CSE 5243 Introduction to Data Mining

Lab 5: Yelp Dataset Challenge
(Executive Summary)

Bharat Suri (suri.40@osu.edu)
Nithin Senthil Kumar (senthilkumar.16@osu.edu)
Taught By: Michael Burkhardt
Date: 11/24/2019

1. Introduction

The aim of this project was to use cluster analysis and text mining techniques on a real-world dataset. The dataset used in this project is the publicly available dataset from the Yelp Dataset Challenge. This dataset is quite vast and was explored thoroughly before narrowing down the search to restaurant reviews in the city of Toronto for the year 2018. Cleaning of the data was done at different levels and the reviews were processed in order to extract features from them. For each restaurant, we have a large set of textual data in the form of reviews. To identify potential patterns or clusters among the reviews, we make use of basic Bag-of-words features and perform text mining and extract some concepts. KMeans Clustering is further applied to the subsets of reviews divided into 4 groups based on the date of reviews for the four quarters in the year of 2018. Finally, the resulting clusters are analyzed to check for trends in the entire year by observing the change of concepts in the reviews and corroborate the concept findings. In conclusion, the final analysis has been discussed along with some recommendations as to how the work can be extended into a large content-based recommender system.

2. Business Understanding

A massive part of a business' success rate depends upon how popular it is amongst its customers. While historically, word of mouth used to determine how good or reliable a business is. In today's world, this has largely been replaced by online reviews from a business' patrons and new customers all over social media. One such social media platform is yelp. Yelp reviews come in the form of star ratings as well as textual descriptions of one's opinion or experience. There is a lot of information that can be derived from the text in these reviews.

Our reasoning behind choosing to analyze reviews comes from the effect that reviews have on our everyday lives. This is because reviews reflect the trends in society both temporally and spatially. For example, reviews about restaurants may start off poor but change to become more favorable over time or vice versa. The number of reviews about a given cuisine or restaurant over a particular period may vary based on cultural trends during that time period. The reviews can also reflect the popular trends in terms of cuisine or culture in different locations over time. There is potential to uncover how consumption of different food/cuisines vary with time and location.

We have chosen to analyze patterns in one geographical location - Toronto. Toronto was chosen because it was the single largest location in terms of the number of restaurants it contained. This was done to reduce the sample size of the number of reviews which was otherwise in the order of millions.

Our objective is to identify clusters that can be formed with the reviews about restaurants in the city of Toronto over the four quarters of the year 2018 and to infer trends from them, these trends may be in terms of categories and labels that may end up being clustered together. In addition to information and clusters obtained from textual analysis of reviews, clusters obtained by including location information, reviews and categories could also be analyzed to identify and investigate interesting patterns that may emerge about the spread of different clusters in Toronto. By combining clusters extracted from reviews, locations, and categories, meaningful information is inferred and trends over time are visualized.

Given that useful trends are found, this information can be used by businesses to determine hotspots around Toronto where restaurants of certain categories are most likely to succeed. For existing restaurants, the information inferred may be used to cater to patrons based on current trends so that the restaurant remains relevant. For the average user, the information extracted may be fed into recommender systems to predict restaurants to visit or locations to frequent for particular interests.

3. Data Understanding and Preprocessing

3.1 Dataset description

The Yelp dataset consists of 5 JSON files. These JSON files which are listed below contain information about businesses that are registered on Yelp, users who have registered on yelp and their review activity.

1. Business.json
2. Review.json
3. User.json
4. Checkin.json
5. Tip.json

Upon preliminary investigation, we found that the number of individual records in these files, which are nothing but the number of rows in each table, is as below:

1. 192609 - business.json
2. 161950 - checkin.json
3. 200000 - photo.json
4. 6685900 - review.json
5. 1223094 - tip.json
6. 1637138 - user.json

Given the nature of our task, we only need the information contained in business.json and review.json. Business.json consists of businesses of every kind and not just restaurants. We proceeded to extract only the businesses which were restaurants and their reviews. In the following sections, we describe how the data in these files were processed to narrow down the search to the relevant subset of data for this project.

The features of records in business.json is as follows.

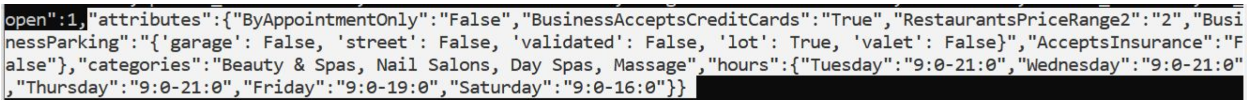
- Business_id: Unique identifier for each business
- Name: Name of the Business
- Address: Address at which the business is located
- City: City in which the business is located
- State: State in which the business is located
- Postal Code: Postal code at which the business is located
- Latitude: Latitude of the location
- Longitude: Longitude of the location
- Stars: Average star rating of the restaurant
- Review Count: Number of reviews that were left for the restaurant
- Attributes: Miscellaneous attributes identifying each restaurant
- Hours: Times of the week for which the restaurant is open
- Categories: Contains a string of labels that serve to broadly describe/categorize a restaurant.
- Is_open: Whether the restaurant is open or not.

The feature of records in review.json are as follows:

- Review id: Unique identifier for each review
- Business_id: Unique identifier for each business
- User_id: Unique identifier for each user that left a review.
- Stars: Star rating of the review
- Useful: The number of reviews left for the restaurant that was marked useful by other users.
- Cool: The number of reviews left for the restaurant that was marked cool by other users.
- Funny: The number of reviews left for the restaurant that was marked funny by other users.
- Text: The actual textual content of the review.
- Timestamp: Time the review was written.

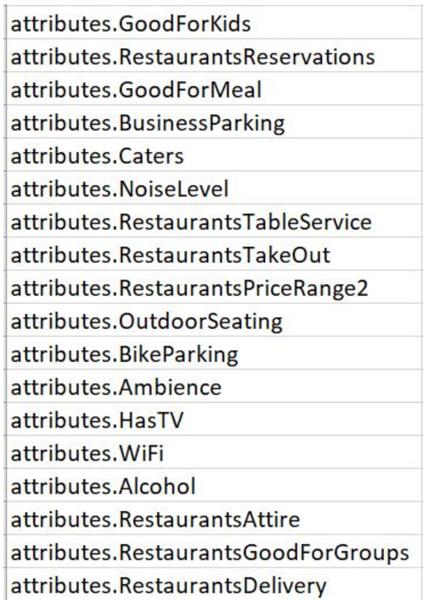
3.2 Preprocessing

Initially, we had to load the data into data frames. The JSON files were not in JSON format, instead, they were large files where each line was a record stored in JSON format. The records were read into data frames with certain columns such as attributes having values stored as JSON objects. These JSON values were then normalized to separate them out into specific columns using the method json_normalize.



```
open":1,"attributes":{"ByAppointmentOnly":"False","BusinessAcceptsCreditCards":"True","RestaurantsPriceRange2":"2","Busi
nessParking":"{'garage': False, 'street': False, 'validated': False, 'lot': True, 'valet': False}","AcceptsInsurance":"F
alse"},"categories":"Beauty & Spas, Nail Salons, Day Spas, Massage","hours":{"Tuesday":"9:0-21:0","Wednesday":"9:0-21:0"
,"Thursday":"9:0-21:0","Friday":"9:0-19:0","Saturday":"9:0-16:0"}}
```

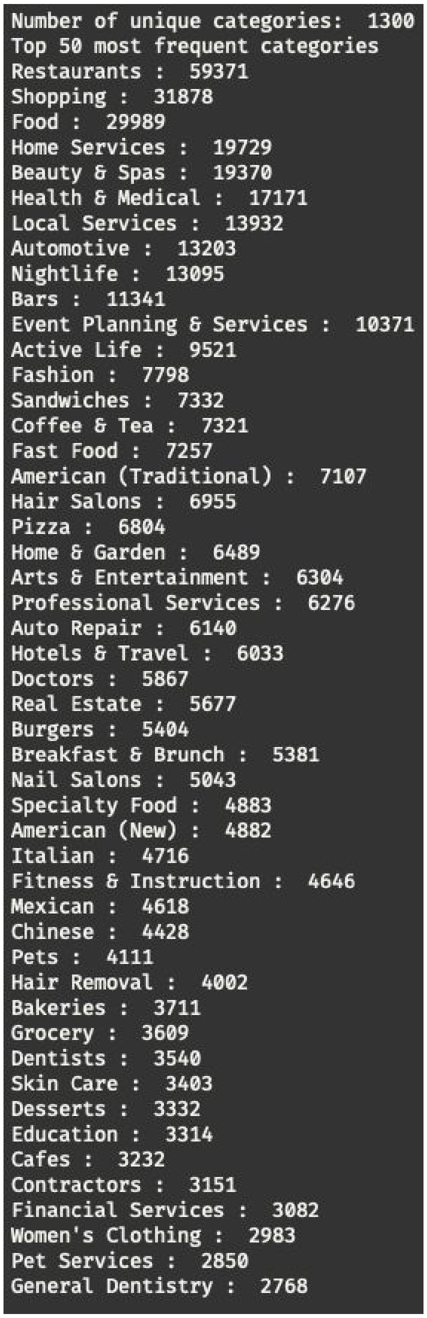
Fig: Attributes as a nested JSON object



attributes.GoodForKids
attributes.RestaurantsReservations
attributes.GoodForMeal
attributes.BusinessParking
attributes.Caters
attributes.NoiseLevel
attributes.RestaurantsTableService
attributes.RestaurantsTakeOut
attributes.RestaurantsPriceRange2
attributes.OutdoorSeating
attributes.BikeParking
attributes.Ambience
attributes.HasTV
attributes.WiFi
attributes.Alcohol
attributes.RestaurantsAttire
attributes.RestaurantsGoodForGroups
attributes.RestaurantsDelivery

Fig: Attributes after using JSON normalize

Secondly, since the dataset had more than 190K businesses and over 6M reviews, we had to narrow down our search to a specific subset of the dataset. For this purpose, the categories occurring in the business file were counted to determine the most occurring category which was ‘Restaurants’. Categories served as an important field since they acted as labels and could be used to further categorize restaurants into groups based on cuisine, type of customers they cater to, lifestyle, etc.



```
Number of unique categories:  1300
Top 50 most frequent categories
Restaurants :  59371
Shopping :  31878
Food :  29989
Home Services :  19729
Beauty & Spas :  19370
Health & Medical :  17171
Local Services :  13932
Automotive :  13203
Nightlife :  13095
Bars :  11341
Event Planning & Services :  10371
Active Life :  9521
Fashion :  7798
Sandwiches :  7332
Coffee & Tea :  7321
Fast Food :  7257
American (Traditional) :  7107
Hair Salons :  6955
Pizza :  6804
Home & Garden :  6489
Arts & Entertainment :  6304
Professional Services :  6276
Auto Repair :  6140
Hotels & Travel :  6033
Doctors :  5867
Real Estate :  5677
Burgers :  5404
Breakfast & Brunch :  5381
Nail Salons :  5043
Specialty Food :  4883
American (New) :  4882
Italian :  4716
Fitness & Instruction :  4646
Mexican :  4618
Chinese :  4428
Pets :  4111
Hair Removal :  4002
Bakeries :  3711
Grocery :  3609
Dentists :  3540
Skin Care :  3403
Desserts :  3332
Education :  3314
Cafes :  3232
Contractors :  3151
Financial Services :  3082
Women's Clothing :  2983
Pet Services :  2850
General Dentistry :  2768
```

Fig: Value counts of top business categories

Next, the dataset was further sampled for the businesses that were situated in the city of Toronto. Finally, for all such restaurants, their respective reviews were extracted from the original 6M review file. Since a majority of the businesses were restaurants, more than 4M reviews were extracted. This is another reason why we limited the search to a particular city - the sheer volume of reviews to analyze. Upon narrowing down the reviews to those in Toronto, the number of reviews had reduced to about 378K textual reviews. Since the aim was to observe concepts in those reviews over a period of time, 55K reviews were sampled for the year 2018 and further resampled quarterly thus giving 4 subsets with 16K, 16K, 16K, and 7K reviews in the four quarters of 2018 respectively.

Below is a visualization of the spread of restaurants in the city of Toronto and the distribution of the star rating. We can observe that most of the restaurants are crowded around Old Toronto and the rest are sparsely distributed across the city. Therefore, simply clustering based on location would not yield the most meaningful results, but the analysis of reviews may include insights about patterns in restaurants, their locations, and features.

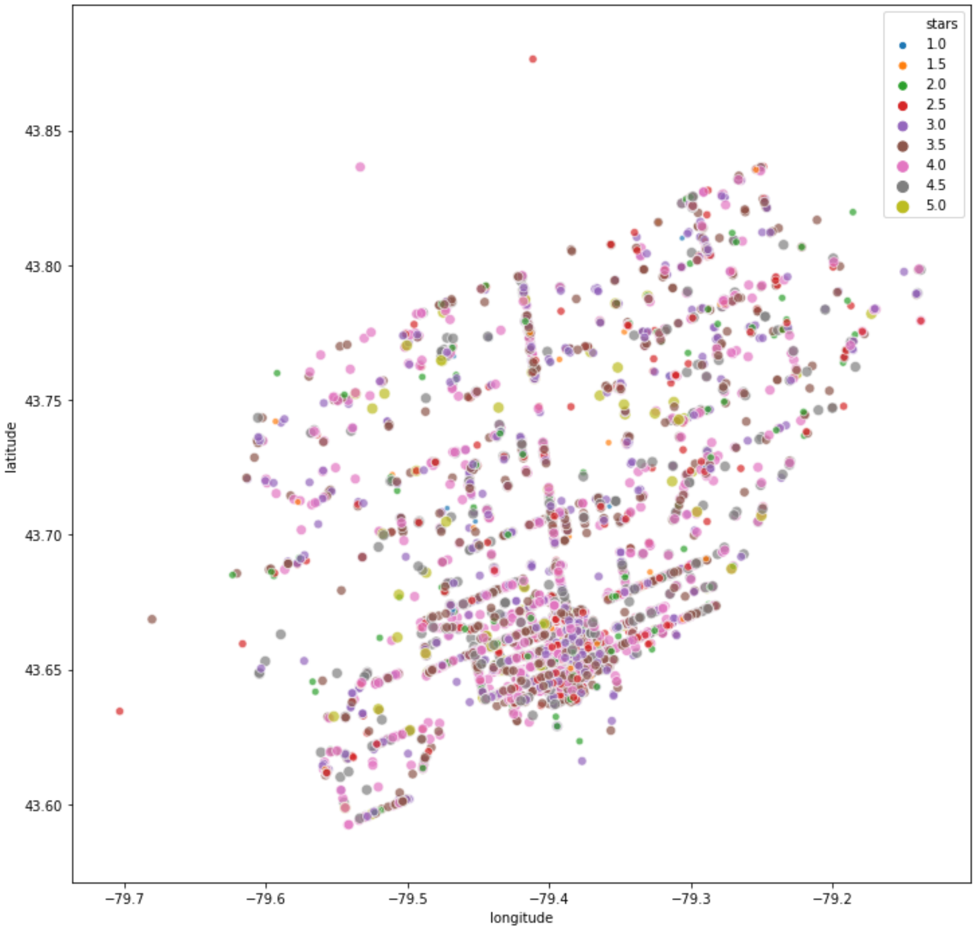
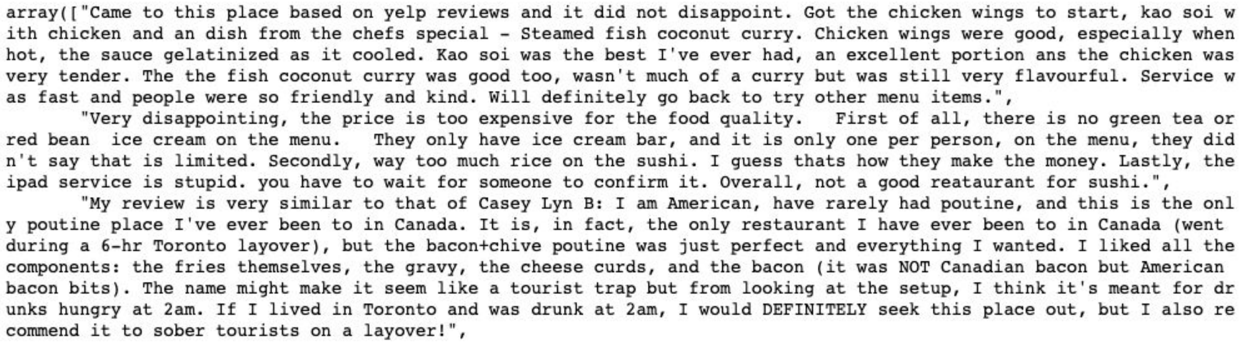


Figure: Plot of the restaurants in the city of Toronto

A key distinction of working with textual data as compared to numerical values is that is is highly subjective to determine which words are meaningless compared to which numerical values are meaningless. A common theme, however, is that words which occur frequently, but in all reviews are meaningless. The first step in preprocessing the reviews was to eliminate such words. Since these were textual reviews, the following steps were taken to clean them:

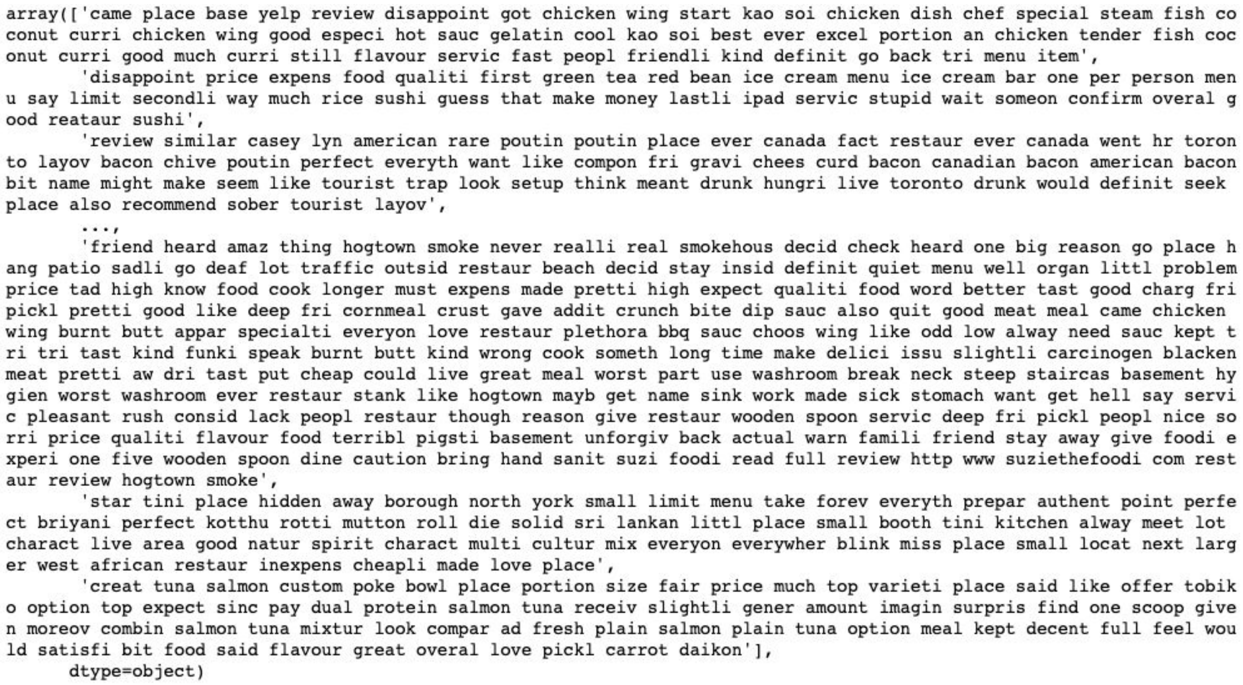
- Remove stop words
- Remove punctuation
- Remove special characters
- Stem the words using Porter Stemmer

Finally, the clean version of the corpus was used to identify concepts.



```
array(["Came to this place based on yelp reviews and it did not disappoint. Got the chicken wings to start, kao soi w
ith chicken and an dish from the chefs special - Steamed fish coconut curry. Chicken wings were good, especially when
hot, the sauce gelatinized as it cooled. Kao soi was the best I've ever had, an excellent portion ans the chicken was
very tender. The the fish coconut curry was good too, wasn't much of a curry but was still very flavourful. Service w
as fast and people were so friendly and kind. Will definitely go back to try other menu items.",
       "Very disappointing, the price is too expensive for the food quality.   First of all, there is no green tea or
red bean  ice cream on the menu.   They only have ice cream bar, and it is only one per person, on the menu, they did
n't say that is limited. Secondly, way too much rice on the sushi. I guess thats how they make the money. Lastly, the
ipad service is stupid. you have to wait for someone to confirm it. Overall, not a good reataurant for sushi.",
       "My review is very similar to that of Casey Lyn B: I am American, have rarely had poutine, and this is the onl
y poutine place I've ever been to in Canada. It is, in fact, the only restaurant I have ever been to in Canada (went
during a 6-hr Toronto layover), but the bacon+chive poutine was just perfect and everything I wanted. I liked all the
components: the fries themselves, the gravy, the cheese curds, and the bacon (it was NOT Canadian bacon but American
bacon bits). The name might make it seem like a tourist trap but from looking at the setup, I think it's meant for dr
unks hungry at 2am. If I lived in Toronto and was drunk at 2am, I would DEFINITELY seek this place out, but I also re
commend it to sober tourists on a layover!",
```

Fig. Corpus



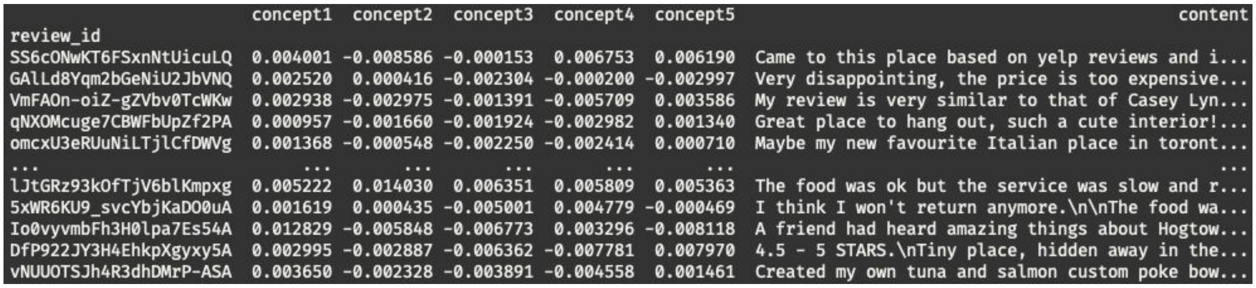
```
array(['came place base yelp review disappoint got chicken wing start kao soi chicken dish chef special steam fish co
conut curri chicken wing good especi hot sauc gelatin cool kao soi best ever excel portion an chicken tender fish coc
onut curri good much curri still flavour servic fast peopl friendli kind definit go back tri menu item',
       'disappoint price expens food qualiti first green tea red bean ice cream menu ice cream bar one per person men
u say limit secondli way much rice sushi guess that make money lastli ipad servic stupid wait someon confirm overal g
ood reataur sushi',
       'review similar casey lyn american rare poutin poutin place ever canada fact restaur ever canada went hr toron
to layov bacon chive poutin perfect everyth want like compon fri gravi chees curd bacon canadian bacon american bacon
bit name might make seem like tourist trap look setup think meant drunk hungri live toronto drunk would definit seek
place also recommend sober tourist layov',
       ...,
       'friend heard amaz thing hogtown smoke never realli real smokehous decid check heard one big reason go place h
ang patio sadli go deaf lot traffic outsid restaur beach decid stay insid definit quiet menu well organ littl problem
price tad high know food cook longer must expens made pretti high expect qualiti food word better tast good charg fri
pickl pretti good like deep fri cornmeal crust gave addit crunch bite dip sauc also quit good meat meal came chicken
wing burnt butt appar specialti everyon love restaur plethora bbq sauc choos wing like odd low alway need sauc kept t
ri tri tast kind funki speak burnt butt kind wrong cook someth long time make delici issu slightli carcinogen blacken
meat pretti aw dri tast put cheap could live great meal worst part use washroom break neck steep staircas basement hy
gien worst washroom ever restaur stank like hogtown mayb get name sink work made sick stomach want get hell say servi
c pleasant rush consid lack peopl restaur though reason give restaur wooden spoon servic deep fri pickl peopl nice so
rri price qualiti flavour food terribl pigsti basement unforgiv back actual warn famili friend stay away give foodi e
xperi one five wooden spoon dine caution bring hand sanit suzi foodi read full review http www suziethefoodi com rest
aur review hogtown smoke',
       'star tini place hidden away borough north york small limit menu take forev everyth prepar authent point perfe
ct briyani perfect kotthu rotti mutton roll die solid sri lankan littl place small booth tini kitchen alway meet lot
charact live area good natur spirit charact multi cultur mix everyon everywher blink miss place small locat next larg
er west african restaur inexpens cheapli made love place',
       'creat tuna salmon custom poke bowl place portion size fair price much top varieti place said like offer tobik
o option top expect sinc pay dual protein salmon tuna receiv slightli gener amount imagin surpris find one scoop give
n moreov combin salmon tuna mixtur look compar ad fresh plain salmon plain tuna option meal kept decent full feel wou
ld satisfi bit food said flavour great overal love pickl carrot daikon'],
      dtype=object)
```

Fig. Stemmed corpus

4. Modeling and Assessment

The two primary modeling concepts that were applied to analyze the data and extract useful information out of it were Singular Value Decomposition to carry out text mining and K-means clustering to find closeness between the reviews and compose them into clusters.

The concepts were identified by reading some of the 55K reviews along with the general ideas discussed in a review of a restaurant. Using SVD and associating each review to a particular concept, a concrete name for the concept was identified and was seen in many examples stated below.



```
                        concept1  concept2  concept3  concept4  concept5                                             content
review_id
SS6cONwKT6FSxnNtUicuLQ  0.004001 -0.008586 -0.000153  0.006753  0.006190  Came to this place based on yelp reviews and i...
GAlLd8Yqm2bGeNiU2JbVNQ  0.002520  0.000416 -0.002304 -0.000200 -0.002997  Very disappointing, the price is too expensive...
VmFAOn-oiZ-gZVbv0TcWKw  0.002938 -0.002975 -0.001391 -0.005709  0.003586  My review is very similar to that of Casey Lyn...
qNXOMcuge7CBWFbUpZf2PA  0.000957 -0.001660 -0.001924 -0.002982  0.001340  Great place to hang out, such a cute interior!...
omcxU3eRUuNiLTjlCfDWVg  0.001368 -0.000548 -0.002250 -0.002414  0.000710  Maybe my new favourite Italian place in toront...
...                          ...       ...       ...       ...       ...                                                 ...
lJtGRz93kOfTjV6blKmpxg  0.005222  0.014030  0.006351  0.005809  0.005363  The food was ok but the service was slow and r...
5xWR6KU9_svcYbjKaDO0uA  0.001619  0.000435 -0.005001  0.004779 -0.000469  I think I won't return anymore.\n\nThe food wa...
Io0vyvmbFh3H0lpa7Es54A  0.012829 -0.005848 -0.006773  0.003296 -0.008118  A friend had heard amazing things about Hogtow...
DfP922JY3H4EhkpXgyxy5A  0.002995 -0.002887 -0.006362 -0.007781  0.007970  4.5 - 5 STARS.\nTiny place, hidden away in the...
vNUUOTSJh4R3dhDMrP-ASA  0.003650 -0.002328 -0.003891 -0.004558  0.001461  Created my own tuna and salmon custom poke bow...
```

Fig. Result of SVD for all documents

The different probable concepts that were observed in the reviews with different singular values were **Food quality**, **Location** and **Ambience**, **Price**, **Menu** and **Variety**, and **Service**. Here are some examples of the reviews that were identified as belonging to different clusters

Food and quality

I ordered the chicken shawarma wrap with fries and it was pretty good. The portion was large enough to completely satisfy. Chicken was well seasoned and veggies fresh. Just wish the sauces were a bit stronger or more flavourful to amp up the taste of the wrap. Love that they are open late too! Overall it was a good meal.

I received my tacos and there were yum. The spicy crema, and the cojita cheese melded perfectly with the tender and juicy meat. My bites were so ferocious the filling began to leak out the back. That's what I get I suppose. I might just stop by again to try their burritos when I pass by this way again.

Location and ambiance

Cute decors... That's all

The ambiance of this coffee shop is super chills and laid back. I love how there's a variety of couches and tables where you can come meet with friends or take your laptop out to do some work--they have free wifi! Prices for the food and drinks are reasonably cheap; I could totally see myself camping out here all day. The staff are super friendly and checked to make sure that my friends and I are having an enjoyable experience.

Price

Drinks and food were overpriced for the quality and portion.

For $5.00, you actually get a lot of food!

Menu and variety

The menu is great for anyone... including vegetarians and those with celeriac disease.

Menu is pretty diverse and offers everything from a delicious looking platter of dips and meze to full big bowls filled with rice, dashes of vibrant raspberry pink coleslaw, and hunks of grilled meats. For a late Sunday morning, I would say this place was pretty filled... there wasn't a table easily available for 2 or 4 but there are lots of big tables that you can use for a communal share.

Service

Service here is friendly and helpful when it comes to their menu. We got the drinks very fast and the bartender/barista knows what she's doing with it comes to the espresso based drink... yeah, call me surprised.

The service here stands out a lot for me. I feel most service today is done by those who don't necessarily see it as a long term option for them... Adam is the opposite of that person.

K-Means

After performing SVD, K-Means was used to cluster the reviews. Since we attempted to find 5 concepts, the value of K chosen was also 5. The main reason behind this was to observe the clusters and the labels to see if the reviews clustered under the same identifier exhibited similar qualities in terms of the concepts observed in the output of the SVD. This output was also used in the final step of this analysis, which was to observe the change in trends over the year.

	Menu	**Location**	**Price**	**Food**	**Service**
2018 Q1	2134	2705	**4299**	3573	3971
2018 Q2	2190	2982	**4409**	3905	2956
2018 Q3	2156	3075	**4607**	3851	3103
2018 Q4	838	1162	**1808**	1474	1227

From the table above, the following observations can be made:

- Price remains the dominant concept overall
- Menu and variety in food as a concept in reviews showed steady numbers throughout the year.
- The concept of location and ambiance showed in more reviews as the year progressed.
- Service showed a decline in terms of the number of reviews talking about it.
- Food quality also showed a steady increase in the overall percentage

Due to the textual nature of the resulting clusters, most of the evaluation was carried out by observing the samples from the clustering outputs. By observing the keywords and general theme of the reviews belonging to a cluster, it was determined that the concepts chosen were in tune with the nature of the reviews in this dataset.

5. Conclusion

Below are the main observation made in this lab:

- The original dataset was quite large and needed to be processed before it became manageable.
- Feature selection was done based on the problem, these features were generated from the reviews through the means of text mining.
- Other features used were the ratings, average ratings, number of reviews and location.
- Time was also used to model the change in concepts and was thus a feature in this analysis.
- The model for finding concepts in the reviews was evaluated by observing the reviews that showed a strong association. The general theme of these reviews and the keywords in them determined which cluster they ended up in. These concepts were decided according to the content of those reviews.

Challenges, Shortcomings, and Room for Improvement

- The first challenge was to decide the scope of the dataset. While including all the 6 MIllion reviews would have been ideal to detect all possible patterns, we were restricted by computing power.

- To arrive at the scope, we presumed that trends may vary according to location and hence finalized on the location with the most reviews to offer.
- However, the biggest challenge of all was to decide upon a meaningful way to test and evaluate our model as the results were very subjective.
- Given that we attempted to perform clustering of reviews, the outcome was groups of reviews.
- To find meaning from these large groups of reviews, we tried to identify patterns based on location, patterns based on categories/labels and patterns based on the underlying concepts of the reviews themselves.
- Developing a meaningful way to evaluate and test these patterns was both a shortcoming and an area for future improvement.
- Furthermore, SVD and the bag of words approach for text mining, while useful, is not the best modeling choice that we could have made.
- A possible alternative was to use word embeddings instead of the bag of words approach and to preserve the meaning of the sentences.
- Given the nuanced nature of reviews, the application of advanced natural language processing techniques can be far more meaningful.
- There are many other clustering techniques apart from K-means clustering which could have been tried and tested to form clusters after SVD was performed.
- Clustering could have included the categories themselves.

There is great potential for real-world applications of such a system that mines concepts from texts. It can be applied to any and all review based systems.

- The problem we tried to tackle is born out of a real-world need to analyze the vast amount of textual data in reviews and identify interesting patterns among them.
- All reviews revolve around one or more concept that is based on the user's experience with the business.
- We hypothesized that this concept can be anything from the ambiance, the service that the user receives, the quality of the food, the price or the location.
- While the rating of restaurants is determined on Yelp based on stars, the actual review has a lot more meaningful information that could be useful for the restaurants to improve their popularity, to improve yelp's platform and for the user to have more personalized recommendations.

- Potential for Restaurants:
    a. Mining concepts from the user reviews and creating scorecards for the restaurants based on the concepts can allow the business to be rating based on multiple parameters.
    b. The trends identified from reviews over time can allow the restaurant owners to adapt to the continually changing of user demands.
    c. Concepts mined from the reviews can also help restaurants be up to date about what are prominently featuring concepts with their competitors.

- Potential for Yelp:

  a. Yelp’s platform is based on user ratings and optional reviews. The simplicity of just a 5-star rating system does not help Yelp’s customers and prospective business clients get a more nuanced opinion of the restaurants/businesses. Concept mining these from the reviews would.
  b. Different businesses will have different concepts or parameters by which users rank them. Having a 5-star rating system for each of these concepts may discourage users from rating honestly. However, the inference of these concepts would be more meaningful to Yelp since reviews often tend to be long and reflect the users’ actual feelings.
  c. The potential for Yelp to sell these insights and trends to businesses is an example of how concept mining of reviews can be incentivized in the real world.
  d. By incorporating and expanding on clustering based on reviews, Yelp can build better recommendation systems for users who are searching for businesses.

- Potential for Yelp’s users:
  a. The users of Yelp are also a target group of people that can directly be impacted by the idea behind this project. This is because of more accurate recommendation systems that can not only be tailored to their past reviews but also the location they are in and the trends they follow.