BIA 660
FINAL PROJECT REPORT

Team 5 : Abhinav Garg, Jian Hui Mai, Nishitha Dodda, Sonali Chavan
Professor Liu

Table of Contents

Section 1: Motivation and Research Question

Motivation/Introduction

We as customers have a plethora of options when we are on the lookout for a new place to eat. What is the best way to find a great restaurant? Ask someone who's been there, of course. If you don't have someone to personally ask, then you can always turn to online reviews. Customers take many factors into consideration when deciding where to eat. It's not just about how great the food tastes but how good the service is, how polite the employees are, and how well maintained the facilities are. The truth is, consumers are trusting advertising less and turning to reviews to find out what dining at a restaurant is really like. They want to know what to expect when trying a new restaurant and who better to tell them than a previous customer. The more individuals hear about your restaurant, the more inclined they will be to dine there. It is known that people are inclined to turn to customer reviews first rather than to decide where to eat. The huge quantity of data in text generated every day in the form of reviews, has no value unless processed. We propose to arrive at a data set that will give a real time experience as to how to deal with this textual data and apply data mining techniques for the effective analysis of the same.

Background and Related Work

Customer satisfaction is an essential concern in the field of marketing and research in terms of consumer behavior. In today's date, there are various techniques that have been utilized to evaluate sentiment underneath the words or expressions. Some of the most used ML algorithms used in NLP fields are Naive Bayes (NB), Maximum Entropy (ME), Support Vector Machine (SVM) and unsupervised learning. Before the rapid development of neural network based methods most recently, Linear SVMs often give the best performance in NLP. It has always been a popular approach to go for the traditional sentiment scores for analyzing the Restaurant's performance. It can be done using just the ratings, what information does that give that can help them improve their business. One approach is to find the aspects/topics for a restaurant. What is the restaurant most talked for or reviewed for? This information can drive the decision making and aid in taking the right path to improvements. This can be achieved using unsupervised machine learning algorithms like LDA and KMeans clustering methods.

Objectives and Research Questions

We hereby posit two research questions by leveraging a dataset about restaurant reviews scraped from Yelp and Google.

Research Question 1: What are the highest rated restaurants in Hoboken?
Research Question 2: What makes each restaurant special? I.e. What characteristics make each restaurant stand out?

People believe each other when it comes to reviews. It gives them an idea of what to expect at a restaurant. We collected reviews from Yelp and Google of popular restaurants in the Hoboken area which we analyzed using various methods. For our first research question, we created graphs based on the ratings given by reviewers. In order to achieve our goals stated in the

research question two, we performed topic modeling on the reviews, used the output from the topic models to perform sentiment analysis and create visual representations of the positive and neutral topic sentiments for all restaurants.

Section 2: Methodology

Brief Description of Methodology

In our project, we leveraged publicly available restaurant reviews from Yelp and Google Reviews to perform topic modeling and sentiment analysis. Each team member scraped data from Yelp and Google Reviews of all the restaurant reviews in the Hoboken area. Since there are many restaurants in the Hoboken area, we created a list of restaurants we will scrape from. We performed preprocessing including tokenization to remove stopwords, lemmatization, any punctuations, special characters and empty lines. Feature extraction was performed on the preprocessed reviews using both CountVectorizer and Term Frequency Inverse Document Frequency (TF-IDF). The algorithm used was determined based on accuracy. These reviews were then trained with the Latent Dirichlet Allocation (LDA) and KMeans algorithms to perform topic modeling. Since our dataset includes only unlabeled reviews, we are limited to only unsupervised training algorithms. The prediction from both our LDA and KMeans model will then be used to perform unsupervised sentiment analysis. A graphical representation of all the positive and neutral sentiments were created for each restaurant.

Algorithms Used

- KMeans
    - It is an unsupervised machine learning algorithm that uses clusters to separate topics. It contains a cluster center which is the mean of all points that belong to the cluster.
- Latent Dirichlet Allocation (LDA)
    - It is an unsupervised machine learning algorithm that uses linear decision boundaries to separate topics. It tries to find topics that a document belongs to based on the words in it.
- Vader Sentiment Analysis
    - VADER sentiment analysis relies on a dictionary that maps lexical features to emotion intensities known as sentiment scores. The sentiment score of a text can be obtained by summing up the intensity of each word in the text.

Input/Output of Each Algorithm

Input for Each Topic/Clustering Algorithm

The input for KMeans and LDA are the same which is the preprocessed dataset with all the reviews. This dataset is created by combining all the preprocessed datasets from both Google Reviews and Yelp.

Output for Each Topic/Clustering Algorithm

The outputs for both KMeans and LDA are different. For LDA, a list of top ten topics is returned. To find the top ten topics, we first located the highest value created by transforming the vectorized reviews of a restaurant. The column containing the highest value will be the top ten topics returned by the LDA model. For KMeans, a list of words for the cluster centroids was returned. To predict the aspects of each restaurant, we transformed the vectorized reviews

using TF-IDF for each restaurant. The result of the prediction is a list of aspects for each restaurant (a list of words).

Input for Vader Sentiment Analysis

The input for Vader Sentiment Analysis will be the topics output from the KMeans and LDA models. The size of the input will be different since KMeans outputs fifteen topics while LDA only outputs ten.

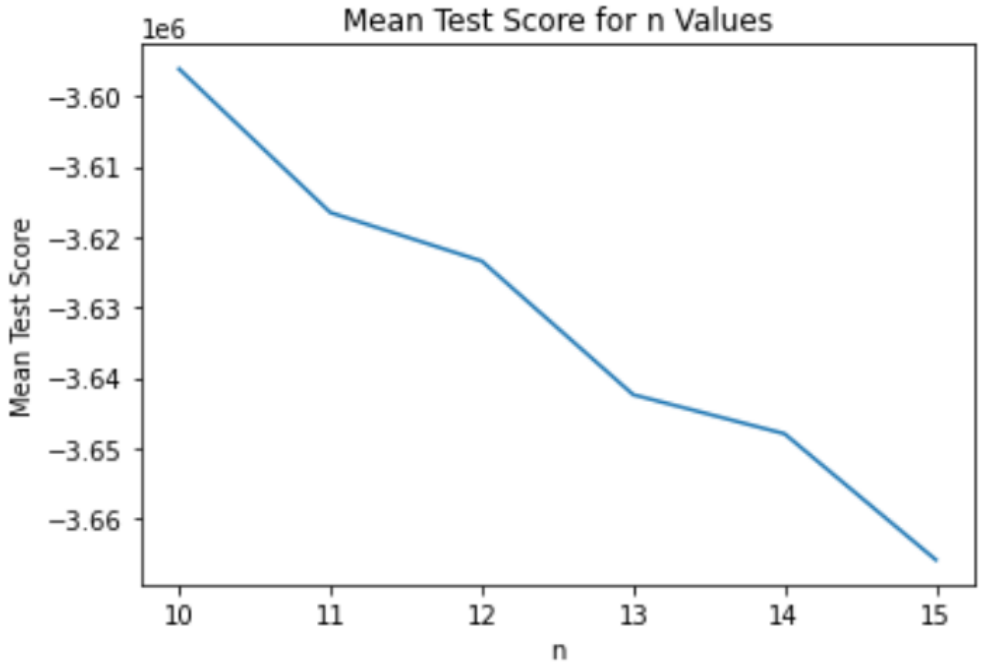
Output for Vader Sentiment Analysis

The output for Vader Sentiment Analysis will be positive and neutral words. We will use this output to create a visual representation of all positive and neutral topics about a specific restaurant.

Performance Metrics

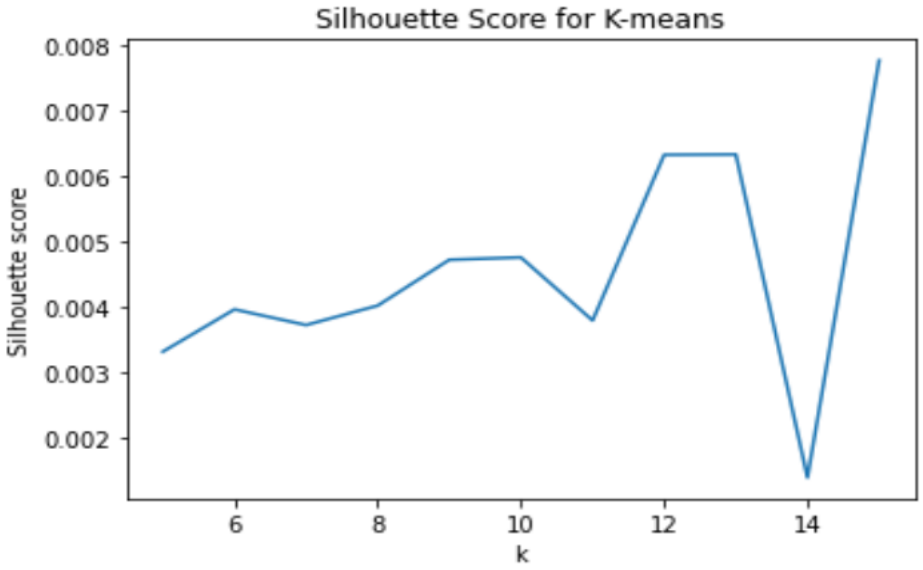
Latent Dirichlet Allocation (LDA)

In our LDA algorithm, we used both the Log Likelihood Score and the Perplexity to determine the best number of components/topics for our dataset. The Perplexity refers to the measure of how well the model predicts a sample. In other words, it is the measure of how surprised the model is when seen new data. Log likelihood refers to the plausibility model parameters given the data. We want Perplexity to be as low as possible and log likelihood to be as high as possible. With this in mind, the number of components is determined by the lowest Mean Test Score. GridSearchCV was used to search the n_components range of 10 to 15. It was determined that 10 n_components using CountVectorizer (for feature extraction) performed the best results.



KMeans

In our KMeans algorithm, we determined the best K through the Silhouette Score. The Silhouette Score is calculated using the mean of the intra-cluster distance and nearest cluster distance. The mean intra-cluster distance refers to the distance between all points to the same cluster while the mean nearest cluster distance refers to the distance of all points to the next nearest cluster. The best K will be determined by having the lowest mean intra-cluster and lowest mean nearest cluster distance. A range of five to fifteen clusters was searched and the corresponding Silhouette Score was graphed. It was found that fifteen clusters with TF-IDF (for feature extraction) performed the best which we used to train our KMeans model.

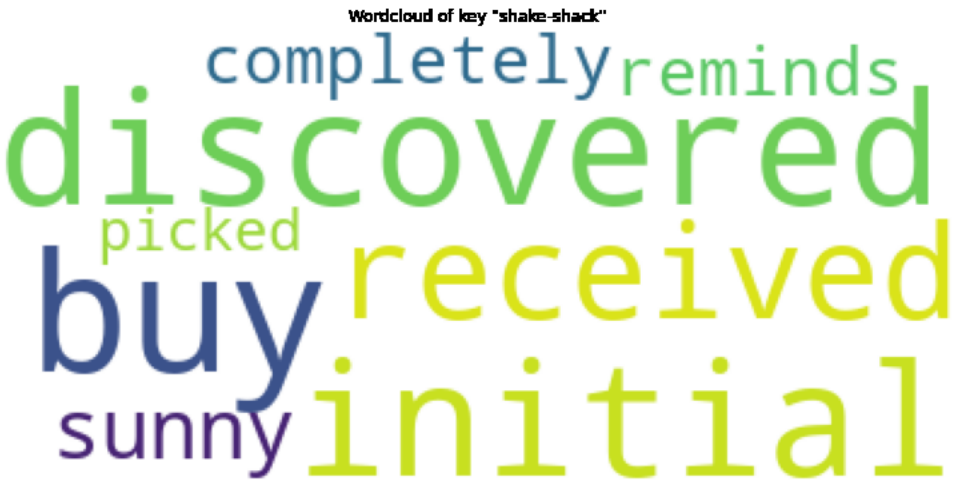


Comparison of Performance

Given the differences in our performance metrics, it is hard to directly compare the algorithms. For example, we cannot use the Silhouette Score for LDA because LDA does not create clusters but instead utilizes linear decision boundaries. With this in mind, we will use the output of both the LDA and KMeans model to perform unsupervised sentiment analysis. The outputs of these algorithms are passed into the Vader Sentiment Analysis algorithm to perform unsupervised sentiment analysis and the output of this algorithm will be compared to determine the best topic model algorithm. From our analysis, we can conclude that KMeans is better with clustering and grouping the restaurant reviews. We believe that LDA is predicting topics that are irrelevant or appear rarely in the reviews. For example, LDA outputs for “Green Pear Cafe” are Thursday and sort but those two words appear once and twice only in the reviews respectively. Additionally, some of the words like stay, suggest, class, cool, stay, and Thursday do not seem to relate to a cafe. Compared to the “Green Pear Cafe” output from KMeans, we see words like iced, breakfast, coffee, latte, pastry, tea, iced, and cup which is more in line with a cafe. LDA’s inaccurate predictions occurs in many of the other outputs for different restaurants also. In Shake Shack, the outputs are completely, reminds, discovered, picked, buy, received, sunny and initial. These words seem to have nothing to do with a restaurant. In comparison, KMeans

outputs love, service, food, good, delicious, service, nice, bowl and great which is more aligned with what you expect from a restaurant. An example of the Shake Shack outputs are below.

LDA



KMeans

Section 3: Data Collected

- Scraped Yelp Reviews
  File(s) location: “/dataset/yelp/reviews”
- Scraped Google Reviews
  File(s) location - “/dataset/google/reviews”
- Preprocessed Reviews for Each Restaurant
  File(s) location - “/dataset/yelp/processed_reviews” &&
  “/dataset/yelp/processed_reviews”
- Dataset with all Preprocessed Restaurants
  File location - “/dataset/all_reviews_processed.csv”

Sample Data

Sample Restaurants


```
SAMPLE_RESTAURANTS=[{'alias': 'fiore-deli-of-hoboken-hoboken', 'name': 'Fiore Deli of Hoboken', 'review_count': 598, 'rating': 4.5, 'transactions': 'delivery', 'location':
'414 Adams St Hoboken, NJ 07030', 'cuisine': 'delis italian'}, {'alias': 'the-cuban-restaurant-and-bar-hoboken-2', 'name': 'The Cuban Restaurant and Bar', 'review_count':
1060, 'rating': 4.0, 'transactions': 'pickup delivery', 'location': '333 Washington St Hoboken, NJ 07030', 'cuisine': 'cuban bars latin'}, {'alias':
'la-isla-restaurant-hoboken', 'name': 'La Isla Restaurant', 'review_count': 908, 'rating': 4.0, 'transactions': 'pickup delivery', 'location': '104 Washington St Hoboken, NJ
07030', 'cuisine': 'cuban'}, {'alias': 'karma-kafe-hoboken', 'name': 'Karma Kafe', 'review_count': 703, 'rating': 4.0, 'transactions': 'pickup delivery', 'location': '505
Washington St Hoboken, NJ 07030', 'cuisine': 'indpak'}, {'alias': 'vitos-italian-deli-hoboken', 'name': "Vito's Italian Deli", 'review_count': 321, 'rating': 4.5,
'transactions': 'pickup delivery', 'location': '806 Washington St Hoboken, NJ 07030', 'cuisine': 'delis'}, {'alias': 'mamouns-falafel-hoboken-hoboken-2', 'name': "Mamoun's
Falafel - Hoboken", 'review_count': 796, 'rating': 4.0, 'transactions': 'pickup delivery', 'location': '300 Washington St Hoboken, NJ 07030', 'cuisine': 'mideastern falafel
sandwiches'}, {'alias': 'amandas-restaurant-hoboken-2', 'name': "Amanda's Restaurant", 'review_count': 602, 'rating': 4.0, 'transactions': 'pickup delivery', 'location': '908
Washington St Hoboken, NJ 07030', 'cuisine': 'breakfast_brunch newamerican'}, {'alias': 'm-and-p-biancamano-hoboken', 'name': 'M & P Biancamano', 'review_count': 208,
'rating': 4.5, 'transactions': '', 'location': '1116 Washington St Hoboken, NJ 07030', 'cuisine': 'grocery delis italian'}, {'alias': 'pilsener-haus-and-biergarten-hoboken',
'name': 'Pilsener Haus & Biergarten', 'review_count': 814, 'rating': 4.0, 'transactions': 'pickup delivery', 'location': '1422 Grand St Hoboken, NJ 07030', 'cuisine': 'german
pubs'}, {'alias': 'anthony-davids-hoboken', 'name': "Anthony David's", 'review_count': 768, 'rating': 4.0, 'transactions': 'pickup delivery', 'location': '953 Bloomfield St
Hoboken, NJ 07030', 'cuisine': 'breakfast_brunch newamerican italian'}, {'alias': 'bwè-kafe-hoboken', 'name': 'Bwè Kafe', 'review_count': 282, 'rating': 4.5, 'transactions':
'delivery', 'location': '1002 Washington St Hoboken, NJ 07030', 'cuisine': 'coffee'}, {'alias': 'zacks-oak-bar-and-restaurant-hoboken', 'name': "Zack's Oak Bar & Restaurant",
'review_count': 549, 'rating': 4.0, 'transactions': 'restaurant_reservation delivery', 'location': '232 Willow Ave Hoboken, NJ 07030', 'cuisine': 'tradamerican bars'},
{'alias': 'grand-vin-hoboken', 'name': 'Grand Vin', 'review_count': 462, 'rating': 4.0, 'transactions': 'pickup restaurant_reservation delivery', 'location': '500 Grand St
Hoboken, NJ 07030', 'cuisine': 'wine_bars italian cocktailbars'}, {'alias': 'ali-baba-hoboken-2', 'name': 'Ali Baba', 'review_count': 317, 'rating': 4.0, 'transactions':
'pickup delivery', 'location': '912 Washington St Hoboken, NJ 07030', 'cuisine': 'mediterranean mideastern vegetarian'}, {'alias': 'choc-o-pain-hoboken-9', 'name': 'Choc O
Pain', 'review_count': 327, 'rating': 4.0, 'transactions': 'pickup delivery', 'location': '157 1st St Hoboken, NJ 07030', 'cuisine': 'bakeries french cookingclasses'},
{'alias': 'benny-tudinos-pizzeria-hoboken', 'name': "Benny Tudino's Pizzeria", 'review_count': 692, 'rating': 3.5, 'transactions': 'pickup delivery', 'location': '622
Washington St Hoboken, NJ 07030', 'cuisine': 'pizza'}, {'alias': 'elysian-cafe-hoboken', 'name': 'Elysian Cafe', 'review_count': 594, 'rating': 3.5, 'transactions': 'pickup
delivery', 'location': '1001 Washington St Hoboken, NJ 07030', 'cuisine': 'newamerican french bars'}, {'alias': 'margheritas-hoboken', 'name': "Margherita's", 'review_count':
402, 'rating': 4.0, 'transactions': 'pickup delivery', 'location': '740 Washington St Ste 1 Hoboken, NJ 07030', 'cuisine': 'pizza italian seafood'}, {'alias':
'la-casa-hoboken', 'name': 'La Casa', 'review_count': 233, 'rating': 4.5, 'transactions': 'pickup delivery', 'location': '54 Newark St Hoboken, NJ 07030', 'cuisine': 'cuban
dominican puertorican'}, {'alias': 'los-tacos-no-1-new-york', 'name': 'Los Tacos No.1', 'review_count': 3245, 'rating': 4.5, 'transactions': 'delivery', 'location': '75 9th
Ave New York, NY 10011', 'cuisine': 'tacos'}, {'alias': 'satay-malaysian-cuisine-hoboken-2', 'name': 'Satay Malaysian Cuisine', 'review_count': 557, 'rating': 4.0,
'transactions': 'pickup delivery', 'location': '99 Washington St Hoboken, NJ 07030', 'cuisine': 'malaysian chinese asianfusion'}, {'alias': 'otto-strada-hoboken-3', 'name':
'Otto Strada', 'review_count': 360, 'rating': 4.0, 'transactions': 'pickup delivery', 'location': '743 Park Ave Hoboken, NJ 07030', 'cuisine': 'italian pizza seafood'}]
```


Sample Scraped Reviews


```
SAMPLE_SCRAPED_REVIEWS = scraped_reviews = [
  {'date': '2021-11-04', 'text': 'A very hyped up spot. Decor and ambiance was amazing, live music and dim lighting. For apps we had the tuna tar tar, calamari, meatballs and
  truffle fries.(Decent) For mains we had the halibut (very good), two different pastas (both kustokay), short rib (very good). The drinks were very good, but pricey! Service
  was also good.', 'rating': 4},
  {'date': '2021-10-14', 'text': "The ambiance of this place is unmatched. Definitely would suggest making reservations ahead of time. Had to sit in a odd but cute location by
  the front door. Had to get a glass of rosé, my fav, to start. For $14 per glass it was definitely a small pour but delicious. Had a charcuterie 6 combo board to share and
  tuna tartare. Both were amazing! Definitely needed more crostinis with the board but loved all the meats and cheeses. Also enjoyed the live music as well. Even though it was
  packed with music playing, it wasn't too loud that you couldn't hear across the table when talking with your friends. Cannot wait to come back and try the short rib and
  risotto. Perfect late night find in Hoboken and open until midnight during the week FYI :)", 'rating': 4},
  {'date': '2021-10-10', 'text': "Disappointed.. Read the reviews and thought this place had potential. It's a nice place for ambience but everything else not so good. We had
  the waiter recommend a wine, he suggested a Malbec. It was gross. I couldn't finish it. We order the stuffed shrimp appetizer- I couldn't eat it. Very fishy taste I did not
  care for it all. So decided to try a different appetizer. I was apprehensive about ordering the scallops appetizer because the shrimp was so bad but we did. It was good.
  After that there's nothing. We waited and waited for our waiter to take our entree order but he never came back. We literally sat there for what seemed like an hour waiting
  to order something else but no one came by. We actually had to go to the bar to ask and Pay our bill. I don't think we be back", 'rating': 2},
  {'date': '2021-10-06', 'text': 'My Fiancé and I had our first date here in 2018. He have come back here for our anniversaries brought friends and family there. The place
  never disappoints', 'rating': 5},
  {'date': '2021-10-05', 'text': "Grand Vin hosted us for our engagement party, and it was honestly one of the best days of our lives! I highly, highly recommend them if you
  are looking to host a shower or any kind of celebration!To start, this restaurant is our happy place. We come here about once a week ether to have dinner or enjoy a drink at
  the bar. So, when we were considering having an engagement party, we knew it would have to be at Grand Vin.Decor reminds me of Napa, like a slightly rustic (but also super
  classy) wine bar. You don't even need much decor because the string lights and greenery throughout the restaurant set the perfect, romantic vibe.The service is top-notch,
  Amanda and Mike (the managers) were communicative and so helpful organizing everything. I was able to drop off decor the day prior and then leave everything fully in their
  hands to execute. We spent about the first 45 mins or so of the party hanging out in the bar area with passed appetizers, then were all seated in the dining room for the
  meal. They timed all of the courses perfectly and the portions were incredibly generous!The food is insanely good; everyone kept talking about it. I dream of that sea bass
  and gnocchi all the time!We had about 45 people seated in their back dining area, but they can fit a bit more than that if you rent out the whole space. There is also a
  parking garage about a block away (a SUPER huge plus for all Hoboken venues!).", 'rating': 5},
  {'date': '2021-08-11', 'text': 'Love this wine bar!! Great service and they have a cute, super cozy vibe!! Their chicken entree with vodka sauce was really good as well as
  the pastas - both the rigatoni and capellini. Great food and drink menu! There was live music on a Friday night.', 'rating': 5},
  {'date': '2021-08-08', 'text': "Don't waste your money here! I don't understand why this place has such high reviews. The food was horrible, none of the staff is
  knowledgeable about the wine and the wine list was tiny, overpriced, and full of nothing but mediocre wines. Hoboken and your wallet both deserve better.", 'rating': 1}
]
```

Data Sources

We scraped data from Yelp using APIs and Google Reviews using Selenium and BeautifulSoup. We did not use any pre-crawled data or Kaggle datasets.

Yelp Reviews

Our data set from Yelp contains a list of 188 restaurants with a variable amount of reviews for each restaurant which accounts for a total of 38,980 reviews.
We're utilizing two separate API's for yelp data.

1. Yelp API (Yelp business search API) - to fetch all the restaurants in hoboken
2. Unwrangle API (API Link) - to fetch the reviews for each restaurant

For Google Reviews

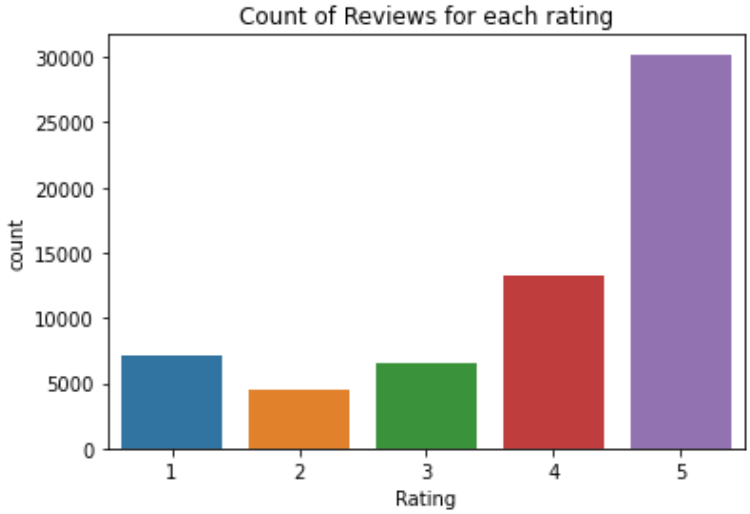
We have a variable amount of reviews for each restaurant and around 100 restaurant's data is scraped from Google's website. We are scraping Google reviews from Google Maps using Selenium and BeautifulSoup.
Our data set from Google contains a list of 185 restaurants with a variable amount of reviews for each restaurant which accounts for a total of 22,550 reviews.

Section 4: Exploratory Data Analysis (EDA)

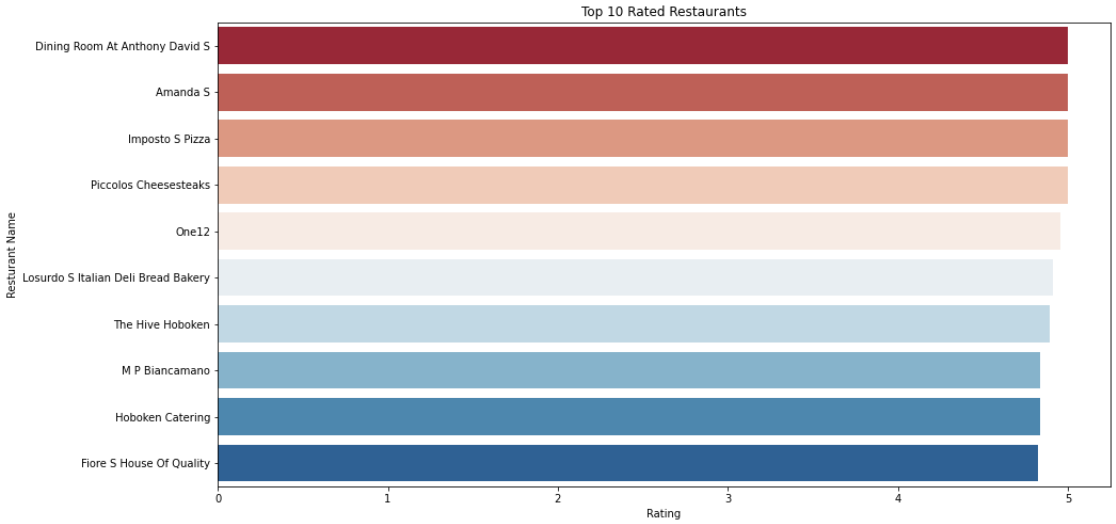
We performed ratings analysis based on the self reported ratings by reviewers. For every review written, a score/star between one and five was assigned by the reviewer. We used those ratings to graph the number of different types of rating for each restaurant.

<u>Research Question 1:</u> What are the highest rated restaurants in Hoboken?

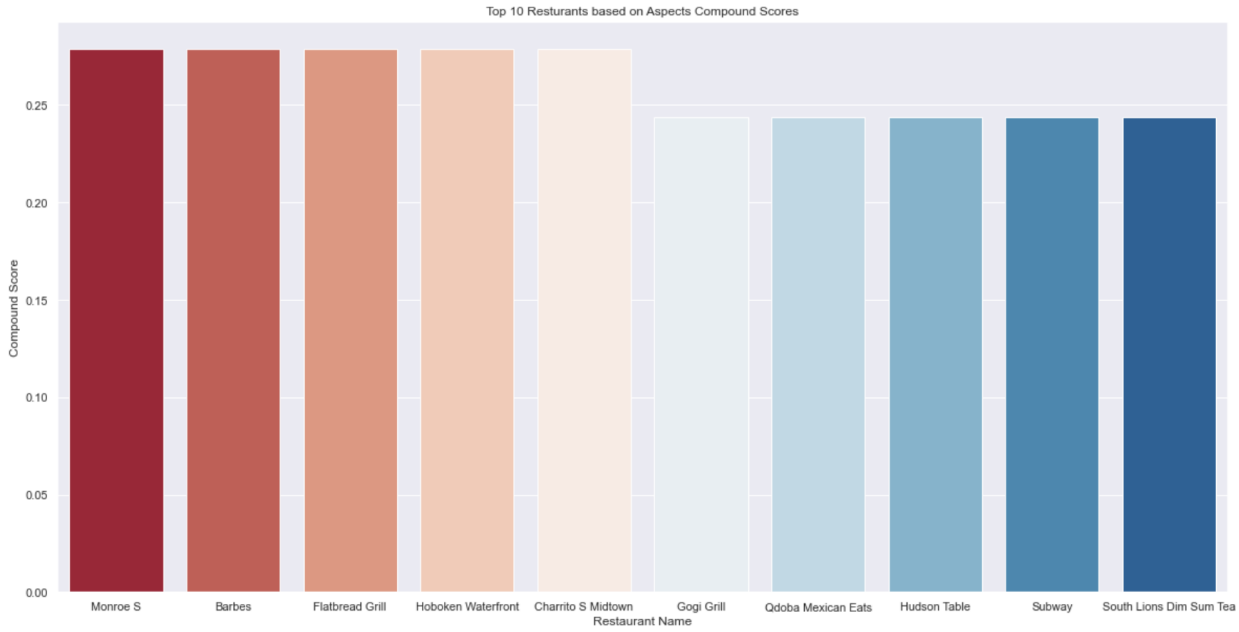
1. Count of Restaurants has been analyzed with respect to their Ratings.



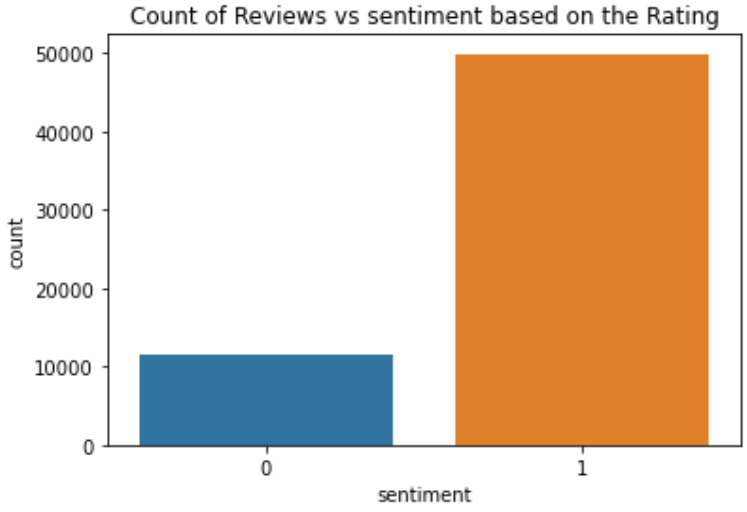
2. Top 10 Highly Rated Restaurants



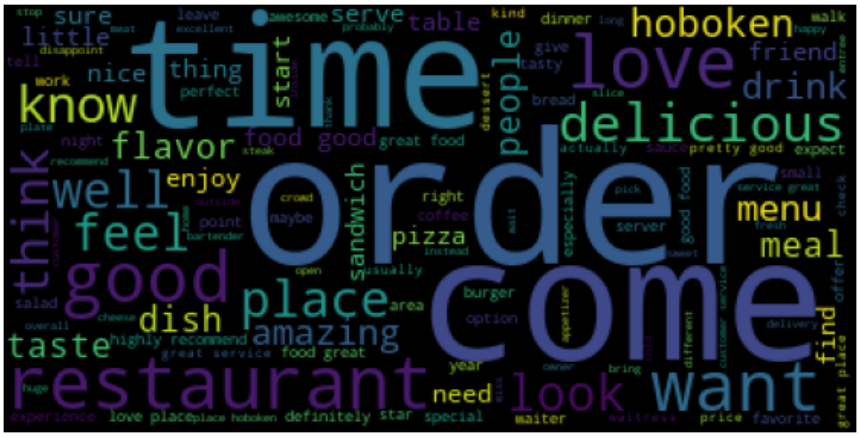
3. Top 10 Highly Restaurants Based on Aspects
   This chart is created from the outputs of the topics of KMeans. We used those topics to determine the average compound score. The compound score is a score determined by the Vader Sentiment Analysis where a higher number represents a more positive sentiment. This has been implemented in rss_kmeans.ipynb
   This connects to the research question 2 and as an extension to research question 2, the sentiments were derived on aspects and then derived holistically for each restaurant.



4. Low vs High Rated Restaurant's Count

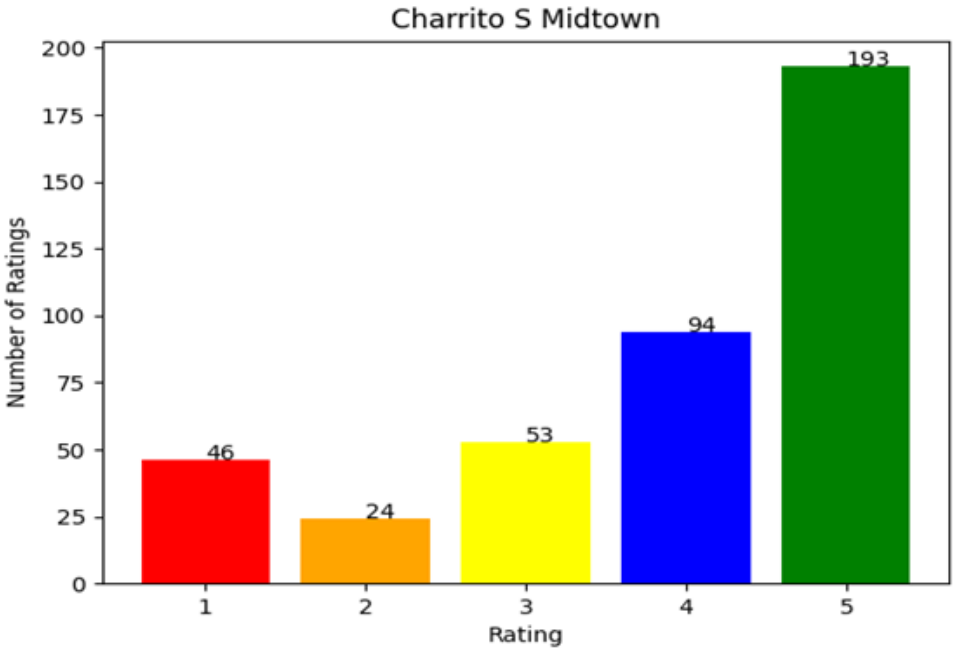


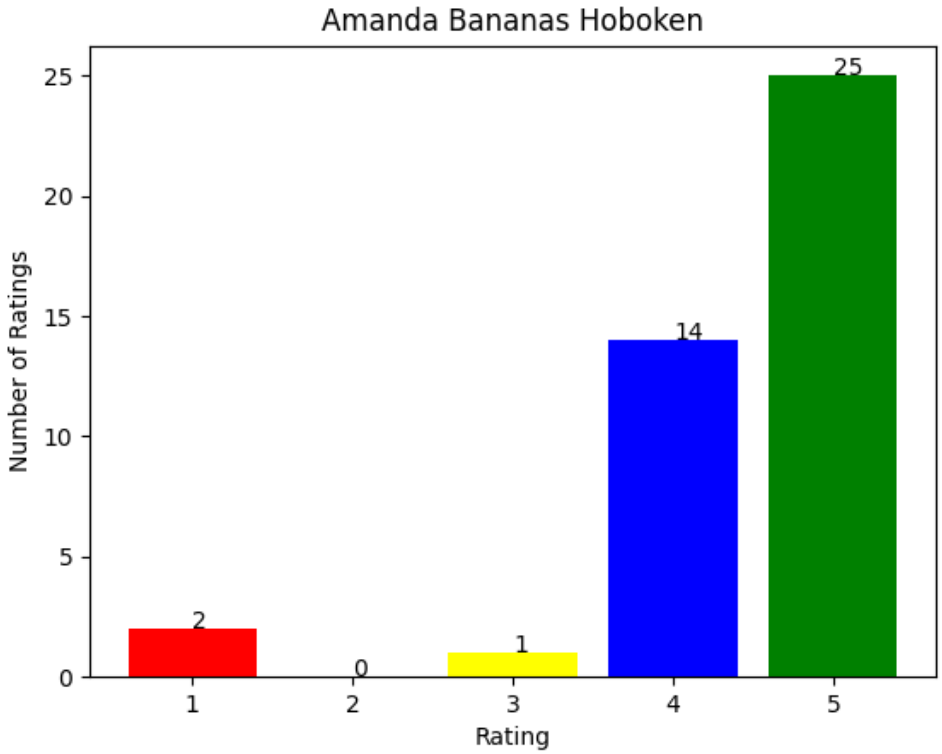
5. Most Common Words in the Reviews



6. Most Common Verbs and Adjectives in the Reviews

7. Sample of Rating Analysis



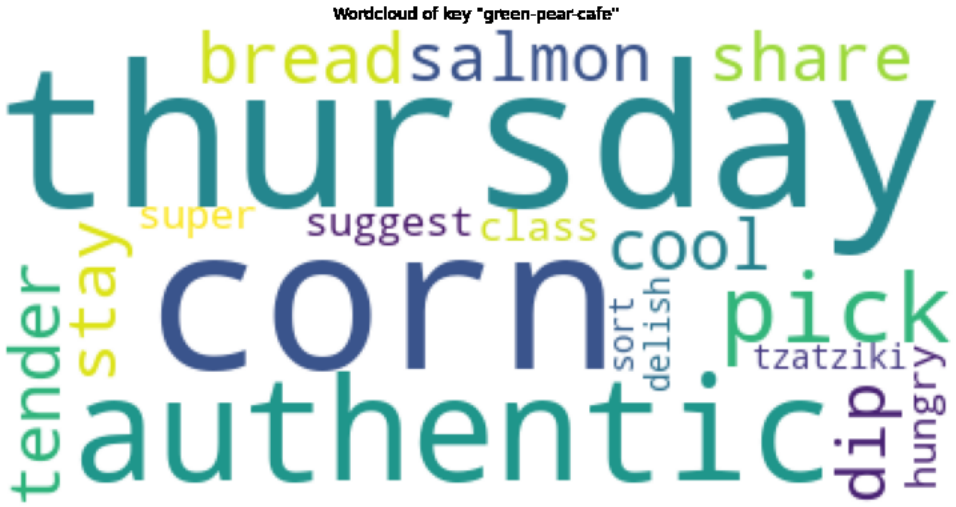


<u>Research Question 2:</u> What makes each restaurant special? I.e. What characteristics make each restaurant stand out?

These graphics were created from the output of each clustering model. (KMeans and LDA). Only the positive and neutral sentiments determined by the Vader Sentiment Analysis were included in the graphics/WordCloud below.

KMeans

LDA



Section 5: Python Scripts

Scraping Data from Yelp

For scraping data from Yelp, we've created a Python script file named "scrape_yelp.py" (located in the "script/yelp/" directory).

Scraping Data from Google

Google scraping has two separate scripts, one for fetching a list of all the restaurants, and another for fetching all the reviews for each restaurant. File names are "restaurants.py" and "reviews.py". Both of these files are stored inside the "script/google" directory.

Steps

1. We searched for all the restaurants in Hoboken City on Google Maps using Selenium and saved the link to visit each restaurant in the "restaurants.csv" file.
2. In the next step, we visited each restaurant link through Selenium and got the restaurant's details and reviews.
3. For reading reviews, we clicked on the reviews link on Google Map page through script, sorted reviews by "Newest" order, expanded reviews and saved them in individual CSV files for each restaurant. For each restaurant, we fetched on average 350-400 reviews.
4. "dataset/google/restaurants_details.csv" file contains the details of all restaurants and the "dataset/google/reviews" folder contains reviews of all restaurants.

Preprocessing of Text

We preprocessed both scraped reviews from Yelp and Google and saved them onto new CSVs. Our preprocessing includes removal of punctuations, stop words, empty lines, and special characters. Additionally, our reviews are normalized using the lemmatization technique. To process each restaurant CSV file, we created a loop that locates all files that end with .csv in a specific folder. We would then go row by row to perform preprocessing where the end results will be saved onto another CSV file appended with "_processed" to signify the processed file. The Python script to preprocess the datasets is located at "script/preprocess.py". Additionally, a combined dataset was created to include all reviews gathered from Google and Yelp. It is called all_reviews_processed.csv and only contains the date, restaurant name, rating, review and source columns. The Python file to create this combined dataset is located at "/script/combine_data.py"

Training of Models

Topic Modeling

LDA

The data containing the processed reviews for all restaurants were read and loaded into a dataframe. Since we are doing topic modeling with reviews only, we filtered the data frame to contain only restaurant reviews. Feature extraction was then applied to the reviews using a CountVectorizer. In order to find the best n_components, GridSearchCV was used to search the ranges of ten to fifteen. The best n_components were found by having the lowest Mean Test Score. It is determined that ten was the best number for n_components. Additionally a graph was created to show the lowest Mean Test Score which also came out to be ten n_components. In an effort to further improve accuracy, the feature extraction algorithm was changed to TF-IDF but it was found to have a lower accuracy than CountVectorizer. Due to this, I trained the model with CountVectorizer and ten n_components. Both the vectorizer and LDA model were saved using pickle to allow for usage later on without retraining. A sample of the topics output is shown below.

```
7-stars-pizzeria_processed.csv
['visiting', 'tot', 'limited', 'wine', 'hungry', 'matter', 'barely', 'rather', 'easily']
80-river-bar-kitchen_processed.csv
['enjoy', 'since', 'protein', 'seated', 'ca', 'said', 'perfect', 'even', 'next']
8th-street-tavern_processed.csv
['extremely', 'overcooked', 'husband', 'try', 'yes', 'love', 'full', 'sausage', 'disappointed']
acai-ya-later_processed.csv
['fall', 'undercooked', 'syrup', 'shocked', 'absolutely', 'code', 'least', 'honey', 'smaller']
acme-markets_processed.csv
['immediately', '4th', 'tabouleh', 'drinking', 'body', 'dancing', 'bonito', 'man', 'frank']
alessio-s-cafe-gelato-pizza_processed.csv
['ok', 'might', 'black', 'kept', 'joint', 'ask', 'world', 'snack', 'salt']
ali-baba-restaurant_processed.csv
['twice', 'note', 'expecting', 'case', 'word', 'indoors', 'combination', 'art', 'churrasco']
amanda-bananas_processed.csv
['crowd', 'portion', 'come', 'min', 'old', 'experience', 'flavor', 'croissant', 'server']
amanda-s_processed.csv
['pleaser', 'passed', 'wound', 'quote', 'chose', 'applied', 'overwhelmingly', '1pm', 'feed']
```

The training script is located in “script/lda/lda.ipynb”. The script to generate the topics above is located in “script/lda/lda_get_topics.ipynb”

KMeans

The data containing the processed reviews for all restaurants were read and loaded into a dataframe. Since we are doing topic modeling with reviews only, we filtered the data frame to contain only restaurant reviews. Features were extracted using the fit_transform method on the TF-IDF vectorizer. Determination of K was done using the silhouette score. K Means model was then run on the 100% data using best K which turned out to be fifteen. The clusters were then grouped and individually saved in csv format. Then, the clusters were intuitively named as below based on the cluster centroids (list of words returned). The training file and topics are generated from “script/kmeans/kmeans_get_topics.ipynb”.

	precision	recall	f1-score	support
0	0.00	0.00	0.00	5
1	0.00	0.00	0.00	6
2	1.00	0.86	0.92	7
3	0.58	0.88	0.70	8
4	0.78	1.00	0.88	7
5	0.88	1.00	0.93	7
6	0.62	0.83	0.71	6
7	0.64	0.78	0.70	9
8	1.00	0.43	0.60	7
9	0.38	0.71	0.50	7
10	0.88	1.00	0.93	7
11	0.00	0.00	0.00	7
12	0.41	0.88	0.56	8
13	0.00	0.00	0.00	7
14	0.50	0.80	0.62	5
accuracy			0.63	103
macro avg	0.51	0.61	0.54	103
weighted avg	0.53	0.63	0.55	103

Cluster	Aspect
0	BAR
1	TASTE
2	RUDE
3	VIBE
4	SUSHI
5	BRUNCH
6	COFFEE
7	SERVICE
8	THAI
9	ITALIAN
10	PIZZA
11	BAD SERVICE
12	FRIENDLY STAFF
13	DELICIOUS
14	RAMEN_BOWL

The classification report was extracted by using only 5,090 reviews as the train set (Due to computational issues) and 100 as the test set by manually labeling. This is something that didn’t work very well with the accuracy scores. We then fetched the topics of each restaurant by using the model created using 100% data as the train set and predicting the aspects of each restaurant. The sample output is shown below. The dataset used for training and testing are located under “script/kmeans/dataset”. The Python script to create the report above and to train is located at “script/kmeans/kmeans_report.ipynb”.

```
C:/Sem 2/web mining/git_latest/BIA-660-main/dataset/yelp/processed_reviews/antique-bar-and-bakery-hoboken-2_processed.csv
['amazing', 'food', 'service', 'great', 'place', 'staff', 'friendly', 'love', 'recommend', 'definitely']
C:/Sem 2/web mining/git_latest/BIA-660-main/dataset/yelp/processed_reviews/apulia-hoboken_processed.csv
['nice', 'place', 'food', 'good', 'great', 'staff', 'service', 'atmosphere', 'really', 'friendly']
C:/Sem 2/web mining/git_latest/BIA-660-main/dataset/yelp/processed_reviews/arthurs-steaks-hoboken_processed.csv
['food', 'time', 'drink', 'bar', 'order', 'table', 'place', 'good', 'service', 'came']
C:/Sem 2/web mining/git_latest/BIA-660-main/dataset/yelp/processed_reviews/augustinos-hoboken_processed.csv
['food', 'time', 'drink', 'bar', 'order', 'table', 'place', 'good', 'service', 'came']
C:/Sem 2/web mining/git_latest/BIA-660-main/dataset/yelp/processed_reviews/ayame-hibachi-and-sushi-hoboken_processed.csv
['sushi', 'roll', 'hoboken', 'place', 'fresh', 'good', 'great', 'best', 'fish', 'service']
C:/Sem 2/web mining/git_latest/BIA-660-main/dataset/yelp/processed_reviews/baking-mama-hoboken_processed.csv
['cake', 'cooky', 'cream', 'ice', 'cupcake', 'chocolate', 'bakery', 'sweet', 'good', 'cookie']
C:/Sem 2/web mining/git_latest/BIA-660-main/dataset/yelp/processed_reviews/bangkok-city-thai-restaurant-hoboken_processe
d.csv
['food', 'place', 'good', 'great', 'love', 'chicken', 'service', 'like', 'hoboken', 'sandwich']
C:/Sem 2/web mining/git_latest/BIA-660-main/dataset/yelp/processed_reviews/barbès-restaurant-hoboken_processed.csv
['food', 'place', 'good', 'great', 'love', 'chicken', 'service', 'like', 'hoboken', 'sandwich']
C:/Sem 2/web mining/git_latest/BIA-660-main/dataset/yelp/processed_reviews/bareburger-hoboken-hoboken_processed.csv
```

Unsupervised Sentiment Analysis using Vader Sentiment Analysis

We are using topics fetched from both LDA and KMeans model for performing sentiment analysis. We are combining topics generated from Google and Yelp reviews and passing them as an input to the VADER's SentimentIntensityAnalyzer(). This returns sentiment scores for positive, negative, neutral and compound. If a specific word is categorized as positive, it will say 1.0 for positive and 0.0 for neutral and negative. The compound score is determined by how positive and how negative a certain word is. The Python script to create this is located at "script/sentiment_analysis.ipynb"

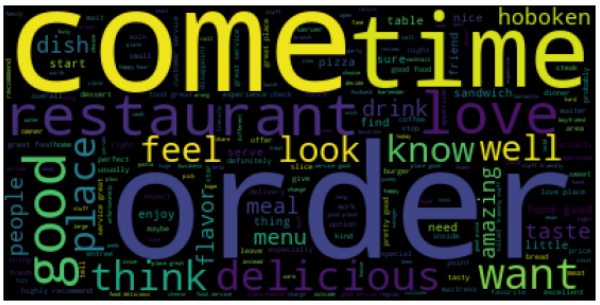
Visual Representations for Exploratory Data Analysis

Rating Graphs

Rating graphs were created for each restaurant (from both Google Reviews and Yelp) using matplotlib. The rating graph counts the amount of one star, two star, three star, four star, and five star ratings. Given the rating, an appropriate color is assigned, one star: red, two star: orange, three star: yellow, four star: blue and five star: green. The height is dependent on the amount of ratings for the star type. Additionally, a data label was placed above each graph. A sample bar graph could be found below under Section 4: Exploratory Data Analysis. The Python script to create this is located under "script/create_rating_graphs.py" .

Word Cloud Created on All Reviews

We created word clouds to visualize the most common words in the reviews and the most common verbs and adjectives in the reviews. Adjectives and verbs talk about the aspects of the restaurants as well as the actions of the customers respectively. It was done using the WordCloud package and WordCloud.generate() method. The Python file to generate WordCloud is located at "script/sentiment_and_wordcloud.ipynb".



Most common words

Most Common Verbs, Adjectives

Top 10 Highly Rated Restaurants Graph

Highly rated restaurants are determined by using group by on the restaurants and ratings column. Values are then sorted in descending order and outputs plotted using matplotlib and seaborn. The Python file to generate WordCloud is located at "script/sentiment_and_wordcloud.ipynb".

Top 10 Highly Restaurants Based on Aspects

The Vader Sentiment Intensity Analyzer was utilized to gather the compound scores of each topic of a restaurant. The average of each restaurant was calculated and gathered with the top 10 average compound scores graphed using Seaborn. The Python file to generate this graphic is located at "script/rss_kmeans.ipynb".

Count of Reviews with Preliminary Sentiment Score as per Ratings Graph

The count of reviews categorized into positive and negative using preliminary sentiment scores from the ratings column. The sentiment score of 0 and 1 are given to the review based on the rating of the restaurant. Threshold was set to 3, <3 was considered a 0 and >=3 considered a 1.

WordCloud to Represent Positive and Neutral Aspects for Restaurants

We are using WordCloud to visualize the importance of positive and neutral topics fetched from the output of Vader Sentiment Analysis. The Python file to generate this graphic is located at "script/sentiment_analysis.ipynb".

Sample word-cloud for the positive topics from one of the restaurant "shake-shack" using k-means and lda models respectively:

Wordcloud of key "shake-shack_processed.csv"

Wordcloud of key "shake-shack"

Section 6: Analysis of Experiment Results

What part of your methodology worked (or didn't work)?

Given that we trained our topic models using unsupervised learning algorithms, it is hard to determine the accuracy of the model. We devised a solution where we manually label about 100 reviews (test dataset) using the clusters that KMeans created and trained another KMeans model with 5,000 reviews only. Using the test dataset, we compared the KMean prediction to the values we labeled, it was determined that there was about a 63% accuracy which is shown below in the classification report.

Classification Report:

	precision	recall	f1-score	support
0	0.00	0.00	0.00	5
1	0.00	0.00	0.00	6
2	1.00	0.86	0.92	7
3	0.58	0.88	0.70	8
4	0.78	1.00	0.88	7
5	0.88	1.00	0.93	7
6	0.62	0.83	0.71	6
7	0.64	0.78	0.70	9
8	1.00	0.43	0.60	7
9	0.38	0.71	0.50	7
10	0.88	1.00	0.93	7
11	0.00	0.00	0.00	7
12	0.41	0.88	0.56	8
13	0.00	0.00	0.00	7
14	0.50	0.80	0.62	5
accuracy			0.63	103
macro avg	0.51	0.61	0.54	103
weighted avg	0.53	0.63	0.55	103

We wanted to train with more than 5,000 reviews and see the accuracy but due to computation limitations, we were able to. With this in mind, we cannot say for certain that our model yields a 61% accuracy since only 5,000 reviews were used and the test set only contains 100 reviews. Since we could not use the classification report, we utilized the Silhouette Score for KMeans and the Mean Test Score for LDA to determine the accuracy of our models. For KMeans, we want the highest Silhouette Score and for LDA, we want the lowest Mean Test Score.

Why did your methodology work or (didn't work)?

Our initial methodology of using semi supervised learning did not work because we were unable to scale both our testing and training datasets. Our testing dataset was limited on how much we can manually label which is time and labor intensive. Additionally, the size of our training dataset was limited to computational limitations of our computer. In regards to why the performance measures worked, it is because we were able to find the best K for KMeans and the best n_components for LDA to create the most accurate model given the parameters. We believe that using these performance measures was successful in finding the appropriate topics and aspects to a good extent. The clusters through KMeans and the aspect derived from the centroid words are in good correlation, which talks about the credibility of the model.This was

determined by skimming through the reviews in each cluster as they were saved in CSV. The reviews were mostly related to the aspect that we derived intuitively using cluster centroid words. For example, if a cluster was named “BRUNCH” based on the cluster centroid words, we observed that the reviews were all/mostly talking about brunch, bagel, sandwich etc. The methodology worked well in finding the top rated restaurants as well as aspect based sentiment scores for each restaurant.

How to improve?

Given that we were unable to use similar performance measures to compare the accuracy of our models, we should label a bigger size of the dataset and use Azure or Amazon Web Services to train our models. We will then compare the classification reports of both LDA and KMeans to determine the “more accurate” model.

How to utilize your results? What business insights can be derived from your analysis?

We can utilize our results to help restaurant owners better understand what areas/aspects they excel in and what areas/aspects their competitors excel in. For example, reviewers say that a certain restaurant has fantastic sushi, the restaurant owner could have promotions to try to lure more customers to try their sushi and potentially gain a long term customer. On the other hand, a restaurant might believe that they excel at something but reviewers say otherwise, the restaurant owner could try to improve that aspect. Restaurant owners could also use the information to try to stand out from the competition. For example, there are two pizzerias across the street from each other and the other restaurant is known for having great and fast deliveries. The other pizzeria can try to stand out by having a better dine in experience.

Section 7: Conclusion and Future Work

Conclusion

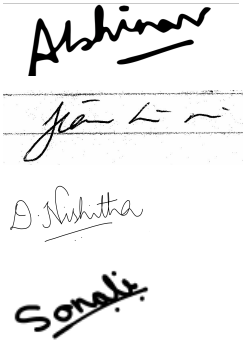
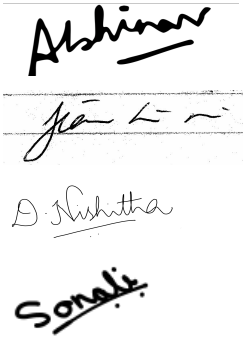
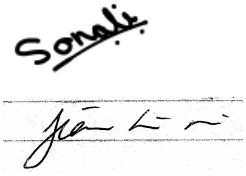
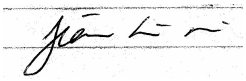
Our project answers both our research questions. As part of Exploratory Data Analysis, we were successful in finding top rated restaurants using the Ratings column in the data. Adding to that, getting to know what the aspects of Restaurants are give us important business insights. The topics/aspects were more insightful in K Means Algorithm. These insights prove to be very essential in decision making with respect to varied aspects.

In order to answer the research questions, we employed web scraping techniques such as APIs, Selenium and BeautifulSoup to procure data, which was followed by the preprocessing and initial Exploratory Data Analysis. Then, we applied the clustering algorithms LDA and K Means to determine the aspects of the restaurants. We then carried out an Aspect based Sentiment Analysis by analyzing the sentiments of the aspects acquired which we used to create word clouds. The Word Clouds helps

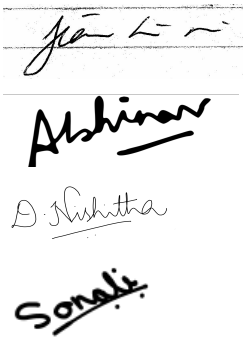
Future Work

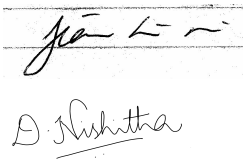
Given the limited timeframe for this project, we were not able to successfully perform semi-supervised machine learning for our topic modeling. In our current implementations, we used two different performance metrics to create the most accurate model. With this in mind, we can not say for certain if the KMeans or the LDA model is better than the other. If we were given more time, we can label about 20% of our dataset to perform semi-supervised learning. We will compare the clusters/topics generated by the KMeans and LDA algorithms to the ground truth labels (topics) that we created. The model with the higher classification accuracy will be the best model. Additionally, we cannot determine if the topics generated by our models are an accurate representation of the reviews. If we had ground truth labels, this would help in that regard also. Ground truth labels would also help us determine if the model is under fitted or over fitted and make appropriate decisions to help fix that whether that is to use a new clustering algorithm or gathering more data.

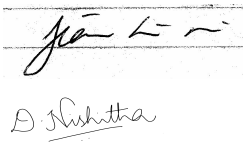
Section 8: Tasks Breakdown

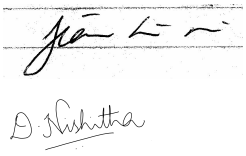
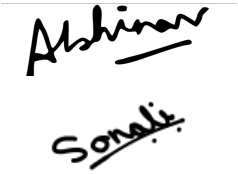
Task	Completed by	Signatures
• List of restaurant review-sources to scrape from and features to include	Abhinav, Jian, Sonali, Nishitha	
• Research of technologies and methodologies	Abhinav, Jian, Sonali, Nishitha	
• Web Scraping of Yelp	Abhinav and Nishitha	
• Web Scraping of Google Reviews	Sonali and Jian	
• Preprocessing ○ Data cleansing ○ Tokenization ○ Lemmatization ○ Stop Words Removal ○ Punctuation removal ○ Removal of empty lines and	Jian	

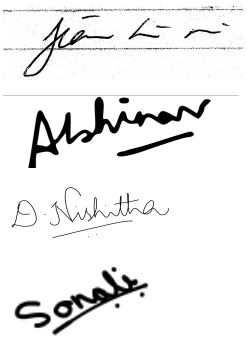
special characters		

• Exploratory Data Analysis
  ○ Bar Graph of count of restaurants and ratings
  ○ Top 10 highly rated restaurants
  ○ Count of reviews with preliminary sentiment score as per ratings
  ○ Word Clouds for highly common words, verbs and adjectives
	Jian, Abhinav, Nishitha and Sonali	

• Vectorization
  ○ Countvectorizer
  ○ Term frequency inverse document frequency (TF-IDF)
	Jian and Nishitha	

• Topic Modeling
  ○ K Means
  ○ LDA
	Jian and Nishitha	

• Test for Accuracy using Performance Metrics	Jian and Nishitha	
• Unsupervised Sentiment Analysis ○ Vader Sentiment Analysis	Abhinav and Sonali	
• Creation of Visual Representation of positive and neutral topics	Abhinav and Sonali	

• Research Paper	Jian, Abhinav, Nishitha and Sonali	
• Presentation Deck	Jian, Abhinav, Nishitha and Sonali