before.

## Content-based

Content based recommendation system uses restaurant rating provided by active user to build user profile and uses it with restaurant matrix to build recommendation list.

If user has never rated restaurant in any specific venue category, content-based recommendation system will never recommend restaurant which falls under that specific venue category to active user. This issue can be resolved by user based collaborative filter recommendation system.

For active user e.g. Marty P, top two recommended restaurants from each give cities and every category which users have visited are listed below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No | City | Venue\_Name | Venue\_Category | User\_Reco\_Rating | User\_Reco\_Rank |
| 1 | Oklahoma City | BJ's Restaurant & Brewhouse | American Restaurant | 3.101529 | 1 |
| 2 | Norman | The Garage | American Restaurant | 2.320319 | 1 |
| 3 | Edmond | Charleston's Restaurant | American Restaurant | 1.8673 | 1 |
| 4 | Norman | BJ's Restaurant & Brewhouse | American Restaurant | 1.979411 | 2 |
| 5 | Oklahoma City | BJ's Restaurant & Brewhouse | American Restaurant | 1.91937 | 2 |
| 6 | Oklahoma City | Iron Star Urban BBQ | BBQ Joint | 2.109672 | 1 |
| 7 | Tulsa | Burn Co. BBQ | BBQ Joint | 1.853291 | 1 |
| 8 | Tulsa | Fassler Hall | Beer Garden | 3.404875 | 1 |
| 9 | Oklahoma City | S&B's Burger Joint | Burger Joint | 2.100532 | 1 |
| 10 | Tulsa | Tally's Good Food Café | Diner | 2.066337 | 1 |
| 11 | Oklahoma City | Ingrid's | German Restaurant | 5.641311 | 1 |
| 12 | Broken Arrow | Chuy's | Mexican Restaurant | 2.043784 | 1 |
| 13 | Oklahoma City | Whiskey Cake Kitchen & Bar | New American Restaurant | 4.481171 | 1 |
| 14 | Tulsa | SMOKE. on Cherry Street | New American Restaurant | 2.27968 | 1 |
| 15 | Tulsa | Andolini's Pizzeria | Pizza Place | 3.233129 | 1 |
| 16 | Tulsa | McNellie's | Pub | 4.706755 | 1 |
| 17 | Tulsa | Kilkennys Irish Pub | Restaurant | 5.547239 | 1 |
| 18 | Broken Arrow | Texas Roadhouse | Steakhouse | 1.989955 | 1 |
| 19 | Oklahoma City | Red PrimeSteak | Steakhouse | 1.879728 | 1 |
| 20 | Tulsa | Yokozuna | Sushi Restaurant | 2.07909 | 1 |
| 21 | Oklahoma City | Big Truck Tacos | Taco Place | 3.520624 | 1 |

## User-based Collaborative Filter

In addition to user profile, collaborative filter algorithm uses likes data for other users and use to identify similarity between active users and other users and generate recommendation list.

Data sparsity (user rates only a limited number of items), cold start (new users or new restaurant usually do not have likes history) and scalability (performance downgrade exponentially as number of users and/or restaurant increase) are typical challenges for collaborative filtering

For active user e.g. Marty P, top two recommended restaurants from each give cities are listed below:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| No | City | Venue\_Name | Venue\_Category | WA\_Reco\_Score | WA\_Reco\_Rank |
| 6 | Norman | BJ's Restaurant & Brewhouse | American Restaurant | 10 | 1 |
| 10 | Norman | Pei Wei | Asian Restaurant | 9 | 1 |
| 5 | Oklahoma City | Iron Star Urban BBQ | BBQ Joint | 10 | 1 |
| 16 | Oklahoma City | Power House | Bar | 7 | 1 |
| 21 | Oklahoma City | The Pump Bar | Bar | 6 | 2 |
| 11 | Oklahoma City | Bricktown Brewery | Brewery | 8 | 1 |
| 15 | Oklahoma City | Roughtail Taphouse | Brewery | 8 | 2 |
| 8 | Oklahoma City | Jazmo'z Bourbon St. Cafe | Cajun / Creole Restaurant | 9 | 1 |
| 12 | Tulsa | P.F. Chang's | Chinese Restaurant | 8 | 1 |
| 13 | Tulsa | Starbucks | Coffee Shop | 8 | 1 |
| 22 | Oklahoma City | Gourmet Donut | Donut Shop | 6 | 1 |
| 3 | Norman | Sergio's Italian Bistro | Italian Restaurant | 10 | 1 |
| 4 | Tulsa | Olive Garden | Italian Restaurant | 10 | 1 |
| 14 | Oklahoma City | Musashi's Japanese Steakhouse | Japanese Restaurant | 8 | 1 |
| 1 | Oklahoma City | Cookies Bar | Karaoke Bar | 10 | 1 |
| 7 | Oklahoma City | Sidecar Barley &Wine bar | Lounge | 10 | 1 |
| 18 | Oklahoma City | Mediterranean Imports & Deli | Mediterranean Restaurant | 7 | 1 |
| 2 | Norman | Tarahumara's Mexican Cafe & Cantina | Mexican Restaurant | 10 | 1 |
| 19 | Norman | Chipotle Mexican Grill | Mexican Restaurant | 7 | 2 |
| 20 | Norman | New York Pizza & Pasta | Pizza Place | 7 | 1 |
| 0 | Oklahoma City | Tamashii Ramen House | Ramen Restaurant | 10 | 1 |
| 9 | Norman | Firehouse Subs | Sandwich Place | 9 | 1 |
| 17 | Oklahoma City | Trapper's Fishcamp & Grill | Seafood Restaurant | 7 | 1 |

Output Comparison

* Only two restaurants are commonly recommended by both systems.

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| No | City | Venue\_Name | Venue\_Category | WA\_Reco\_Score | WA\_Reco\_Rank | User\_Reco\_Rating | User\_Reco\_Rank |
| 0 | Oklahoma City | Iron Star Urban BBQ | BBQ Joint | 10 | 1 | 2.109672 | 1 |
| 1 | Norman | BJ's Restaurant & Brewhouse | American Restaurant | 10 | 1 | 1.979411 | 2 |

* User-based collaborative recommender able to recommend top 2 restaurants from more categories than content-based recommender.

|  |  |  |  |
| --- | --- | --- | --- |
| **Content Based Recommender** | | **User-based Collaboration Filter Recommender** | |
| **Venue\_Category** | **Count** | **Venue\_Category** | **Count** |
| American Restaurant | 4 | American Restaurant | 1 |
| BBQ Joint | 2 | Asian Restaurant | 1 |
| Beer Garden | 1 | BBQ Joint | 1 |
| Burger Joint | 1 | Bar | 2 |
| Diner | 1 | Brewery | 2 |
| German Restaurant | 1 | Cajun / Creole Restaurant | 1 |
| Mexican Restaurant | 1 | Chinese Restaurant | 1 |
| New American Restaurant | 2 | Coffee Shop | 1 |
| Pizza Place | 1 | Donut Shop | 1 |
| Pub | 1 | Italian Restaurant | 2 |
| Restaurant | 1 | Japanese Restaurant | 1 |
| Steakhouse | 2 | Karaoke Bar | 1 |
| Sushi Restaurant | 1 | Lounge | 1 |
| Taco Place | 1 | Mediterranean Restaurant | 1 |
|  |  | Mexican Restaurant | 2 |
|  |  | Pizza Place | 1 |
|  |  | Ramen Restaurant | 1 |
|  |  | Sandwich Place | 1 |
|  |  | Seafood Restaurant | 1 |

* User-based collaborative recommender recommended restaurant from the following category which were not matching with active user profile.

|  |  |
| --- | --- |
| **Venue\_Category** | **Count** |
| Cajun / Creole Restaurant | 1 |
| Chinese Restaurant | 1 |
| Italian Restaurant | 2 |
| Japanese Restaurant | 1 |
| Karaoke Bar | 1 |
| Lounge | 1 |
| Mediterranean Restaurant | 1 |
| Ramen Restaurant | 1 |
| Sandwich Place | 1 |

# When limited first 40 recommended restaurants (order by respective rating index), It is observed that,

# Content based recommender recommended more restaurant in most of the venue category e.g. American Restaurant

* + Top restaurant selection in each category was also different between two recommendation systems e.g. Chuy's Mexican restaurant was selected by content-based recommender vs Tarahumara's Mexican Cafe & Cantina was suggested by user-based collaborative filter-based recommender
  + Some category, same top restaurant has been recommended by both systems e.g. Iron Star Urban BBQ

# Conclusion

In this study, I have analyzed restaurants located in selected five cities of Oklahoma state, users who have visited these restaurants, their ratings for these restaurants and their relationship. I have built two recommendations systems content based and collaborative filter based. Both systems have provided recommendation list for active users, however, they widely different in nature. It is observed that very few restaurants are comment in both recommendation list. Former system mainly selected restaurants which are tightly linked to similarity between restaurant visited by active users. Hence it ignored restaurant associated with categories which is never visited by active user. Later recommendation system overcomes this limitation and does suggest restaurants which are highly rated by other users who have similar preference as active users. Both algorithms suffer if active user doesn’t have history e.g. new user. In addition to it, former system suffers if restaurant is lack of accurate content e.g. category data and later suffer if other users did not have rating history for common restaurant with active user as well as rating history for other restaurants

# Future Improvement

Model can be further improved by utilizing additional meaningful content of restaurants, real rating for restaurant, restaurant rating provided by active users’ friends. Algorithm can further improve accuracy of recommendation list by using menu details e.g. items and ingredients, rating associated with other dimensions of restaurant e.g. customer service, food quality, attitude of attending personal etc., and sentiment analysis of tips text.