Text Mining and Recommendation System for Yelp Dataset

Aayushma Pant*, Shrijak Dahal*
Department of Electronics and Computer Engineering, Thapathali Campus
Institute of Engineering, Tribhuvan University
Kathmandu, Nepal

***Abstract*—This project is carried out using the Yelp dataset from where we extracted the business and user review datasets for performing restaurant recommendations. The recommendation system is created using hybrid content and a collaborative based filtering method. Here we performed the Exploratory Data Analysis on the business data followed by data cleaning and used dimension reduction techniques PCA (Principal Component Analysis) with K-means clustering on the highest variance explained components. Also, the same content-based filtering approach was taken to recommend the restaurants based on high review star count. Furthermore, the project is enhanced by implementing the TF-IDF model for text mining the reviews and recommending the restaurants based on the user text reviews following the collaborative filtering approach.[1]**

***Index Terms*—EDA, K-Means Clustering, Gradient Descent, PCA, TF-IDF, Recommendation System, Yelp Dataset**

I. Introduction

Recommendation System has been one of the leading technologies that have successfully exploited the available information and data enhancing profits, sales, views etc. Bigger companies like Amazon, YouTube, Facebook, Netflix have enforced different approaches of recommendation systems to improve the performance. It uses this technology in increasing market trends and finding potential customers by recommending with help of basics of reviews and their common interests. Different approaches like collaborative and content-based filtering method have been used in this field. Here in this project, we present the hybrid model following both this techniques on the Yelp Datasets.

Yelp is one of the largest online searching and reviewing systems for various kinds of businesses, including restaurants, shopping, home services et al. Our project is aiming to create a restaurant commendation system using the business and review yelp data. The Content-based filtering is used to recommend the restaurants based on their features and content which was procced using K-Means clustering on the principal components of the given restaurant features. From this clustering, we could recommend the similar restaurants having same features and even recommend the nearest restaurant lying on same clusters using geographical information.

Similarly used we constructed the collaborative filtering on the text review data using the TFIDF model followed by matrix factorization and Gradient descent optimization technique to recommend the restaurants based on the given text reviews. The application of this project can extend the use of yelp to a social networking level, which allows users to find new restaurants having high review and better features.

II. Literature Review

Several recommendation systems have been approached defining certain applications. In this study, we reviewed a few papers related to this project. Most of them have applied either content-based or collaborative approach methods. So, to eliminate the limitations of the individual_filtering_method, our method combines multiple filtering techniques to improve accuracy.

Arai and Barakbah [1] used a hierarchical method clustering algorithm to find the best centroids in the set. This method though generated a good result for higher dimensions but took a long time to run. Another method K-Means++ was proposed by Arthur and Vassilvitskii [2] that chooses initial centroids uniformly randomly, and choose the subsequent centroid with weighted probability proportional to the squared distance from its closest existing centroid. This method improves the speed and accuracy of the K-Means algorithm, which we use this initialization scheme in our project for grouping the restaurants having similar features. Some of the papers even used the matrix factorization method for finding latent relations between users and items which was easier to implement but stills takes a long processing time. We interpreted this matrix factorization method differently to present a collaborative approach that is different from the proposed systems.

For measuring similarity values, there are many methods that can be used, such as Pearson Correlation, Spearman Rank, Discounted Similarity, and others. Previous research [3] has tried to see how popular and good these methods are in measuring the similarity value. Based on the results of research that has been done, the Weighted Pearson Correlation produces a good accuracy value in prediction but the completeness of the data used is a critical issue. Following this problem, we purposed the Cosine similarity method to tackle this issue which worked accurately and precisely for given yelp datasets.

*They have equal contribution in this work

[1] The source code and data are available at https://github.com/Ayushma00/Text-Mining-and-Recommendation-System-for-Yelp-Dataset-

III. Theoretical and Mathematical Modeling

A. Cosine Similarity

Cosine similarity is the cosine of the angle between two vectors with dimension of n. For the project, 256047 was the dimension of data which is just the collection of users with the rating to that restaurant. There were total of 1232 restaurants and calculation of cosine similarity of restaurants was done. If two vectors or data are highly correlated then angle between them will be zero. The cosine similarity for those two data will be 1. The formula to calculate cosine similarity is given by (1).

$$\cos(d_1, d_2) = \frac{d_1 \cdot d_2}{\|d_1\| \times \|d_2\|} \qquad (1)$$

where · indicates vector dot product and $\|d\|$ is the length of vector d.

B. Euclidean Distance

Euclidean distance is the length of a line segment between two points in n-dimensional space. Its general formula is given by (2).

$$\mathrm{d}(x, y) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \qquad (2)$$

For the project, it is used for calculating distance between two restaurants with dimensions (latitude, longitude).

C. K-Means Clustering

K-means clustering aims to partition data into k clusters in a way that data points in the same cluster are similar and data points in the different clusters are farther apart. It is a partitional, center-based clustering approach:

--Data points belong to exactly one cluster.

--Each cluster is associated with a centroid (center point).

--Each point is assigned to the cluster with the closest centroid.

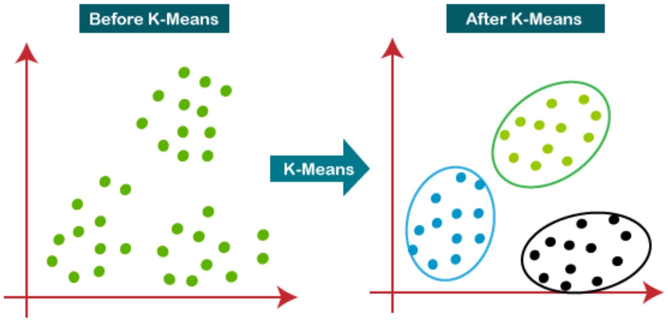


Figure III-a: K-Means Clustering with K=3

In K-Means Clustering, initial centroids are chosen randomly. So, cluster created on same data points using K-Means produces different cluster in each run.

During the fitting of model, centroid points are calculated using mean of the points in a cluster which is given by (3).

$$m_j = \frac{\sum_{x \in C_j} x}{|C_j|} \qquad (3)$$

where m_j is the Centroid of jth Cluster, C_j is the jth Cluster and x is the Objects contained in the jth Cluster.

The centroid of each cluster converges to finite value after each iteration by performing following procedures:

--Assign each point to the cluster with the nearest centroid

--Iteratively re-compute each centroid as the mean of the points assigned to it

--Ideal stopping criteria is when all centroids do not change position

The evaluation for good cluster is mainly done using SSE (Sum of Squared Error). Its formula is given by (4).

$$\mathrm{SSE} = \sum_{i=1}^{k} \sum_{x \in C_i} dist^2\,(m_i, x) \qquad (4)$$

where m_j is the Centroid of ith Cluster, C_i is the ith Cluster and x is the Objects contained in the ith Cluster and k is the total number of clusters.

To find optimal number of k for clustering, elbow method is used. The K versus SSE plot is called elbow plot. SSE falls rapidly until the optimum K value and then changes little.

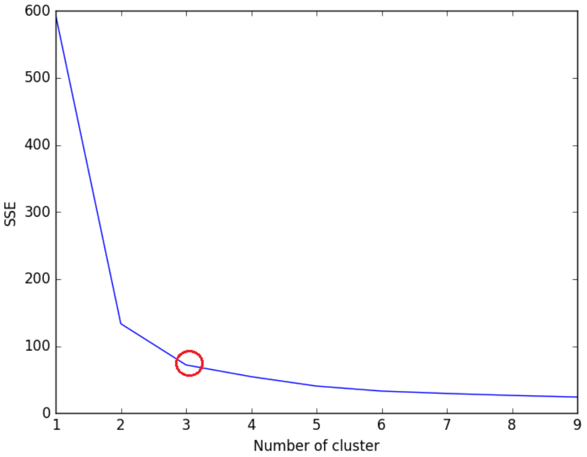


Figure III-b: Elbow Plot

D. PCA

Principal component analysis (PCA) is a technique for reducing the dimensionality of dataset, such that it increases interpretability but at the same time minimizes information loss. It does so by creating new uncorrelated variables that successively maximize variance [4].

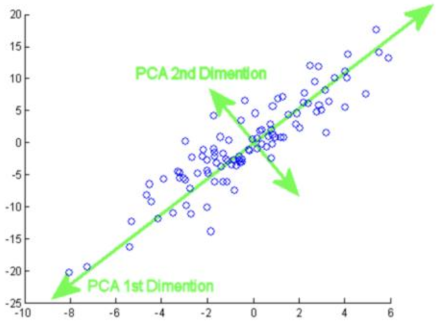


Figure III-c: PCA

First step in performing PCA is calculating covariance matrix which is given by (5).

$$S_x = \frac{1}{n-1} X X^T \qquad (5)$$

where X is m×n matrix, m is no. of data points and n is the dimension of each point.

The covariance matrix S_x is a square symmetric (m×m) matrix. The diagonal terms in S_x represents variance of measurement types and off-diagonal terms represents covariance between measurement types.

The goal of PCA is to change X to Y with matrix transformation P such that S_y is a diagonal matrix. To achieve this goal, eigen vectors of S_x is calculated and they are arranged in matrix column wise. This produces diagonal matrix as required.

For reducing dimension, PCA process keeps top principal components which explains most of variance. This is calculated by using eigenvalues corresponding to eigenvectors. By discarding remaining components, the dimension of the original data is reduced.

E. TF-IDF

Term Frequency-Inverse Document Frequency (TF-IDF) is a numerical statistic that is intended to reflect how important a word is to a document in a collection of documents. It is used as a weighting factor in searches of information retrieval, text mining. This is done by multiplying two metrics: how many times a word appears in a document, and the inverse document frequency of the word across a set of documents [5].

The term frequency (tf) of a word in a document is the first metric. There are several ways of calculating this frequency. A simple method is raw count of instances a word appears in a document. Then, there are ways to adjust the frequency, by length of a document, or by the raw frequency of the most frequent word in a document.

$$tf(t,d) = \log(1 + freq(t,d)) \qquad (6)$$

where t is the word in document d.

The inverse document frequency (idf) is the second metric. This means, how common or rare a word is in the entire document set. The closer it is to 0, the more common a word is. This is calculated by taking the total number of documents, dividing it by the number of documents that contain a word, and calculating the logarithm.

$$idf(t,D) = \log\left(\frac{N}{count(d \in D: t \in d)}\right) \qquad (7)$$

where D is the document set and N is total no. of documents.

So, if the word is quite common and appears in many documents, this number will approach 0. Otherwise, it will approach 1.

$$tf\,idf(t,d,D) = tf(t,d).idf(t,D) \qquad (8)$$

F. Gradient Descent

Gradient Descent is an iterative optimization algorithm which finds local minimum of a differentiable function.

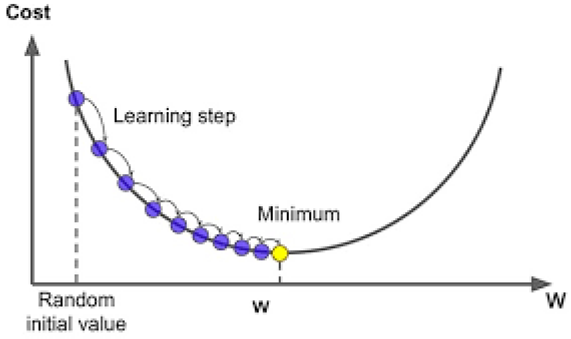


Figure III-d: Gradient Descent

In figure III-d, x axis is represented by W and y axis is represented by cost function. Let us suppose, the minimum cost is at W=w. So, if we initialize W at a random point, we want to reach to W=w. For this purpose, gradient descent is used.

For given data points y we create a function that will predict the value of points. The error in real vs. obtained value is called cost function and our goal is to minimize it.

$$\begin{array}{ll} \text{Hypothesis:} & h_\theta(x) = \theta_0 + \theta_1 x \\ \text{Parameters:} & \theta_0, \theta_1 \\ \text{Cost Function:} & J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} \left(h_\theta(x^{(i)}) - y^{(i)}\right)^2 \\ \text{Goal:} & \underset{\theta_0, \theta_1}{\text{minimize}}\, J(\theta_0, \theta_1) \end{array} \qquad (9)$$

Gradient descent algorithm takes partial derivative of cost function with respect to parameters. This value is multiplied by learning rate called alpha which is step size at which parameter will be updated. Then, obtained value is subtracted from old parameter value to obtain new parameter value that will hopefully reduce cost function. This process is repeated iteratively until cost function is nearly 0.

Gradient descent algorithm

$$\begin{aligned}&\text{repeat until convergence } \{\\ &\quad \theta_j := \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta_0, \theta_1)\\ &\qquad (\text{for } j = 1 \text{ and } j = 0)\\ &\}\end{aligned} \tag{10}$$

The learning rate or step size value is initialized by user. This value needs to be chosen carefully. In figure III-, the loss versus. epoch graph is shown. For increasing epoch, loss seems to be increasing for very high learning rate. For low learning rate the loss is decreasing very slowly. So, choosing good learning rate decreases loss very fast after every epoch and will reach to local minima quickly.

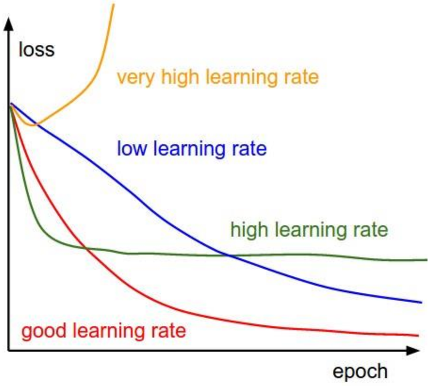


Figure III-e: Loss vs. Epoch for different learning rates

IV. Dataset Exploration

The dataset for the project is collected from Yelp Dataset. It includes about 42,153 businesses, 252,898 users, and 1,125,458 reviews, which include star ratings in the range of 1 to 5 and users' opinions in text. This dataset includes businesses other than restaurants, which is not what we want. We only took those restaurants whose review count are greater than 15 and performed data cleaning. After all the trimming, we reduced our dataset size to >26500 restaurants and >400,000 reviews.

Looking at the restaurant data we visualized certain features and properties of datasets. On plotting the geographical location of restaurants, we discovered most of the restaurants lies on the North America.

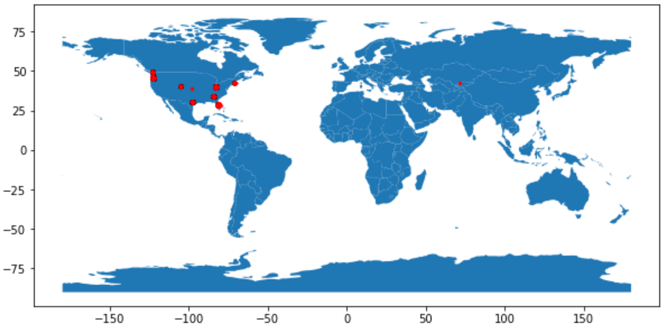


Figure IV-a: Yelp Dataset Restaurant geographic locations

Similarly, based on the state, the total counts of the restaurant in a different state of North America can be seen below. Here from the plot, it is found that Massachusetts has a larger number of restaurants.

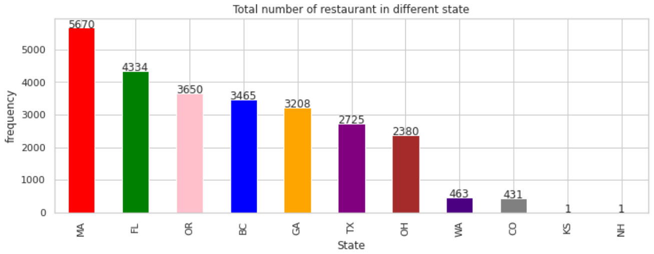


Figure IV-b: Total no. of Restaurants in different states

Moving forward to the ratings, more than 7000 restaurants have a maximum of 4.0 ratings and less than 60 restaurants e lesser ratings comparatively. From this, we can estimate most of the restaurants are good and have better quality and food to offer.

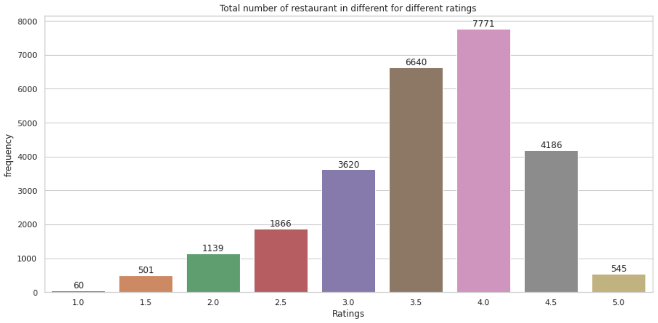


Figure IV-c: Rating Distribution in Yelp Dataset

Comparative study between the ratings gained by different restaurants across different states suggests, Massachusetts have the higher percentages of reviews as well as better counts, so we can also interpret Massachusetts as the busiest place with good reviewed stared restaurants.

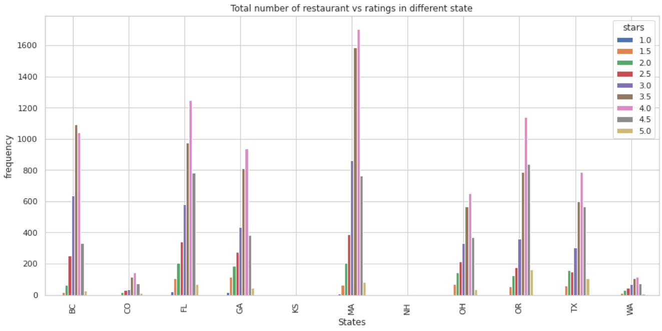


Figure IV-d: Total no. of restaurant vs. rating in different state

Now moving towards the various categories like burgers, Korean food etc., we identified top 25 categories. Among them the top 3 categories are restaurants, food and nightlife that are mostly popular in united states.

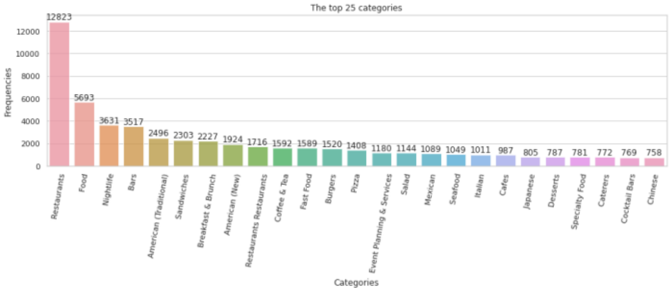


Figure IV-e: Top 25 categories of all of the Restaurants

Finally, we even look closer view towards top 20 cities from where we got most of the reviews of the restaurants.

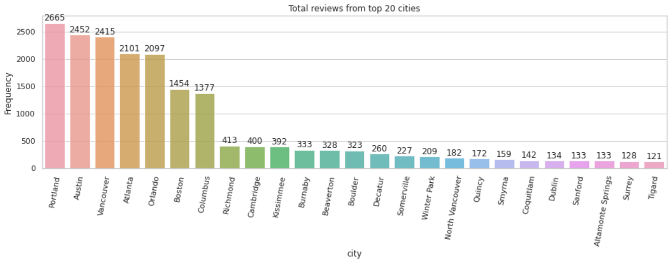


Figure IV-f: Top 20 cities which got most of the reviews

These are some of the exploratory data analyses of restaurants. Similarly, for the review's datasets, we encountered only those reviews of restaurants which are present in restaurant datasets.

V. System Block Daigram

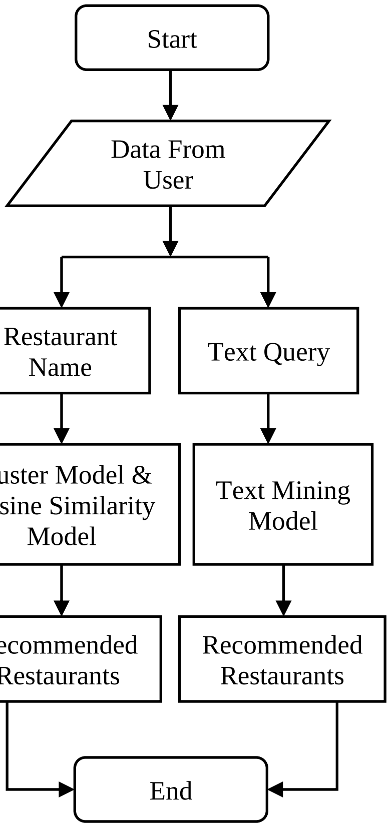


Figure V-a: System Block Diagram

The above system block diagram shows data from user interaction. Users can view restaurants from the interface and according to the restaurant name, the cluster model and cosine similarity model will give recommended restaurants list as the output for the user. Furthermore, in this model, if the user wants to search the restaurant with his/her preference like “I want to have dinner with a beautiful view.” then this input is sent to the text mining model which will recommend the restaurants.

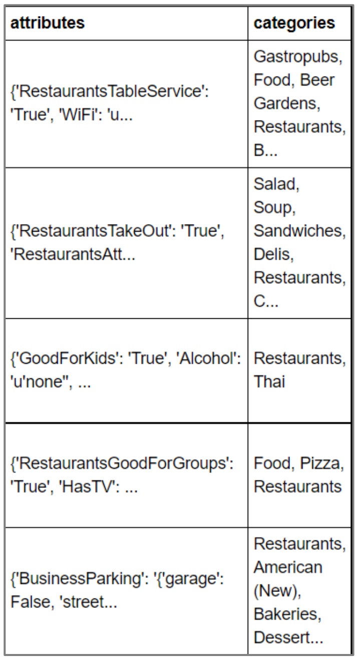
VI. Implementation Details

A. *Data Preprocessing*

The first step in doing the project was to do exploratory analysis of data. We visualized the data extensively and check if there were features that were unnecessary to our model. We removed different columns from the business table like postal code, date etc., trimmed and cleaned the datasets for better processing.

This project focuses on two data Frames, business datasets and review datasets. The business table was huge with 1000000 rows. Processing the whole table poses a huge challenge in RAM usage and CPU utilization. So, we filtered the table such that only businesses which are open and have a review count greater than 15. Again, the business table contained diverse types of business such as salon, movie hall, restaurant etc. We planned to analyze only on restaurant data so, the business table which does not contain restaurants in its category was discarded. After applying these reprocessing steps only 26328 restaurants were left on the table.

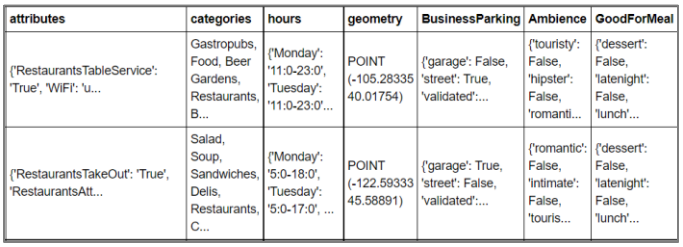
For categories in restaurant, it was in the form of string with each category separated by comma which can be seen in figure VI-a.



attributes	categories
{'RestaurantsTableService': 'True', 'WiFi': 'u...	Gastropubs, Food, Beer Gardens, Restaurants, B...
{'RestaurantsTakeOut': 'True', 'RestaurantsAtt...	Salad, Soup, Sandwiches, Delis, Restaurants, C...
{'GoodForKids': 'True', 'Alcohol': 'u'none'', ...	Restaurants, Thai
{'RestaurantsGoodForGroups': 'True', 'HasTV': ...	Food, Pizza, Restaurants
{'BusinessParking': '{'garage': False, 'street...	Restaurants, American (New), Bakeries, Dessert...

Figure VI-a: Table containing Attributes and Categories

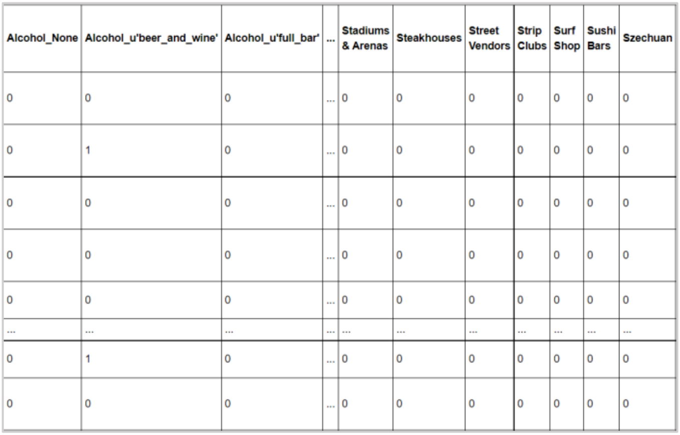
Similarly, in figure VI-the attributes column have data in form of nested dictionary. So, first we created new column with nested dictionary as a separate column to form simple dictionary as in figure VI-b.



attributes	categories	hours	geometry	BusinessParking	Ambience	GoodForMeal
{'RestaurantsTableService': 'True', 'WiFi': 'u...	Gastropubs, Food, Beer Gardens, Restaurants, B...	{'Monday': '11:0-23:0', 'Tuesday': '11:0-23:0'...	POINT (-105.28335 40.01754)	{'garage': False, 'street': True, 'validated':...	{'touristy': False, 'hipster': False, 'romanti...	{'dessert': False, 'latenight': False, 'lunch'...
{'RestaurantsTakeOut': 'True', 'RestaurantsAtt...	Salad, Soup, Sandwiches, Delis, Restaurants, C...	{'Monday': '5:0-18:0', 'Tuesday': '5:0-17:0', ...	POINT (-122.59333 45.58891)	{'garage': True, 'street': False, 'validated':...	{'romantic': False, 'intimate': False, 'touris...	{'dessert': False, 'latenight': False, 'lunch'...

Figure VI-b: Nested dictionary into separate columns

Then all the unique category in categories column was made separate column. Similarly, attributes containing dictionary was also made in separate column as shown in figure VI-c.



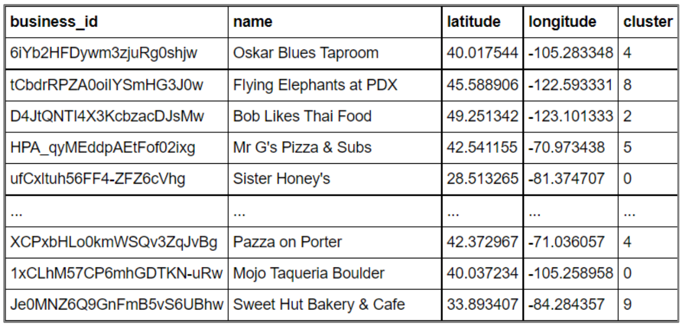
Alcohol_None	Alcohol_u'beer_and_wine'	Alcohol_u'full_bar'	...	Stadiums & Arenas	Steakhouses	Street Vendors	Strip Clubs	Surf Shop	Sushi Bars	Szechuan
0	0	0	...	0	0	0	0	0	0	0
0	1	0	...	0	0	0	0	0	0	0
0	0	0	...	0	0	0	0	0	0	0
0	0	0	...	0	0	0	0	0	0	0
0	0	0	...	0	0	0	0	0	0	0
...	...	...	...	...	...	...	...	...	...	...
0	1	0	...	0	0	0	0	0	0	0
0	0	0	...	0	0	0	0	0	0	0

Figure VI-c: Categorical and Attributes data all in separate column

After converting all the data, its dimension was 1189. We then applied PCA to the data and found at around 200 principal components the cumulative explained variance is more than 95%. So, dimension of data was reduced from 1189 to 200.

B. Model Training

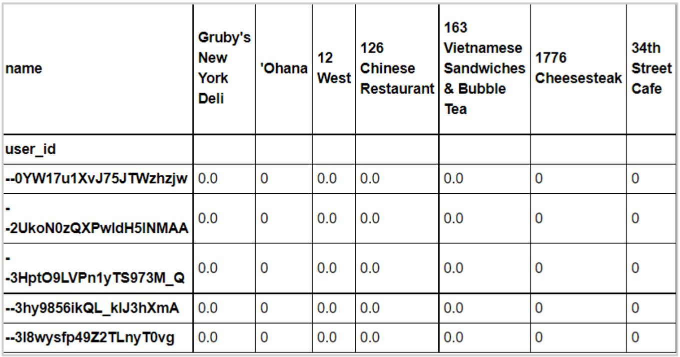
We applied K-Means Clustering algorithm with K ranging from 2 to 29. And, from elbow plot we found optimal K as 10. Then, the trained cluster points were added to the table.
This is seen in figure VI-d.



business_id	name	latitude	longitude	cluster
6iYb2HFDywm3zjuRg0shjw	Oskar Blues Taproom	40.017544	-105.283348	4
tCbdrRPZA0oilYSmHG3J0w	Flying Elephants at PDX	45.588906	-122.593331	8
D4JtQNTI4X3KcbzacDJsMw	Bob Likes Thai Food	49.251342	-123.101333	2
HPA_qyMEddpAEtFof02ixg	Mr G's Pizza & Subs	42.541155	-70.973438	5
ufCxltuh56FF4-ZFZ6cVhg	Sister Honey's	28.513265	-81.374707	0
...	...	...	...	...
XCPxbHLo0kmWSQv3ZqJvBg	Pazza on Porter	42.372967	-71.036057	4
1xCLhM57CP6mhGDTKN-uRw	Mojo Taqueria Boulder	40.037234	-105.258958	0
Je0MNZ6Q9GnFmB5vS6UBhw	Sweet Hut Bakery & Cafe	33.893407	-84.284357	9

Figure VI-d: Final Restaurant table with cluster cloumn

For Cosine similarity, the users review table and business table was merged and created such that the users rating was row of the table and restaurant was column as shown in figure VI-e.



name	Gruby's New York Deli	'Ohana	12 West	126 Chinese Restaurant	163 Vietnamese Sandwiches & Bubble Tea	1776 Cheesesteak	34th Street Cafe
user_id							
--0YW17u1XvJ75JTWzhzjw	0.0	0	0.0	0.0	0.0	0	0
--2UkoN0zQXPwldH5lNMAA	0.0	0	0.0	0.0	0.0	0	0
--3HptO9LVPn1yTS973M_Q	0.0	0	0.0	0.0	0.0	0	0
--3hy9856ikQL_klJ3hXmA	0.0	0	0.0	0.0	0.0	0	0
--3l8wysfp49Z2TLnyT0vg	0.0	0	0.0	0.0	0.0	0	0

Figure VI-e: Users in row and restaurants in column

Passing this data to cosine similarity function we got a matrix of 1232×1232 which is cosine similarity matrix of restaurants with restaurants.

C.1 Preprocessing Textual Data

For text mining, the reviews table was used. Reviews given by each user to each restaurant was stored on the table. First, we filtered the table by review count, Restaurants with a review count of less than 500 were discarded. Then, the reviews text was processed by removing punctuations, removing stop words. We, aggregated the review of a user for all the restaurant that user have reviewed. Similarly, reviews got by a restaurant from all the users was also aggregated into single review.

C.2 Textual Data Model Training

On formation of two reviews table for user and restaurants, TF-IDF model was implemented. The TF-IDF vectorizer with max features of 3000 was used So only the top 3000 words were saved in column.

User review vector had shape of (114622, 3000) and restaurant review vector had shape of (166,3000). These two matrices were saved as P and Q. Also, the user rating matrix for restaurant was created for text mining training. This matrix was saved as R. Then using Gradient Descent Algorithm on matrices P and Q with R as actual value the model was trained.

D. Model Serving

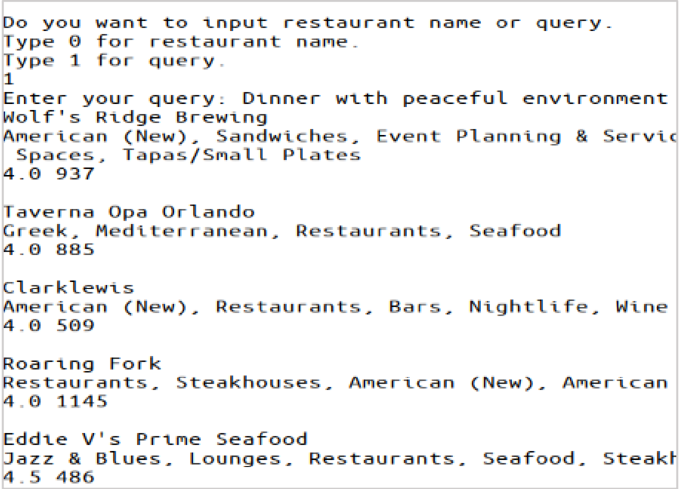
Finally, the two models were saved on csv files and pickle file as required. These models were then used in recommendation system which an recommend restaurant based on restaurant name or the query asked.

VII. Results

The project gave a decent result. Different models were trained based on data mining techniques such as PCA, K-Means Clustering, Cosine Similarity, Gradient Descent etc. which effectively brought a robust result for recommending the restaurants.

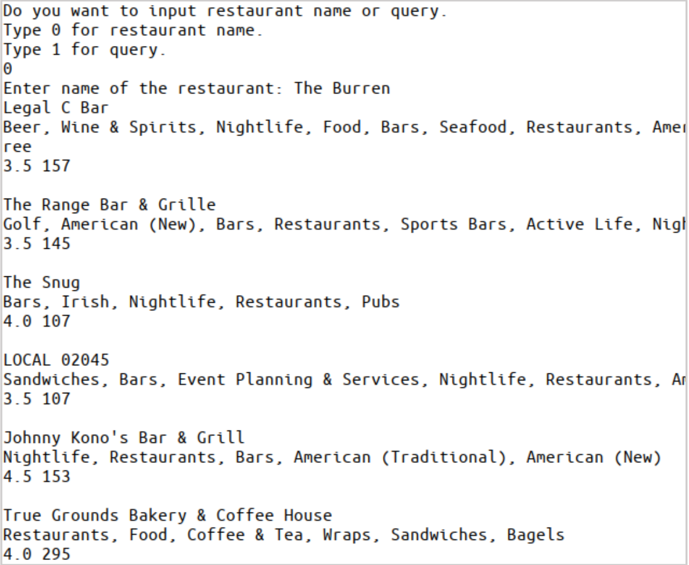
Some of the snaps of the results can be seen below. Here, in figure VII-a the result were from the collaborative filter approach method of text mining analysis. It takes query from user as input and recommends the restaurant based on the query.

Similarly in the figure VII-b the result was from the content filtering approach method created using the combination of K-Means clustering and cosine similarity matrix. It takes restaurant name as user input and recommends the restaurant based on its features, similarity and location.



```
Do you want to input restaurant name or query.
Type 0 for restaurant name.
Type 1 for query.
1
Enter your query: Dinner with peaceful environment
Wolf's Ridge Brewing
American (New), Sandwiches, Event Planning & Servic
 Spaces, Tapas/Small Plates
4.0 937

Taverna Opa Orlando
Greek, Mediterranean, Restaurants, Seafood
4.0 885

Clarklewis
American (New), Restaurants, Bars, Nightlife, Wine
4.0 509

Roaring Fork
Restaurants, Steakhouses, American (New), American
4.0 1145

Eddie V's Prime Seafood
Jazz & Blues, Lounges, Restaurants, Seafood, Steakh
4.5 486
```

Figure VII-a: Query from user and Recommended Restaurants



```
Do you want to input restaurant name or query.
Type 0 for restaurant name.
Type 1 for query.
0
Enter name of the restaurant: The Burren
Legal C Bar
Beer, Wine & Spirits, Nightlife, Food, Bars, Seafood, Restaurants, Amer
ree
3.5 157

The Range Bar & Grille
Golf, American (New), Bars, Restaurants, Sports Bars, Active Life, Nigh
3.5 145

The Snug
Bars, Irish, Nightlife, Restaurants, Pubs
4.0 107

LOCAL 02045
Sandwiches, Bars, Event Planning & Services, Nightlife, Restaurants, Ar
3.5 107

Johnny Kono's Bar & Grill
Nightlife, Restaurants, Bars, American (Traditional), American (New)
4.5 153

True Grounds Bakery & Coffee House
Restaurants, Food, Coffee & Tea, Wraps, Sandwiches, Bagels
4.0 295
```

Figure VII-b: Recommended Restaurants based on current restaurant name

VIII. Discussion and Analysis

Principal component Analysis, one of the techniques for dimensional reduction is been used in content-based filtering method. First, we used all the columns for describing the components. Upon doing this, we discovered the variance explained by the components that starts to decline and gets saturated after 200 components.

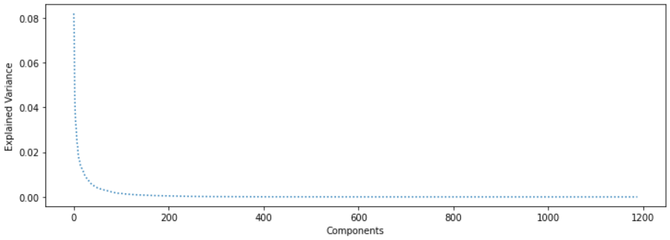


Figure VIII-a: Components vs. Explained Variance

So, following the previous analysis, on plotting the cumulative explained variance we found 95% of variance is explained by up to 200 components and thus used 200 PCA components for explaining the features of given restaurants having more than 1100 columns.

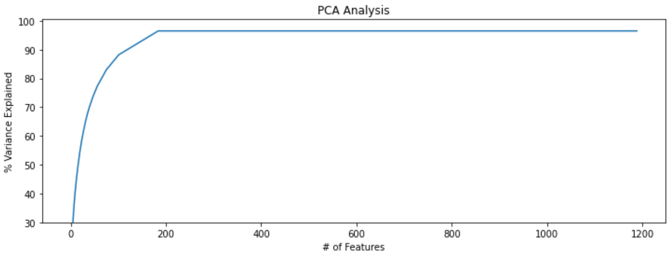


Figure VIII-b: Components vs. Cumulative Explained Variance

Now, to estimate the K-Means clusters we preferred the elbow plot method. From which we created total of 10 clusters by initializing the centroids randomly on the given PCA components.

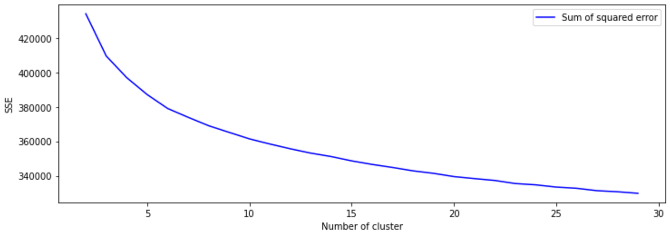


Figure VIII-c: Elbow Plot for different values of K vs. SSE

From the above approach, we created a restaurant recommending system that recommends the restaurants which are of same clusters.

Similarly, we also worked on Cosine Similarity Technique to recommend the restaurants based on the highest reviews. We created a 1232 by 1232 similarity matrix of restaurants for recommending similar restaurants based on the highest score. These two methods are merged to give the best recommendation of restaurants.

Furthermore, supporting a collaborative approach, individual TF-IDF vectorizers were created for the restaurant's reviews and the user reviews separately. Such a model has more than 3000 features from which a user rating matrix was constructed based on the user id as rows and restaurants id as columns. This is optimized using Gradient Descent Optimizations using 25 steps and a 0.001 learning rate. This took more than 8 hours for training.

Hence on completion, the final prediction tends to be good and efficient.

IX. Conclusion

Recommendation System using the hybrid model followed by the algorithms like K-Means Clustering, Cosine Similarity on the restaurant data and textual mining using TF-IDF model on review text has shown a decent result. The output tends to give a recommendation of restaurants based on two different queries like similar restaurants or recommendations based on the input review text. We have explored various models like matrix factorization, optimization techniques like Gradient Descent, similarity measuring techniques like cosine similarity and many more. However, better models and more other features are still needed to be discovered. So, in future, such terminology will be explored and researched to create a more robust model.

References

[1] K. Arai and A. R. Barakbah. “Hierarchical K-means: an algorithm for centroids initialization for K-means”. *Reports of the Faculty of Science and Engineering Saga University*, vol. 36, No.1, 2007, pp. 25-31.

[2] D. Arthur and S. Vassilvitskii, “k-means++: the advantages of careful seeding,” in SODA ’07: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms, New Orleans, Louisiana, 2007, pp. 1027--1035.

[3] A. Munaji and A. Emanuel, "Restaurant Recommendation System Based on User Ratings with Collaborative Filtering", IOP Conference Series: Materials Science and Engineering, vol. 1077, no. 1, p. 012026, 2021. Available: 10.1088/1757-899x/1077/1/012026 [Accessed 11 November 2021].

[4] I. Jolliffe, Principal component analysis. New York: Springer Science+Business Media, LLC, 2010.

[5] A. Rajaraman, J. Leskovec and J. Ullman, Mining of Massive Datasets. Cambridge University Press, 2011.