Yelp Data Set User Review text analysis

Introduction:

For this Project, problem selected is finding patterns in the user reviews which are useful for businesses to assess not their individual strength and weaknesses but also be able to match their current offerings with people expectations in their business category.

Problem Statement and Approach:

Problem 1: Finding out what people care most for a business category in general which leads to difference in star ratings. *Finding these attributes for a business category will be helpful for all the businesses in that category.*

Problem 2: Finding out what people like or dislike in your business based on user review and review ratings: *Finding these attributes will be helpful for a businesses to understand its weakness and strength.*

Approach Problem 1:

Data selection Using Pandas:

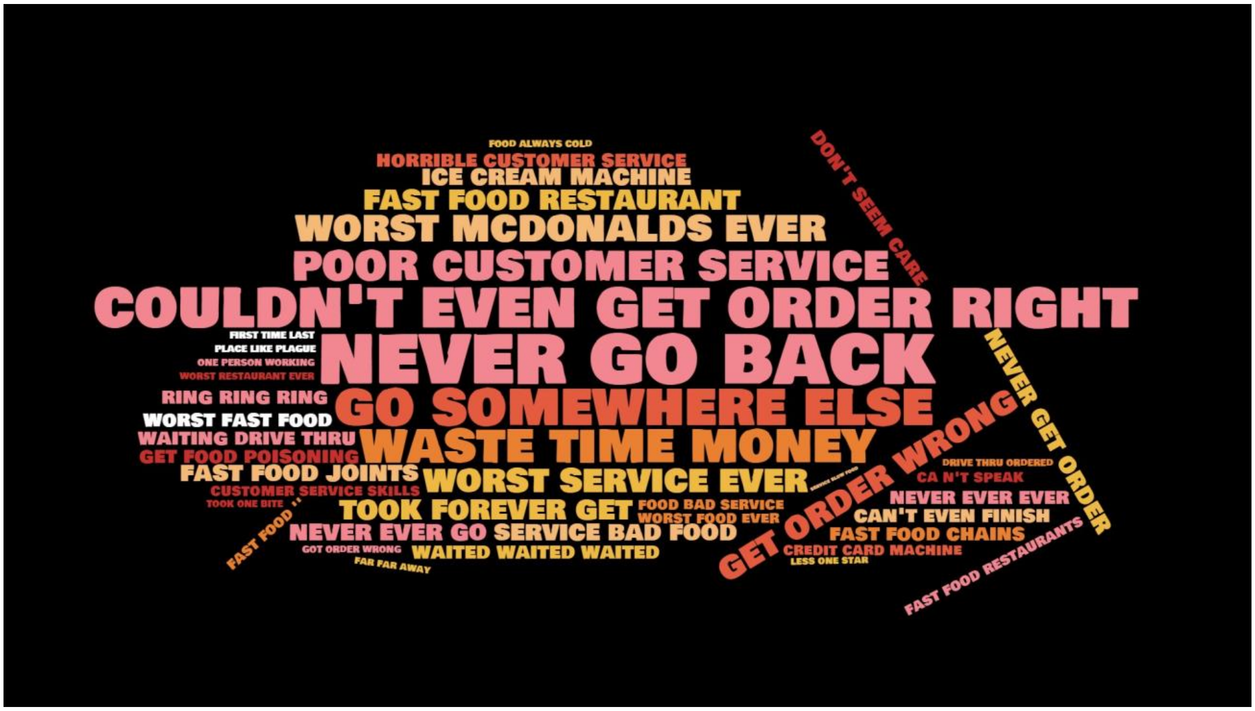
1) Divided the business data set in two groups based on their star ratings:
   a. Low Rating (<=1.5)
   b. High Rating (>=4)
2) For each group above select the data based on different business categories. Some of the businesses in one category Restaurant are like:
   [Breakfast & Brunch', 'Restaurants'], ['Cafes', 'Restaurants']
   ['Indian', 'Restaurants'], ['Restaurants', 'Mediterranean', 'Turkish']
3) Join filtered data by category with review dataset. So data now contains:
   a. All restraint businesses whose rating is >= 4
   b. All restraint businesses whose rating is <=1.5
4) Select the review text from this joined data so we get data like:
   a. All review texts for restaurant businesses whose rating is >= 4
   b. All review texts for restaurant businesses whose rating is <= 1.5

Finding patterns in review text using NLTK (Natural language tool kit):

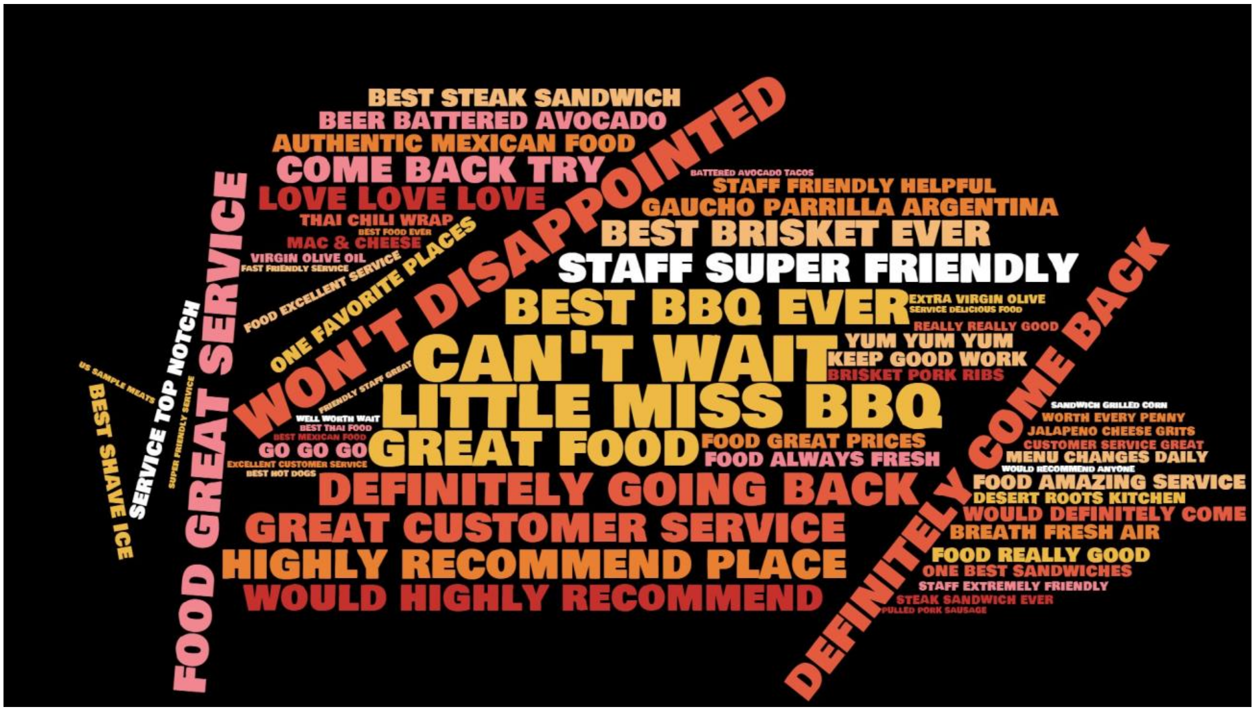
5) Use the review data from step 4 for generating tokens using NLTK.
6) Preprocessing of review text by converting it in lowercase and decoding in UTF-8 format.
7) Cleaning the tokens by removing stop words and punctuations from the tokens.
8) Generating N-grams from the tokens available now using n = 3.
9) Generate frequency dictionary out of these N-grams.
10) For each business category generate the word cloud of N-grams using these dictionaries in lower and higher rating segment.
11) Now based on these visual word clouds, try to identify the pattern in user reviews for good and poor ratings.

Business Category: Restaurant

Lower rating reviews:

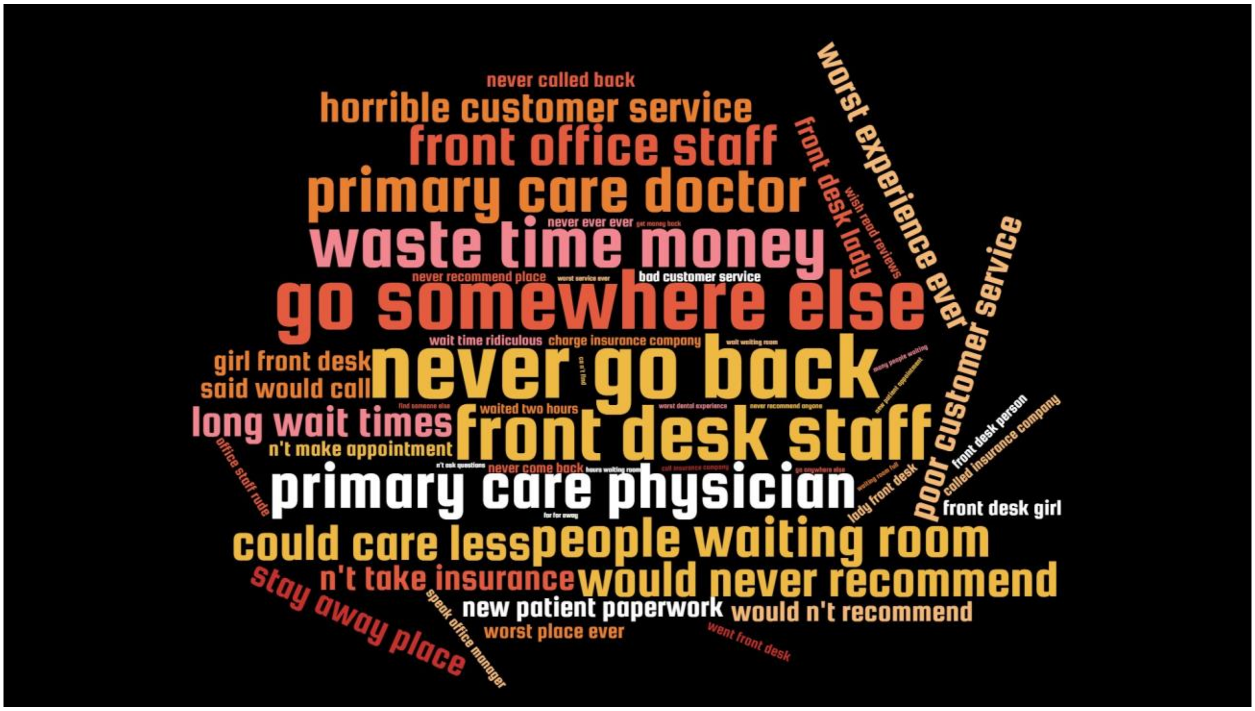


Higher rating reviews:

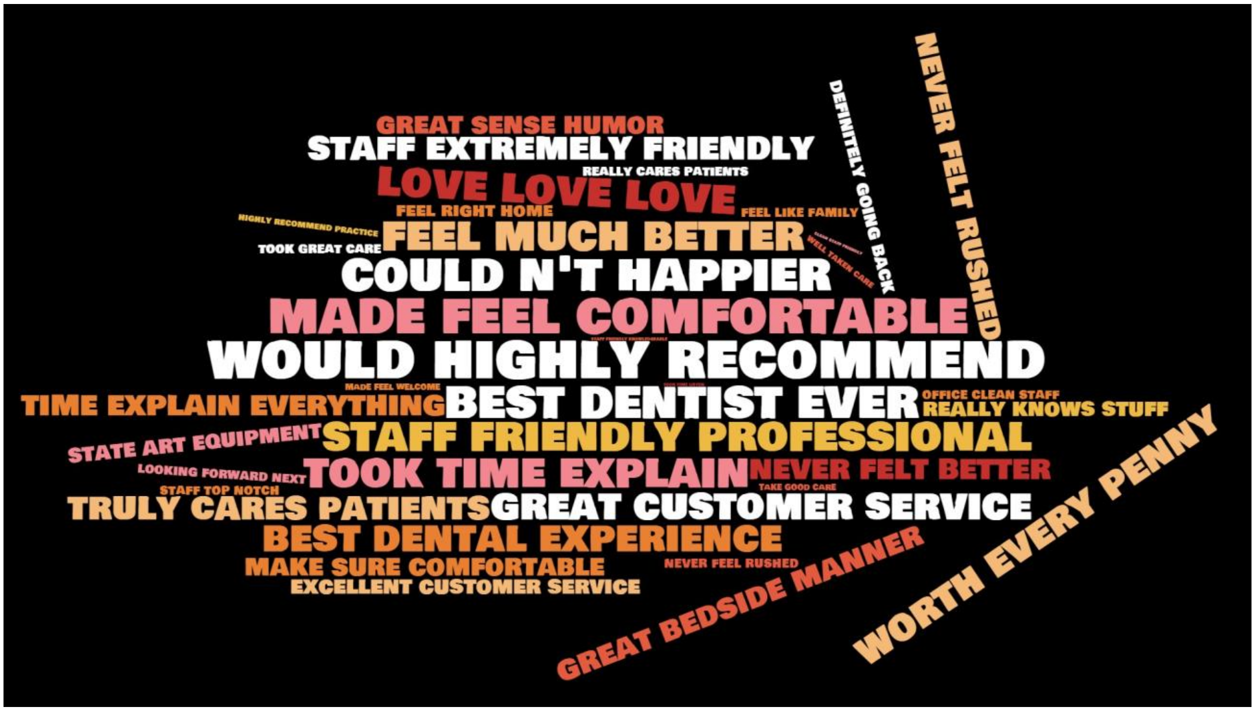


Business Category: Health Services

Lower rating reviews:

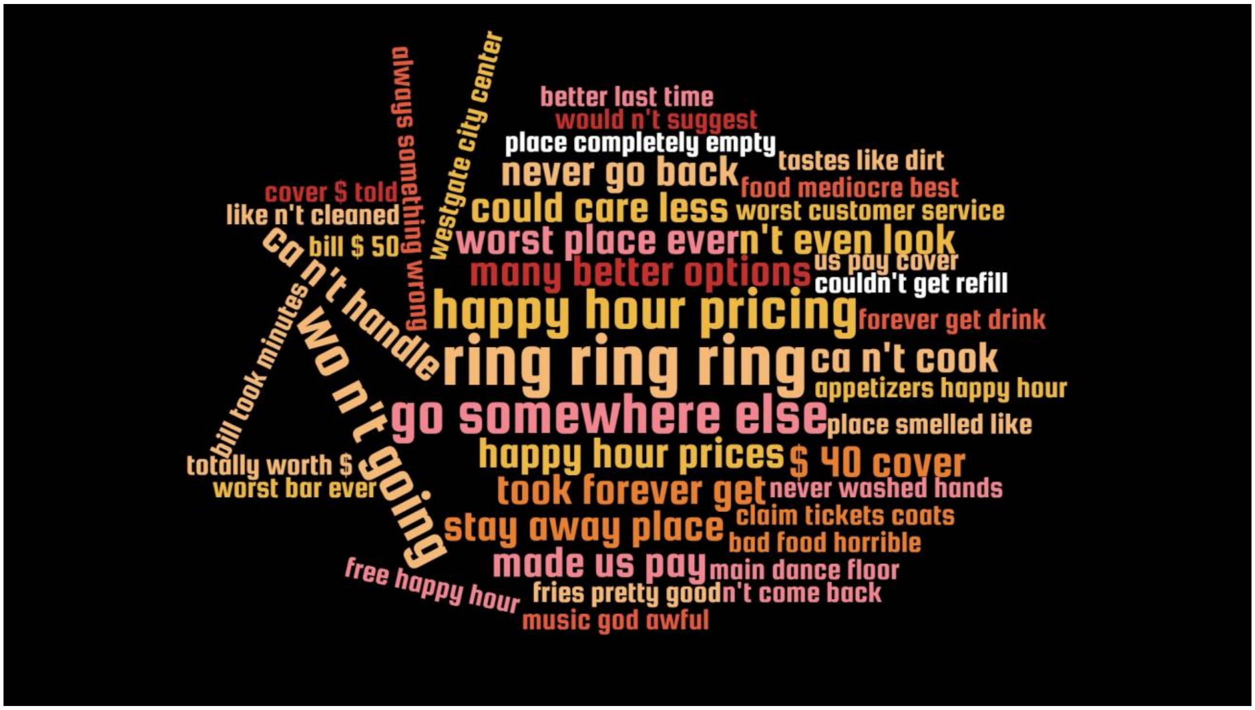


Higher rating reviews:

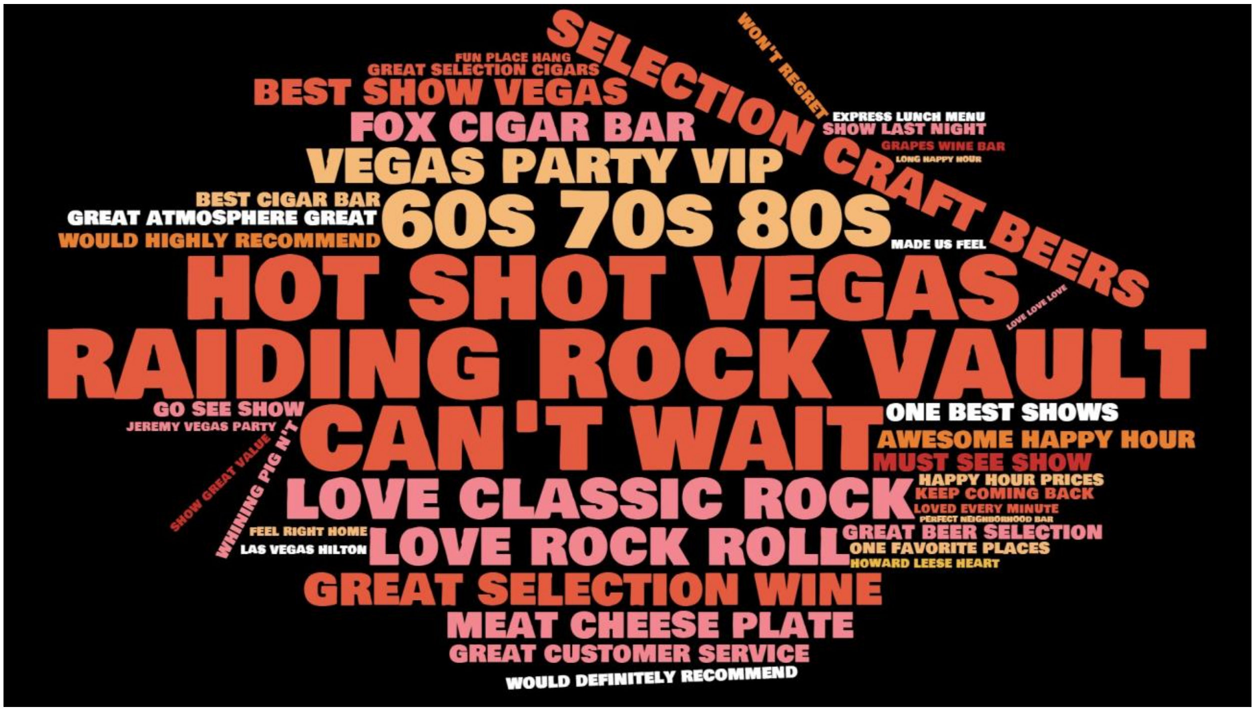


Business Category: Night Life

Lower rating reviews:



Higher rating reviews:



Rishabh Agrawal : 110467487

Results: (Business Categories)

Restaurant Business:

	Common Patterns in higher rating	Common Patterns in lower rating
1	Great Food	Wrong order served.
2	Great BBQ	Delay in serving food or not delivered the order at all.
3	Super friendly staff	Fast food chains are having maximum complaints.
4	Service top notch	Poor food quality, cold Food.
5	Good food in great price	Got food poising.
6	Food always fresh	Staff doesn't seem care.
7	Fresh Air	Phone ringing but nobody receiving call.

Health and Medical:

	Common Patterns in higher rating	Common Patterns in lower rating
1	Great Customer service	Waiting room hours(wait time)
2	Make feel comfortable	Front desk person
3	Helpful and knowledgeable staff	Primary care doctor
4	Take time and explain things	Poor customer service
5	State of Art equipment's	Can't take appointment
6	Doctor is one of the best	Rude staff
7	Office cleaning	Can't ask questions
8	Never feel rushed	Paperwork for new patient

Night Life:

	Common Patterns in higher rating	Common Patterns in lower rating
1	Vegas!!	Ring ring ring!!
2	Rock n Roll show	High cover charge
3	Classic Rock shows	Many better options
4	60s, 70s, 80s music shows	Delay in serving drink
5	Great collection of Wine, Beer and Cigar	Awful Music
6	Happy Hour prices	Empty place or smelly place
7	Great atmosphere	High Happy Hour prices
8	Great customer service	Food mediocre or taste like dirt

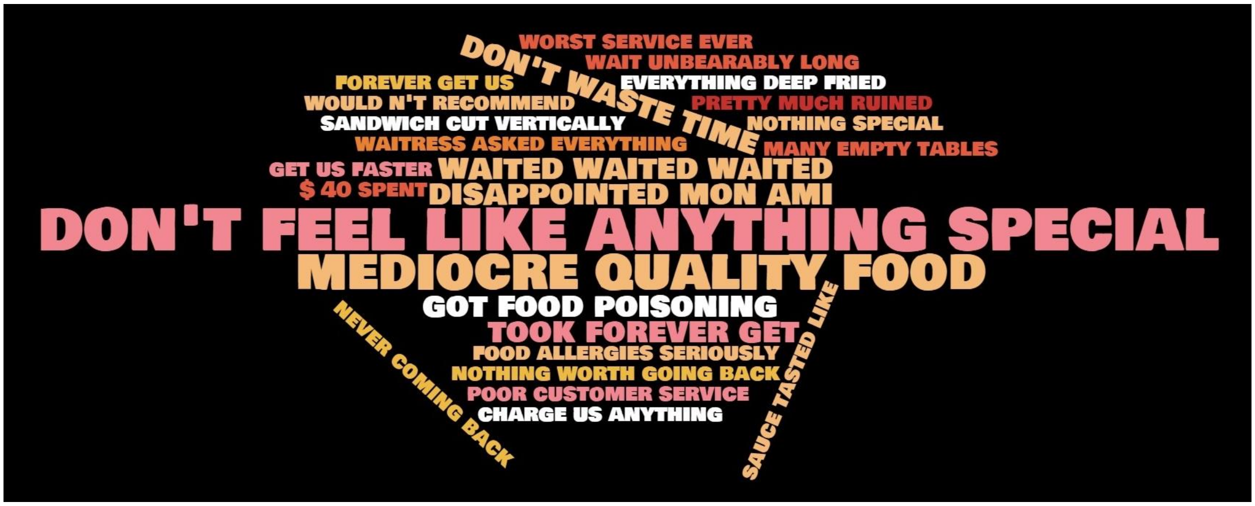
Problem-2 Approach:

In this data gathering part with pandas changes but the n-grams extraction and rest of the pipeline remains same.

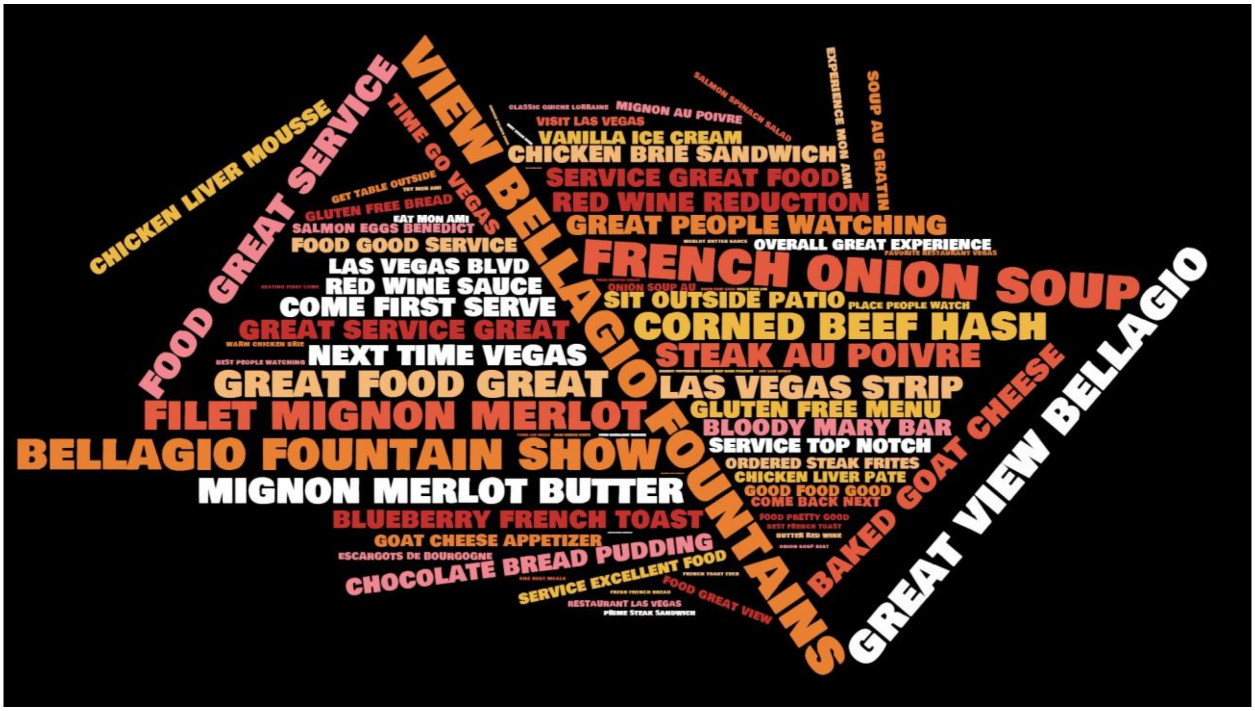
Example Single Business unit: (**Business ID**: 4bEjOyTaDG24SY5TxsaUNQ)

Avg rating: 4. **Category:** ['Breakfast & Brunch', 'Steakhouses', 'French', 'Restaurants']

Lower rating reviews: (<=2.0)



High rating reviews: (>=4)



Results: (Individual business Firm)

	Common Patterns in higher rating	Common Patterns in lower rating
1	View Bellagio fountain across the street	Longer wait time and empty tables
2	French onion soup	Medicare food quality lead to food poisoning
3	Baked goat cheese	Poor Customer Service
4	Filet mignon merlot	Charges are high
5	Corned Beef Hash	Everything is deep fried

Conclusions:

So First with the above text analysis on three business categories, we can find out the interesting patterns in user review. Here we can easily see the likes and dislikes for each category is different.

Yet there are some common attributes as **Customer service, staff behavior, delay in order completion** etc. which can be seen having their footprints across the categories. We can also mark them as basic necessity for a successful business of any category.

Along with these common traits, each business category is having its very own attributes which people like or dislike. **Nightlife** is a good example where we have found really interesting things which people like the most and these are **Vegas, Music Shows and Fine collection of drinks and cigars.** Similar interesting feature traits have been observed for other categories.

In Problem second solution we divided the reviews in two data sets: high and low rated reviews and tried to find out **what are the main attractions and areas of improvement** for that business unit.

Based on these findings we can confirm the importance of user reviews and finding patterns in the data which can be useful for businesses to survive and growth.

Other Possible use cases with this approach:

Predicting the best attributes for any business at a specific location based on user reviews to improve the chances of success in that business.

References:

Word Cloud:

http://www.wordclouds.com/

NLTK:

https://blogs.princeton.edu/etc/files/2014/03/Text-Analysis-with-NLTK-Cheatsheet.pdf

http://www.nltk.org/book/ch05.html

http://streamhacker.com/2010/05/24/text-classification-sentiment-analysis-stopwords-collocations/

Constructing Frequency tables of n-grams:

http://stackoverflow.com/questions/11763613/python-list-of-ngrams-with-frequencies

http://stackoverflow.com/questions/24289553/python-nltk-ngrams-filtering-and-excluding?noredirect=1#comment37580272_24289553

Rishabh Agrawal : 110467487