Restaurant Review Analysis on Yelp data

Omkar Reddy, Sayan Biswas, Sneha Agarwal, Varun Jagadeesh

Summary

Yelp is an online social business review platform where users can review the service of any business by rating them and providing detailed feedback. Among all business categories, the project focuses on the Restaurant business.

The motivation behind the project is to provide meaningful insights to the restaurant owners and users. The first objective is to provide the most positive and negative reviews to the users for each restaurant. This was later integrated into Shiny application so that the users and restaurants can select any restaurant and see the most positive and negative reviews. The second objective is to provide aspect based rating for each restaurant which will help the users in selecting the restaurants as well as help restaurant owners take business decisions. The third objective is to tell the users about the most talked about dishes in any restaurant.

The data set for the project is downloaded from the Kaggle website and it consists of seven data files that give information about 174,000 businesses and 5,200,000 user reviews. The size of the data was around 3.5GB. The project focuses on reviews for restaurants in Las Vegas since it has the most restaurants and the highest number of user reviews. Yelp Reviews and Yelp Business are the two CSV files used. The business file contains business name, unique identifier, star rating, review count available for each business as well as latitude and longitude coordinates. The review file contains a unique identifier for each business, a unique identifier for each review, the actual review text and the star rating for each review. For working towards the objectives a new tibble was created joining review table and business table on 'business_id', see appendix 1.A.

The text was split into tokens and hash_sentiment_senticnet lexicon was used to arrive at the word-wise sentiment which was then averaged to arrive at review wise sentiment. The reviews with the highest value of sentiment were considered the most positive and the ones with the lowest value the most negative.

The aspects in the reviews were discovered after initial EDA using word clouds and manual inspection of reviews. The aspects food, service, ambiance and likelihood of coming back were identified. A list of seed words for each aspect was identified to be used in the Guided Latent Dirichlet Allocation technique to assign each word to a respective topic and also each document to a topic. The same lexicon was used to estimate the sentiment scores for each review of a restaurant, later averaging them to get the overall sentiment scores for each aspect of the restaurant.

The results of the Guided LDA topic modeling were used to identify trigrams related to aspect "food" to fetch the most mentioned food dishes.

Methods

Approach for finding top positive & negative reviews:

The dataset contains reviews in multiple languages, only reviews belonging to English were selected using detect_language() function available in the cld3 package. The contraction words such as “isn’t”, ”couldn’t”, etc. were replaced with their full forms “is not”, “could not”, etc. to ensure that the sentiment of these words was not missed while calculating sentiment score. In-depth analysis was done on the reviews before the customized stop words were removed from the reviews using the removeWords() function from the ‘tm’ package. The reviews were then tokenized into unigrams[0].

Initially, AFINN lexicon was chosen to assign the sentiment score to the reviews where a polarity score was assigned to each word in the review. Upon using AFINN lexicon a lot of reviews were lost because the lexicon contains ~3300 words which are relatively very less for the diversity and number of words present in the review data. The same approach was validated using two different lexicons[2] namely hash_sentiment_jockers_rinker and hash_sentiment_senticnet as they contained much more number of words ~ 23000. Overall, hash_sentiment_senticnet was found to give better results.

The overall review score was found by averaging the sentiment scores for all the words in the review. The same was done across all reviews for all restaurants The reviews with the highest value of sentiment score was considered the most positive and the ones with the lowest value the most negative reviews. Since the context was not taken into consideration the results were not very accurate, hence, the reviews were tokenized into bigrams instead of unigrams. The sentiment score for the polarized word was negated if the first word of the bigram was a negation word. The results were more accurate when the context was considered.

Integrating with Shiny application:

The objective was to integrate the results of the sentiment analysis performed above into a shiny application to provide insights to the users and the restaurant owners using an interactive tool. Shiny[3] is an open source web application development framework for R. It can be leveraged to create interactive map interfaces by integrating it with Leaflet[4] map layers. The landing page of the application has an interactive leaflet map panel showing clusters of restaurants in Las Vegas, the size of clusters varying with the change in zoom level, see appendix 1.B.

A slider filter for restaurant ratings was provided allowing the user to select restaurants within a range of 1-5 divided over the intervals of 0.5. There is an individual location marker, see appendix 1.C, for each restaurant and is located on the map using latitude and longitude values available in the business table of the dataset. The restaurant name is displayed within a label when the user hovers over a restaurant marker. The user is redirected to a restaurant summary page upon clicking on a restaurant marker.

Restaurant summary page displays the restaurant name, top positive and top negative reviews, see appendix 1.D and 1.E, this gives the user a convenient way of getting a good overview of the restaurant rather than the arduous task of going over all the reviews. Additionally, for each selected restaurant a word cloud is created to visualize the most frequent words appearing in the reviews for each restaurant, see appendix 1.F.

Approach for calculating aspect-based rating using topic modeling:

The objective was to estimate an individual rating for each of the aspects like food, service, ambiance, etc. for each restaurant depending on the sentiment of the reviews. Post exploratory data analysis and manually inspecting a good sample of reviews it was very evident that most of the reviews comprised of a mixture of topics/aspects namely food, service, ambiance and if the user recommends it or not. Majority of the reviews had a higher proportion of the content related to food and service compared to ambiance and recommendations. Hence the reviews were classified into the above mentioned topics using Latent Dirichlet Allocation, a topic modeling algorithm for unsupervised classification of text data into discovered categories or topics.

Latent Dirichlet Allocation (LDA), treats each document as a mixture of topics, and each topic as a mixture of words or terms. Rather than being assigned to a distinct topic, this allows documents to overlap in topic.

The LDA() function from the topicmodels[5] package in R can be used to fit the topic model to the data, see appendix 1.G. The LDA() function in R takes in a document-term matrix, a number of topics to be used for segregation, control parameters such as setting seed for reproducibility and the method which is Variational Expectation Maximization (VEM) algorithm by default.

In pursuit of creating a Document Term matrix the below mentioned steps were followed:

- Each review was tokenized and broken down into sentences. Each sentence was considered a document.
- The sentences were cleaned by removing any digits or numbers from the sentences.
- The sentences were then tokenized into words.
- A customized set of stop words were removed from the list of tokenized words.
- Words with a minimum frequency of occurrence greater or equal to 100 in the overall document were considered.
- Words were stemmed to their root words by stemming.
- The document term matrix was created using the cast_dtm() function available in tidytext package.

The document term matrix created had each sentence of the review in a row, each term as a column, and each value in a cell as the count or frequency of a term in that particular sentence.

LDA method was fitted using the document term matrix created above, the number of topics was chosen to be 4 as our data or the reviews consisted of mainly four topics namely food, service, ambiance and likelihood to return, and a seed was set for reproducibility. The method used was “VEM” which is set by default.

The above algorithm for LDA works fine to give better results when the proportion of words belonging to topics is almost equal. But because of the proportion imbalance of words belonging to topics/aspects in the reviews, the topics were not segregated properly and could not be used for accurately providing an aspect based rating for restaurants. Also, VEM is a deterministic algorithm and is generally faster but it only gives local optimum.

In spite of the proportion imbalance, the need to classify the documents and words with minimal overlap of topic(s) allowed us to use the Guided Latent Dirichlet Allocation[6] (Guided

LDA) approach for topic modeling. LDA leverages the concept of seed words in which we can assign prior weights to a list of words to belong to a particular topic by using the method as Gibbs[7][8] while fitting the LDA model. Now the algorithm will have a higher count for seeded words belonging to the seeded topics and thus the probability of the word belonging to the seeded topic will be higher. Hence, the method "Gibbs" was used which allowed providing seed words to the model. Also, Gibbs follows the stochastic approach and is slower but approaches true distribution and provides better results, see appendix 1.H.

Seed words let us "lock" topics numbers by giving us the ability to know in advance which topic number corresponds to what words. It also helps speed up the algorithm convergence by providing a starting point. A sparse matrix was created with the number of rows equal to the number of topics and number of columns equal to the number of terms in the document-term matrix with values in the cell specifying weights for seed words for topics, see appendix 1.I.

The tidy() method from the tidytext package was used to extract the per-topic-per-word probabilities, called β ("beta"), from the model. The results were better than the VEM model and the terms were more accurately mapped to the respective topics. Then the per-document-per-topic probabilities, called γ ("gamma"), was examined with the matrix = "gamma" argument to tidy(). After obtaining the document to topic probabilities, each document(sentence of a review) was classified to belong to topic(s) depending on the maximum probability of belonging to a topic. If the document belongs to multiple topics with the same maximum probability, the document was classified to multiple topics as well. This provided us the classification of each document(sentence of a review) to aspect(s). Each row represented a sentence, each column represented an aspect and the value in each cell was either NA or "1" indicating the aspect to which the sentence belonged to, see appendix 1.J.

Sentiment analysis was done for each sentence of a review for calculating the sentiment score for the aspect (topic) which the sentence belonged to. sentimentr[1] package was used to quickly approximate the sentiment (polarity) of text by sentence. sentimentr attempts to take into account valence shifters (i.e., negators, amplifiers, de-amplifiers and adversative conjunctions) while maintaining speed. Simply put, sentimentr is an augmented dictionary lookup. The lexicon used to calculate the sentiment of the sentences is lexicon::hash_sentiment_senticnet which provides a sentiment score for ~23500 words. The sentiment scores were calculated for each sentence and the sentiment score was multiplied with the indicator values for each sentence (for each row), thus providing sentiment score of the aspect. The sentiment scores were normalized using appropriate customized intervals using cut() method to bring the range of scale between 1 and 5. Grouping and summarizing the sentiment score of the sentences of each review, the score for all the aspects for each review was approximated and further the same was done across each review to get the aspects rating of a restaurant.

Approach for finding the most mentioned food dishes:

The intent was to give the users the most frequently mentioned food dishes that appear in the positive reviews. These can be fairly assumed to be the best dishes the restaurant offers. For a restaurant, all the sentences which belonged to the aspect food were filtered. The sentences were then tokenized into trigrams using the unnest_tokens() method in the tidytext package. Each token of the trigrams was processed and filtered to remove stop words by anti-joining a list of customized stop words. These trigrams provided the list of most ordered dish/most mentioned dishes across all reviews of a particular restaurant.

Results

From Figure 1 it is observed that the restaurants with good rating have better sentiment scores across all aspects which also correlates with the overall restaurant rating whereas, for a lower rated restaurant the sentiment score for each aspect do not correlate with the overall rating individually.

	Name	Rating	Food	Service	Ambience	Recommended
1	Diablo's Cantina	3	3.646990	3.697770	3.595171	3.459011
2	Jose Cuervo Tequileria	2	3.337432	3.426718	3.352651	3.244262
3	J Karaoke Bar	5	4.056497	4.160630	4.125253	3.825666
4	Mon Ami Gabi	4	3.914593	3.852329	3.765863	3.688201
5	Bachi Burger	4	3.857731	3.587277	3.594510	3.503537
6	Pho Little Saigon	2	3.318328	3.278447	3.163285	3.226492
7	Zenaida's Cafe	5	3.900709	3.960843	3.741748	3.781457
8	Lady D's Pizzeria	1	3.000000	3.377778	3.625000	3.052632
9	Godfathers Pizza	1	2.909091	3.250000	2.000000	2.647059
10	Gordon Ramsay BurGR	4	3.923544	3.722984	3.652526	3.493873

Figure 1

From Figure 2, it can be seen that the median sentiment score of the aspects is increasing with an increase in the star rating of the restaurant. This supports the general hypothesis, the more the star rating the more individual aspect rating with some outliers. It was also observed that the 5 star reviewed restaurants had more individual service rating when compared to lower rated restaurants.

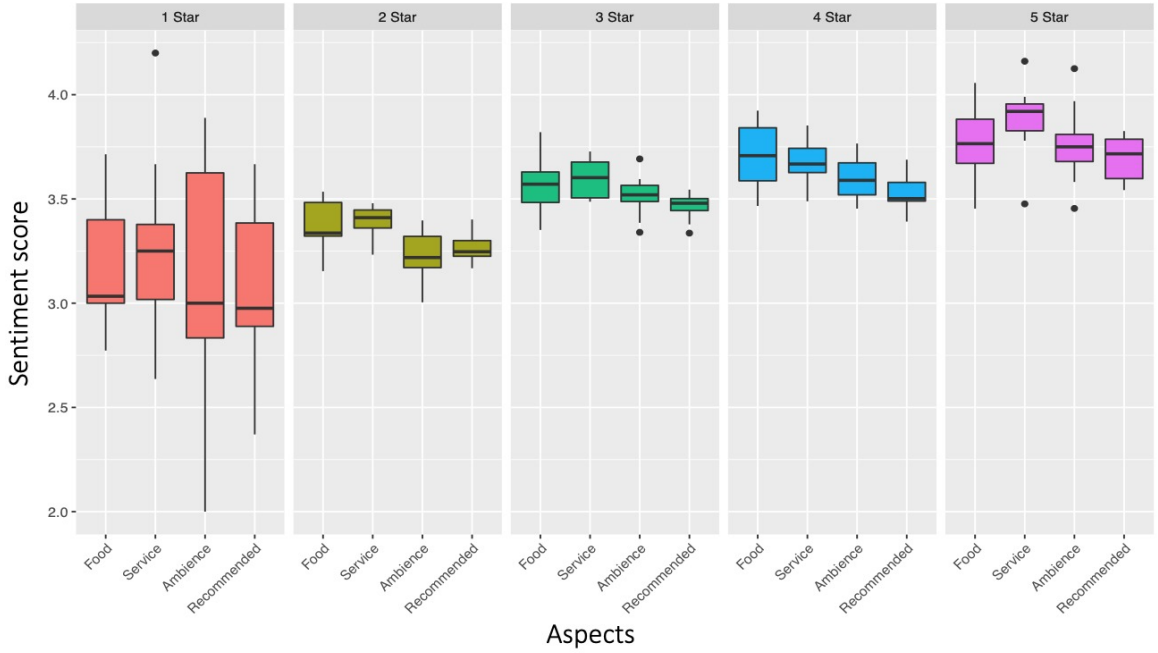


Figure 2 – Sentiment score distribution across aspects for each restaurant star rating

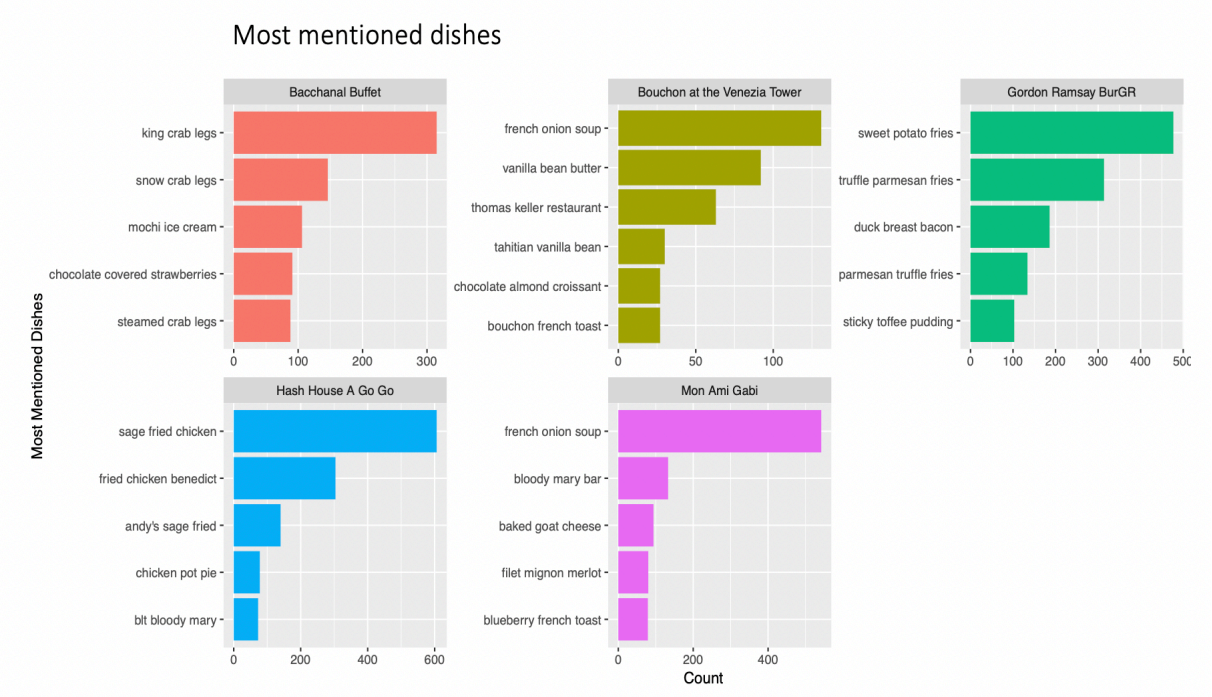


Figure 3 shows the most mentioned dishes for each of the restaurants

Discussion

The results from Figure 1 show that the overall rating may not be indicative of the rating across every aspect. The score across every aspect is a reflection of how the restaurant is doing in that particular aspect which is insightful to both customers and restaurant owners. For example, the restaurant “Godfathers Pizza” has a rating of 1, and the sentiment scores signify that the restaurant is doing fairly in terms of service but is doing poorly in remaining areas.

From Figure 2, it can be inferred that 1 star rated restaurants have a wider range of distribution of the sentiment scores for each aspect. It is observed that for most of these restaurants, even though one of the aspects has good sentiment score the other aspects having not so good sentiment score brought down the overall rating to 1. The same was observed and verified through manual inspection of the reviews. It can also be seen that 5 star rated restaurants are doing good in all aspects signifying that users rate restaurants as 5 star when they are satisfied and happy with every aspect of the restaurant. 5 star rated restaurants provide better service compared to other aspects.

Initially, the review summary page of any individual restaurant displayed top positive and negative sentences so as to excuse the user the hefty task of reading through long paragraphs of text. It was however noticed that most of these sentences were coming from the same review and the diversity of user’s perspective was not getting captured. It was therefore decided to retain the entire content of the original reviews.

While calculating the sentiments for the reviews, the hash_sentiment_senticnet lexicon was used which contained ~23,500 words. The words were not exhaustive enough and also were not domain specific. This reduces the accuracy of the calculation of sentiments. This can be improved by adding more words to make the list exhaustive. Next, we were not able to capture sarcasm while calculating sentiment scores.

Few of the signature dishes had lengthy names(5 words or more), as per the approach used it would come out as 2 separate dishes as trigrams were used to recommend dishes. This could be improved if we could scrape and store the menu for each restaurant from it's website and then later use it as a reference while recommending the dishes.

References

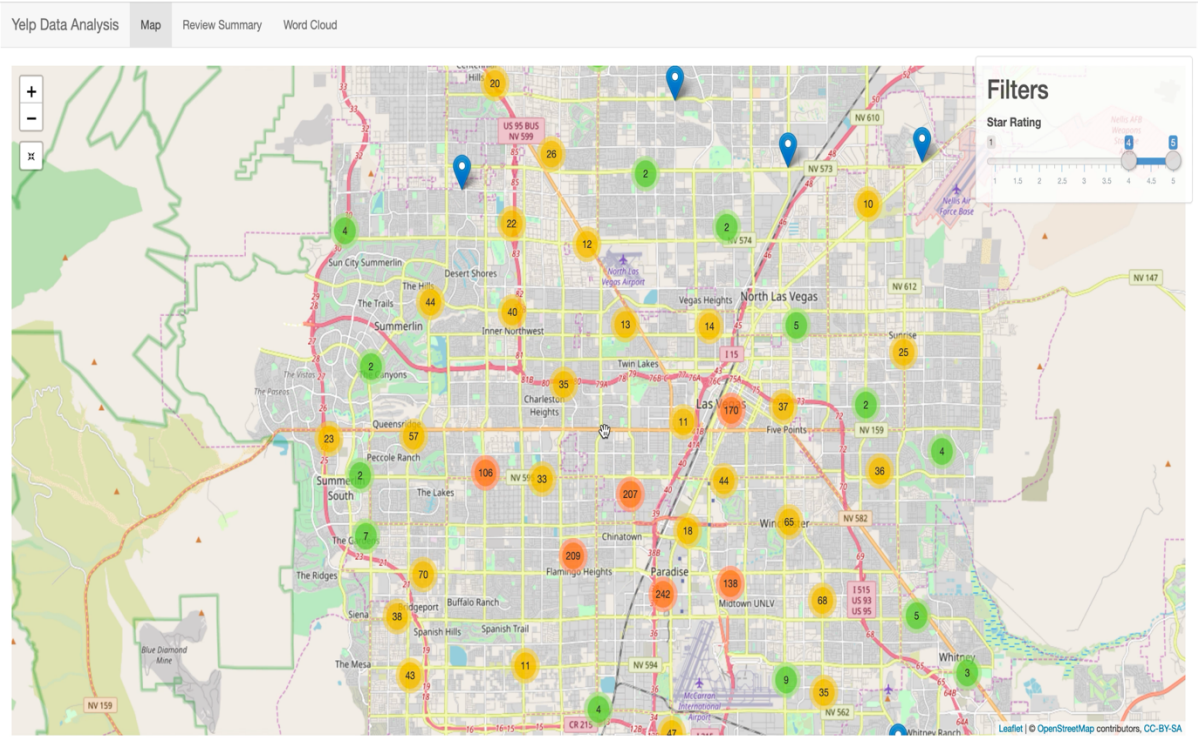
[0] https://www.tidytextmining.com
[1] https://cran.r-project.org/web/packages/sentimentr/sentimentr.pdf
[2] https://cran.r-project.org/web/packages/lexicon/lexicon.pdf
[3] https://shiny.rstudio.com
[4] https://www.datacamp.com/courses/building-web-applications-in-r-with-shiny
[5] https://cran.r-project.org/web/packages/topicmodels/vignettes/topicmodels.pdf
[6] https://medium.freecodecamp.org/how-we-changed-unsupervised-lda-to-semi-supervised-guidedlda-e36a95f3a164
[7] https://en.wikipedia.org/wiki/Gibbs_sampling
[8] https://ethen8181.github.io/machine-learning/clustering_old/topic_model/LDA.html

Appendix

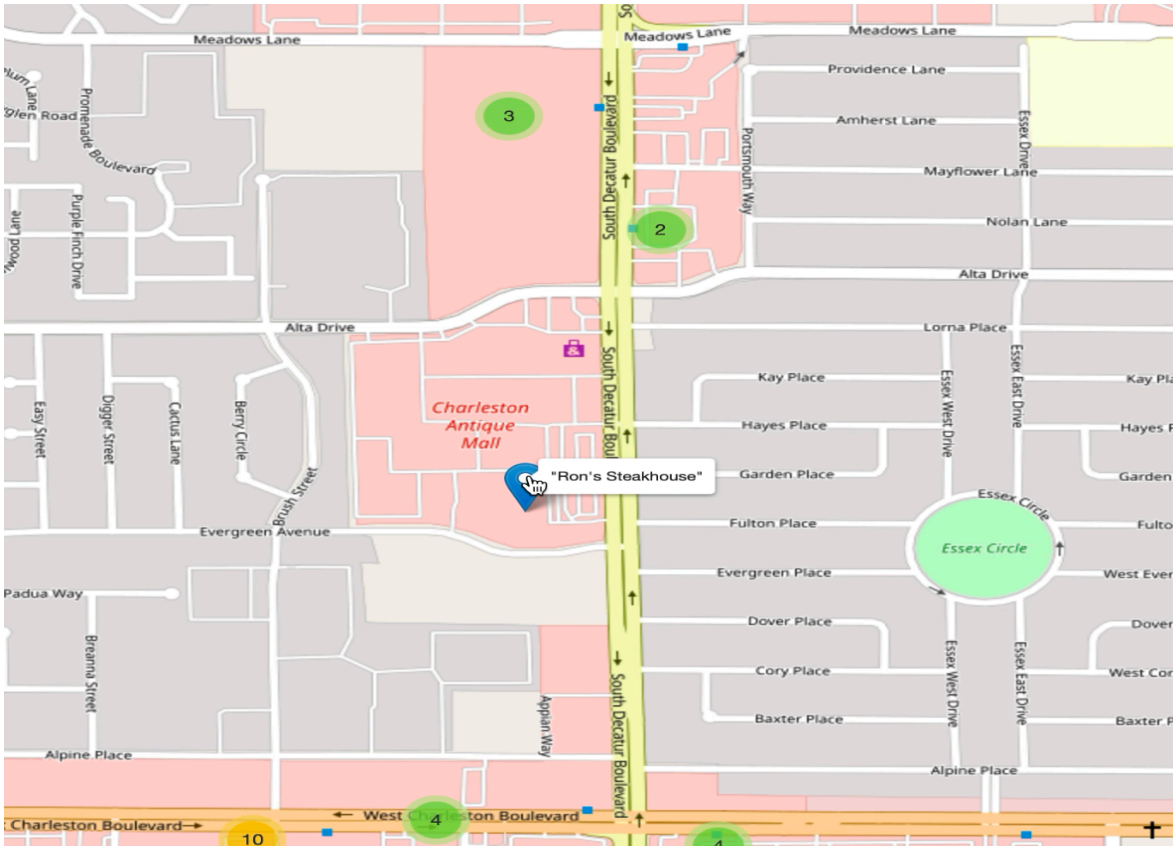
1.A shows the sample part of dataset used for the project

	review_id	business_id	stars	text
1	QgSf2JvYz-M4PU2yuJjxNQ	9Jc3W0aR9Xf2gcHI0rEXsw	1	After being scared away from Rock & Rita's, we ended...
2	gN6GARS_BRr5UX2D3WAH0w	xVEtGucSRLk5pxxN0t4i6g	5	We got recommendations for this place from my pare...
3	t4oXDPN4S4USIhBGpuSD8A	2LZGeJy8qByYKB71ML-jcw	2	We got a coupon to eat here when we checked in: $6....
4	R9w7GeMX_KZTV23gmI8Zjg	RhV7sraRUB3km-gF-tmDow	3	I have eaten here numerous times and am still amaze...
5	oncT7W70CFwzzJkQoz3T5Q	NaZVUOzqk5b-l0mlki-9Og	4	We stopped in here for lunch this afternoon. Staff was...
6	9dWoAJGcJHWscv2ZAdzkNg	tJzf6H1dkuUbL-t8bzL3dw	5	I was looking for a nice place to take the family to din...
7	pcszB9oTZE2DNylbbXIZAg	yLiaMaJFq03JxXPk4puloQ	3	I have stopped in here several times. it is always busy...
8	ulhK5rQ5FoUqIV2z_9TyWA	ZibmYdOPKLlqDM9oR6xz...	2	I live very near this place and have been curious to try...
9	bcHKbijnCQfOQfMEqs3h1g	7_F6dA9xh2lydTtr1LCtlQ	2	I live very close to this place and when I need a bit to ...
10	qOkKKBgIvpg3xzxJitCoCQ	Trbt0Ex85yvwT8DHoEFCvg	1	Ordered food to go from here. They send you an ema...

1.B Shows different restaurant clusters and slider filter



1.C Shows the location marker and restaurant name in the hover pop-up



1.D Top positive reviews for 'Ron's Steakhouse' in the summary page

Yelp Data Analysis | Map | Review Summary | Word Cloud

Restaurant: "Ron's Steakhouse"

Top positive reviews

Show 25 entries | Search:

text
One of the best flavorful filets I had in a long time , waiter was good everything was great . Took away one star for haven having one undercooked shrimp in my scampi , but overall well be back !!!
Great place! I had never eaten here and decided to try it is a hidden gem ! Great steak Service was wonderful Maria was our waitress and she is an asset to your restaurant !! I will recommend it to everyone !
We had an outstanding dinner with outstanding service from Gheorghe ask for him by name and u will not be disappointed by him or the food we will return. Thanks for a wonderful evening
Excellent food, friendly staff, and great service. We go here usually when anyone has a birthday in the family. Very happy every time.
We are sorry we waited so long to visit you. We are indeed impressed. Great bread basket, wonderful oso buco ,big tasty ,well prepared. Great salmon and dessert well price . Good service nice and quiet . We be back soon.
My family ate there for Thanksgiving and it was pretty good, I had the Filet Ocsar which was good. Several of us had the Turkey Dinner and they said it was good except the dressing. I know it is not ever like homemade for Turkey but not bad. Service was very good as well and we had a private room which was nice.
Omg this place was delicous! Our fillets were cooked to perfection. The amount of food was perfect! I know it is a bit out of the way but it was worth the trip!
i really enjoy this place the last few times ive been here. the food is good, i do however wish the sides were bigger. the steak is good. however, what probably has to be my favorite, is there staff. i really cant speak any higher of them. there really wonderful and attentive.
This is old school Las Vegas with great food, unbelievable service and an intimate atmosphere. They prepare many dishes table side like Caesar Salad, Lobster Bisque and Banana Foster. The prices are reasonable, there are intimate booths for a romantic dinner and tables for a fun time with family and friends. As a bread eater, the bread basket with 3 different butters was awesome. There is a good wine list with knowledgeable staff.
Went for the hubs 60th birthday. Had a groupon..(paid $30 for a $50 value). We were greeted with the bread basket and compound butters. Honey butter is great! Had the salads, one ranch and one Italian. I think it has raw beets in it, excellent. Both had filet medium. Cooked perfectly, tender, topped with a pat of garlic butter. YUM! Sides of asparagus and potatoes au gratin. I would probably pass on the potatoes next time. Not bad, just not great. Too cheesy. Brought a huge dish of vanilla ice cream with a lit candle out while singing Happy Birthday. Great service and food. We WILL GO BACK. Also won $300. Great evening. Only paid an additional $17 after the groupon. Plus tip, of course. Worth every penny.
text

Showing 1 to 10 of 10 entries | Previous 1 Next

Top negative reviews

Show 25 entries | Search:

1.E Top negative reviews for ‘Ron’s Steakhouse’ in the summary page

Top negative reviews

Show 25 entries Search:

text
Granted Arizona Charlie's is a local's dive as far as casinos go, but it typically has a higher payout than most casinos. Considering the casino, I should not expect much from its "steakhouse". Based on the medium range price of the menu items, I thought maybe it was a hidden gem. It is not. I had the Porterhouse Pork Chop on the prix fixe menu. The potato soup was ok, very gravy-like and less soup like.I had a 6pm reservation, so it is not like it was the end of the day. Had to ask for the bread basket that was evident on other tables. I wanted to take my leftover soup home and when the server brought the container back, the outside had globbers of soup all over it. No attempt to wipe off the "dribbles". The meaty 1-1\2 inch chop was under-cooked and tough. The demi glace on the top tasted like powdered gravy mix. Temperature of all the hot food was tepid. Service was slow and careless.They were not that busy and the place is not all that big. The scant communication I had with the waiter was as he was walking passed the table. By the time I could catch his attention about the chop, I would been waiting 30 minutes. At that point I no longer wanted to wait for the food to be remade. I ordered the desert and coffee, which were both marginal. After I would pushed everything the to opposite side of the table and placed my credit card on the table, it took another 30 minutes to get the check and I had to ask for it too. Epic fail Ron's.
I use to Love this place but last night was the worst. Our Waitress spoke broken English do not really know if she understood half of what I was saying. She made my ceasor salad tasted like olive oil over lettuce could not eat it. Forgot my husbands salad. Next their famous bread basket came 45 min later. It was not busy she had us and 1 other table. Crab cakes were great. Stuffed mushrooms tuff. Finally he gets his prime rib but she forgets his baked potato. after an hour of sitting there filled up on bread asked for to go container we were done just wanted to get out of there. Now we wait 20 minutes for our check then went up front to wait still no check .Everything was just rediculas. I have never in my life not tipped but no tip was left as I ate bread. This use to be a really good place maybe it was just the night but I am still taste olive oil in my mouth. I do not know if I will go back. The Cafe is a better choice fast friendly service in the cafe. This place has sold I hope the bad service is not a result of the new owners. So DISAPPOINTED
Went to Ron's Steakhouse this weekend after getting a $50 off coupon in the mail. Haven't been to AZ Charlie's in years, but gave it a try. While my visit to Charlie's was decent, Ron's was not. The staff was not very friendly, our waiter was very short n rude, almost like on autopilot, and when he checked back (once) to see how our meal was and saw my wife was not eating her food.. asked why. The prime rib was off in color, extremely overcooked, and just looked gross. After expressing this he insisted to offer a replacement. He took it back to the kitchen and returned with the SAME piece of meat.. All they did was re-dip it, and seared the edge where her previous two bites were. Absolutely the same piece of meat. How awful. To be honest it was not worth it even with the $50 off, and we left a side tip for the bread guy (who was also the busser) as he was very inviting, friendly, casual, fun, and got us extra bread with our to-go boxes. All while we never saw our waiter again. Doubt we will be returning. Steaks were average at best.
Meh truly sums it up. High point was the excellent bread basket. Cream of mushroom soup was decent. Stuffed mushrooms were odd (topped with what seemed to be squares of packaged, pre-sliced mozzarella) but acceptable. Chicken piccata's lemon-caper sauce had a scarily gelatinous consistency--not unlike that of bad lemon chicken at a low-end Chinese restaurant. And, finally, the true low point: Caesar salad prepared tableside. What is the point of preparing the salad tableside if, when the customer states her preference for light dressing, the server snottily declares that there is only one way to dress the salad? And, when the customer clarifies that she does not like her lettuce swimming in dressing, the server states emphatically, "We do not do that"? In the end, the Caesar was drenched in dressing, so that even if the romaine had been fresh and good, the salad was, for my purposes, ruined. The only upside of the overly dressed salad was that the pre-cut, obviously bagged romaine did not seem worth eating in the first place. Beyond the food, the service was fairly attentive, but that attention was overshadowed by the fact that the server seemed to be sneering under his saccharine veneer. I would have liked to have pointed out to him that if you can not, when asked, keep from drowning the Caesar, you have no right to cop an attitude over someone's ordering ketchup with his steak. (The last sentence should NOT be confused with a defense of my husband's use of ketchup on steak.) Finally, the atmosphere was quiet and somewhat austere (attractive but bare table tops, for example), with overly bright lighting, as if incandescent bulbs had been replaced with LED bulbs of the same wattage.
We had a poor experience here both poor food and poor service. The ribeye was edible and that is the best thing I can say. Tasted like it came out of the freezer and thrown on the grill. The server was the nicest guy you could ever meet but he was literally running around trying to keep up with all his tables. It was a bad start when the salads arrived before the cocktails. I definitely give this place a thumbs down!
My girlfriend took me here yesterday for my birthday. The service was the slowest I had ever experienced at a restaurant. Meal took over 2 hours. My porterhouse sat under a heat lamp for an hour, then microwaved. It was the worst steak I have ever had at a steak house. I was not happy and the server could tell, but I bit my tongue. To make matters worse the server was trying to hustle a big tip at the end of this shitty meal.
I think the place is cute but was not happy at all with the meal or service. We all ordered the same thing the prime rib Ron's cut well done with a baked potato. None of them were well done they were red and bloody these were the fattiest pieces of prime rib I have ever seen, I filled my bread plate with fat. I could only eat half of my baked potato because it was also [illegible] will never return for a meal over $100 and nothing was cooked right then the service the bus boy was great but our waiter sucked

1.F Word cloud of the most frequently used words in the reviews of a restaurant

1.G Shows default LDA method provided by R.

LDA(x, k, method = "VEM", control = NULL, model = NULL, ...)

- x: is an object of class "DocumentTermMatrix" with term-frequency weighting or an object coercible to a "simple_triplet_matrix" with integer entries.
- k: is the number of topics.
- Method: The method to be used for fitting; currently method = "VEM" or method= "Gibbs" are supported.
- Control: A named list of the control parameters for estimation or an object of class "LDAcontrol"
- Model: Object of class "LDA" for initialization.
- … Optional arguments. For method = "Gibbs" an additional argument seedwords can be specified as a matrix or an object of class "simple_triplet_matrix"; the default is NULL.

1.H Shows the Gibbs sampling method in LDA used for this project

LDA(dtm4, k=4, method="Gibbs",seedwords=seedwords,
control=list(alpha=60,iter=1000,seed=12345, thin=50))

control : Tuning parameters provided are

- alpha, which is the basic idea behind the parameters for the Dirichlet distribution. The higher the value the more likely each document is to contain a mixture of most of the topics instead of any single topic.
- seed: For reproducibility a random seed can be set which is used in the external code.
- iter: The number of times the process to be repeated to allow the algorithm to converge.
- thin: These parameters control how many Gibbs sampling draws are made. thin controls the number of iterations omitted during the training. This serves to prevent correlations between samples during the iteration.
- burnin: The number of initial iterations to be discarded, as the starting point of Gibbs sampling is chosen randomly, thus it makes sense to discard the first few iterations. But in our case providing seedwords provides a starting point to the algorithm, hence the burnin was kept to be its default value i.e. 0.

1.I Sample of seed words used for guided LDA

Topic1 (Food)	Topic2 (Service)	Topic3 (Ambience)	Topic4 (Recommended)
Food, Rice, Chicken, Dish, Menu, Noodles, Delicious	Waiter, Queue, Service, Friendly, Staff, Minutes, Wait	Place, Décor, Experience, Ambience, Light, Romantic, Spot	Recommend, Worth, Back, Come, Price

1.J Topic assignment to each sentence of the review along with the sentiment scores

	Review_Rating	sentence	document	Food	Service	Ambience	Recommended	ave_sentiment
1	3	service was excellent!!!	1	NA	1	NA	NA	0.771339960
2	3	jalapeño chicken tacos were great....steam burrito was g...	2	NA	NA	1	NA	0.279742190
3	5	my wife and i went here for the first time for lunch on w...	3	NA	1	NA	NA	0.720269047
4	5	it was by far the best mexican food we've ever had.	4	1	NA	NA	NA	-0.003316625
5	5	i had the jalapeno tacos.	5	1	NA	1	NA	0.000000000
6	5	while she had the chicken fajitas.	6	1	NA	NA	NA	0.013063945
7	5	we were seated right away, and the service was great.	7	NA	1	NA	NA	0.584388912
8	5	i would never visit vegas again without making this a re...	8	NA	NA	NA	1	-0.486994952
9	5	the chips and salsa were great as well.	9	1	NA	NA	NA	0.848881691