Mini Project 3

Shreya Bhatia

110619150

Introduction

Yelp is a crowd-sourced local business review and social networking site. In this project we will be aiming at predicting the primary category of the businesses present in Yelp.

Problem Statement

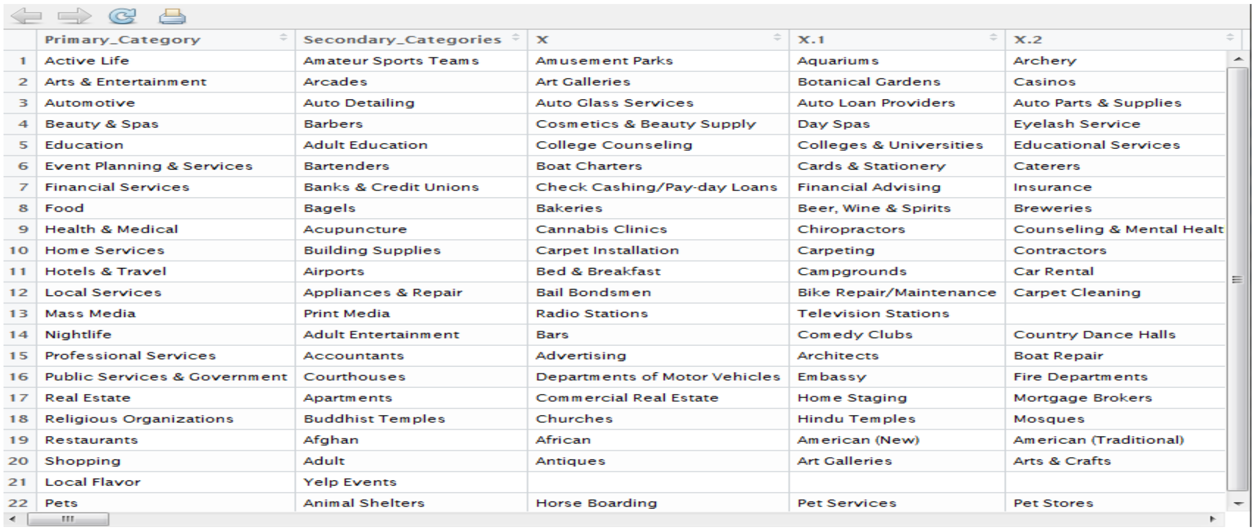
We will try different classification algorithm to automatically predict the Primary category of business based upon features like check-in information, stars, reviews. We will evaluate this model on the test data and calculate its accuracy.
We also propose to find correlation between different business categories by means of clustering. We are also interested in finding the type of words contained in different starred reviews.

Result and Discussions

- First task was to map the entries in the yelp_business file to the primary category of the business. Example:

Business ID	**Secondary Categories**	**Primary Category**
b12U9TFESStdy7CsTtcOeg	Auto Repair: Automotive	Automotive

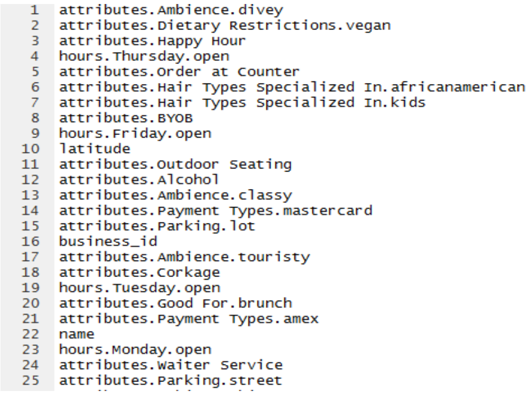
  - For this, mapping of secondary categories and Primary categories was used. [3]
  - All the entries were tagged by running this mapping.



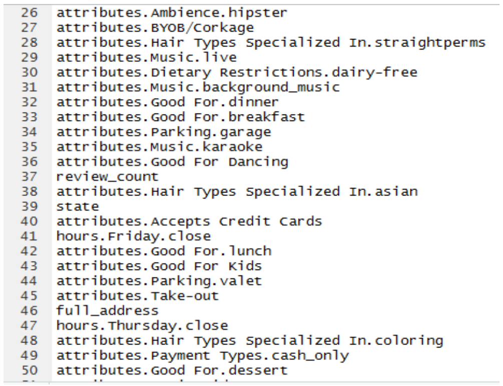
	Primary_Category	Secondary_Categories	X	X.1	X.2
1	Active Life	Amateur Sports Teams	Amusement Parks	Aquariums	Archery
2	Arts & Entertainment	Arcades	Art Galleries	Botanical Gardens	Casinos
3	Automotive	Auto Detailing	Auto Glass Services	Auto Loan Providers	Auto Parts & Supplies
4	Beauty & Spas	Barbers	Cosmetics & Beauty Supply	Day Spas	Eyelash Service
5	Education	Adult Education	College Counseling	Colleges & Universities	Educational Services
6	Event Planning & Services	Bartenders	Boat Charters	Cards & Stationery	Caterers
7	Financial Services	Banks & Credit Unions	Check Cashing/Pay-day Loans	Financial Advising	Insurance
8	Food	Bagels	Bakeries	Beer, Wine & Spirits	Breweries
9	Health & Medical	Acupuncture	Cannabis Clinics	Chiropractors	Counseling & Mental Healt
10	Home Services	Building Supplies	Carpet Installation	Carpeting	Contractors
11	Hotels & Travel	Airports	Bed & Breakfast	Campgrounds	Car Rental
12	Local Services	Appliances & Repair	Bail Bondsmen	Bike Repair/Maintenance	Carpet Cleaning
13	Mass Media	Print Media	Radio Stations	Television Stations	
14	Nightlife	Adult Entertainment	Bars	Comedy Clubs	Country Dance Halls
15	Professional Services	Accountants	Advertising	Architects	Boat Repair
16	Public Services & Government	Courthouses	Departments of Motor Vehicles	Embassy	Fire Departments
17	Real Estate	Apartments	Commercial Real Estate	Home Staging	Mortgage Brokers
18	Religious Organizations	Buddhist Temples	Churches	Hindu Temples	Mosques
19	Restaurants	Afghan	African	American (New)	American (Traditional)
20	Shopping	Adult	Antiques	Art Galleries	Arts & Crafts
21	Local Flavor	Yelp Events			
22	Pets	Animal Shelters	Horse Boarding	Pet Services	Pet Stores

Fig 1: Mapping Of Primary Category to Secondary Categories.

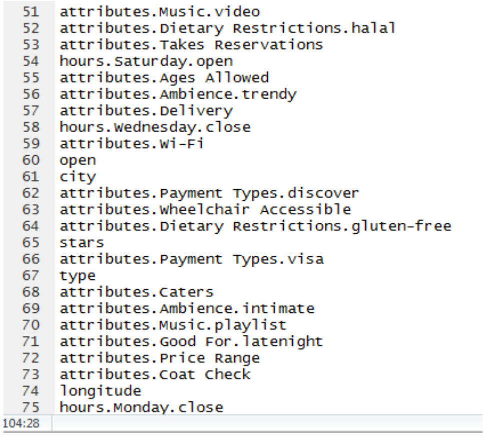
- **Prediction Using Naive Bayes Classifier**
  - In this, we predicted the primary category of the business using Naïve Bayes Classifier.
  - Features Used are :



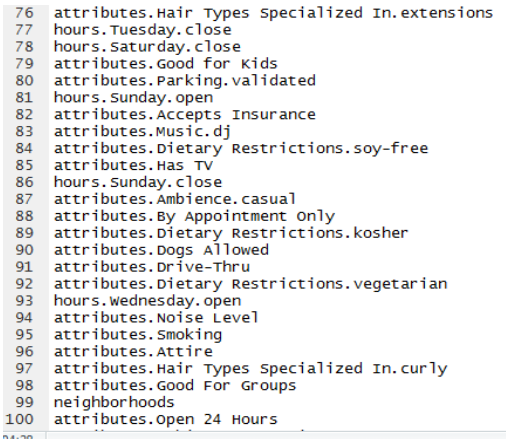
```
 1  attributes.Ambience.divey
 2  attributes.Dietary Restrictions.vegan
 3  attributes.Happy Hour
 4  hours.Thursday.open
 5  attributes.Order at Counter
 6  attributes.Hair Types Specialized In.africanamerican
 7  attributes.Hair Types Specialized In.kids
 8  attributes.BYOB
 9  hours.Friday.open
10  latitude
11  attributes.Outdoor Seating
12  attributes.Alcohol
13  attributes.Ambience.classy
14  attributes.Payment Types.mastercard
15  attributes.Parking.lot
16  business_id
17  attributes.Ambience.touristy
18  attributes.Corkage
19  hours.Tuesday.open
20  attributes.Good For.brunch
21  attributes.Payment Types.amex
22  name
23  hours.Monday.open
24  attributes.Waiter Service
25  attributes.Parking.street
```



```
26  attributes.Ambience.hipster
27  attributes.BYOB/Corkage
28  attributes.Hair Types Specialized In.straightperms
29  attributes.Music.live
30  attributes.Dietary Restrictions.dairy-free
31  attributes.Music.background_music
32  attributes.Good For.dinner
33  attributes.Good For.breakfast
34  attributes.Parking.garage
35  attributes.Music.karaoke
36  attributes.Good For Dancing
37  review_count
38  attributes.Hair Types Specialized In.asian
39  state
40  attributes.Accepts Credit Cards
41  hours.Friday.close
42  attributes.Good For.lunch
43  attributes.Good For Kids
44  attributes.Parking.valet
45  attributes.Take-out
46  full_address
47  hours.Thursday.close
48  attributes.Hair Types Specialized In.coloring
49  attributes.Payment Types.cash_only
50  attributes.Good For.dessert
```



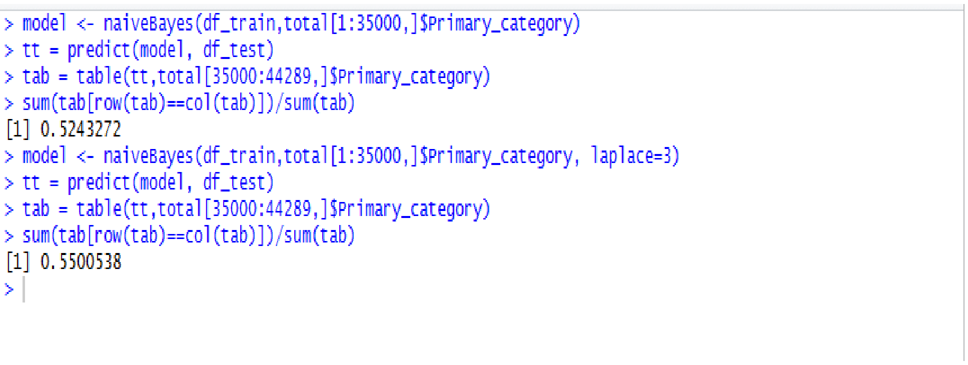
```
51  attributes.Music.video
52  attributes.Dietary Restrictions.halal
53  attributes.Takes Reservations
54  hours.Saturday.open
55  attributes.Ages Allowed
56  attributes.Ambience.trendy
57  attributes.Delivery
58  hours.Wednesday.close
59  attributes.Wi-Fi
60  open
61  city
62  attributes.Payment Types.discover
63  attributes.Wheelchair Accessible
64  attributes.Dietary Restrictions.gluten-free
65  stars
66  attributes.Payment Types.visa
67  type
68  attributes.Caters
69  attributes.Ambience.intimate
70  attributes.Music.playlist
71  attributes.Good For.latenight
72  attributes.Price Range
73  attributes.Coat Check
74  longitude
75  hours.Monday.close
104:28
```



```
 76  attributes.Hair Types Specialized In.extensions
 77  hours.Tuesday.close
 78  hours.Saturday.close
 79  attributes.Good for Kids
 80  attributes.Parking.validated
 81  hours.Sunday.open
 82  attributes.Accepts Insurance
 83  attributes.Music.dj
 84  attributes.Dietary Restrictions.soy-free
 85  attributes.Has TV
 86  hours.Sunday.close
 87  attributes.Ambience.casual
 88  attributes.By Appointment Only
 89  attributes.Dietary Restrictions.kosher
 90  attributes.Dogs Allowed
 91  attributes.Drive-Thru
 92  attributes.Dietary Restrictions.vegetarian
 93  hours.Wednesday.open
 94  attributes.Noise Level
 95  attributes.Smoking
 96  attributes.Attire
 97  attributes.Hair Types Specialized In.curly
 98  attributes.Good For Groups
 99  neighborhoods
100  attributes.Open 24 Hours
```

  - With the above features, checkin values are also used for prediction from the file yelp_academic_dataset_checkin. Secondary categories were not used in these features.
  - Data was divided into ratio of 70-30% of the data set.

- **Naive Bayes Results**



```
> model <- naiveBayes(df_train,total[1:35000,]$Primary_category)
> tt = predict(model, df_test)
> tab = table(tt,total[35000:44289,]$Primary_category)
> sum(tab[row(tab)==col(tab)])/sum(tab)
[1] 0.5243272
> model <- naiveBayes(df_train,total[1:35000,]$Primary_category, laplace=3)
> tt = predict(model, df_test)
> tab = table(tt,total[35000:44289,]$Primary_category)
> sum(tab[row(tab)==col(tab)])/sum(tab)
[1] 0.5500538
> 
```

Fig 2: Accuracy of Naïve Bays Classifier Model

In the above figure,

- Accuracy of the classifier (Precision) without Laplace smoothing is 52%.
- Accuracy of the classifier (Precision) with Laplace smoothing is 55%.

- **Check-in Based Business Clustering**
  - We will be using K-means clustering for this. First Task is to find K, number of clusters. We will be using elbow test for this. [1]

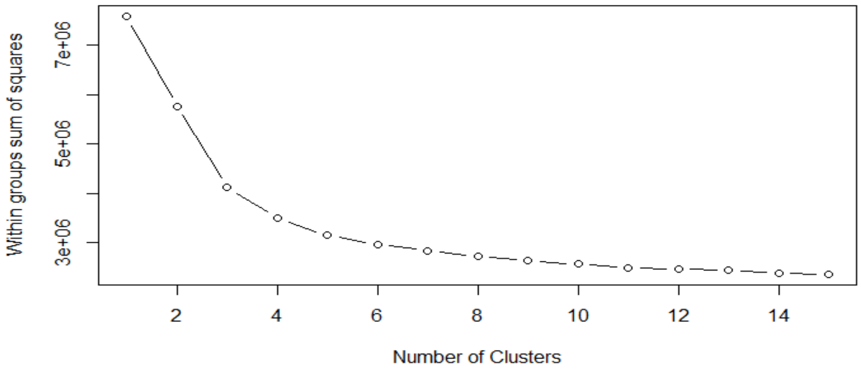


Fig 3: Plot within group of sum of squares VS number of clusters extracted

- The sharp decreases from 1 to 3 clusters (with little decrease after) suggest a 3-cluster solution.

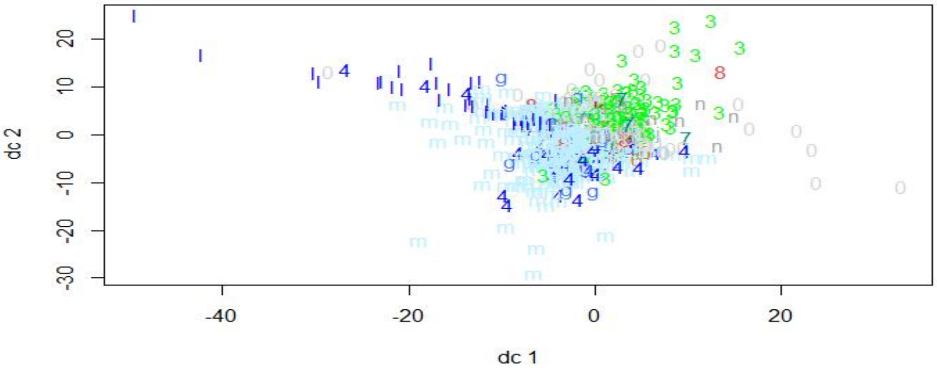


Fig 4: Plot of businesses grouped by Primary Category.

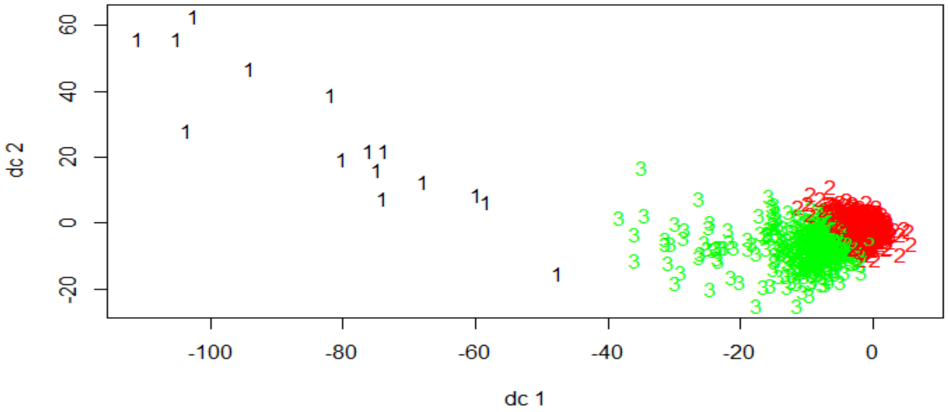
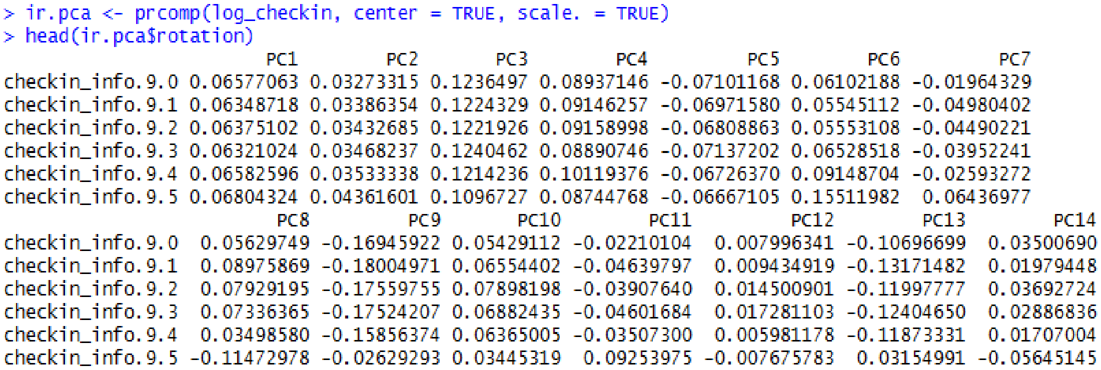


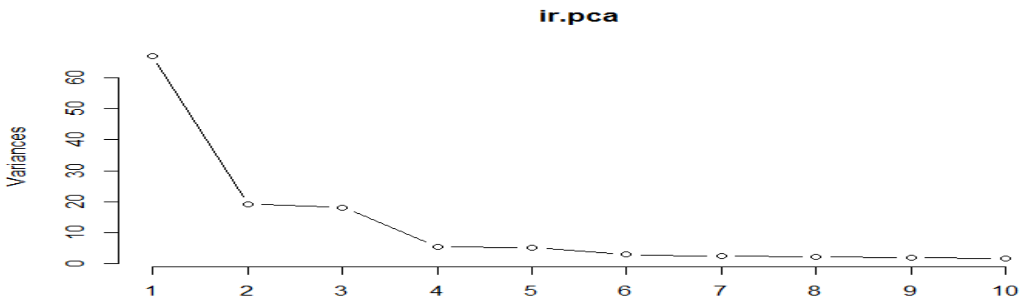
Fig 5: Plot Of clusters from K means.

- Since the data contains large number of variables, we will be applying PCA (dimensionality reduction) method to make the information easier to analyze and visualize



```
> ir.pca <- prcomp(log_checkin, center = TRUE, scale. = TRUE)
> head(ir.pca$rotation)
                        PC1        PC2       PC3        PC4         PC5        PC6         PC7
checkin_info.9.0 0.06577063 0.03273315 0.1236497 0.08937146 -0.07101168 0.06102188 -0.01964329
checkin_info.9.1 0.06348718 0.03386354 0.1224329 0.09146257 -0.06971580 0.05545112 -0.04980402
checkin_info.9.2 0.06375102 0.03432685 0.1221926 0.09158998 -0.06808863 0.05553108 -0.04490221
checkin_info.9.3 0.06321024 0.03468237 0.1240462 0.08890746 -0.07137202 0.06528518 -0.03952241
checkin_info.9.4 0.06582596 0.03533338 0.1214236 0.10119376 -0.06726370 0.09148704 -0.02593272
checkin_info.9.5 0.06804324 0.04361601 0.1096727 0.08744768 -0.06667105 0.15511982  0.06436977
                         PC8         PC9       PC10        PC11         PC12        PC13        PC14
checkin_info.9.0  0.05629749 -0.16945922 0.05429112 -0.02210104  0.007996341 -0.10696699  0.03500690
checkin_info.9.1  0.08975869 -0.18004971 0.06554402 -0.04639797  0.009434919 -0.13171482  0.01979448
checkin_info.9.2  0.07929195 -0.17559755 0.07898198 -0.03907640  0.014500901 -0.11997777  0.03692724
checkin_info.9.3  0.07336365 -0.17524207 0.06882435 -0.04601684  0.017281103 -0.12404650  0.02886836
checkin_info.9.4  0.03498580 -0.15856374 0.06365005 -0.03507300  0.005981178 -0.11873331  0.01707004
checkin_info.9.5 -0.11472978 -0.02629293 0.03445319  0.09253975 -0.007675783  0.03154991 -0.05645145
```

Fig 6: Result Of applying PCA to check-in data



Axis : X – Number of components Y - Variance

Fig 7: After applying PCA to check-in Data and plotting the results

- From the above figure we see, first four principle components explain 85% or greater variation in data. So we will pick these components from all the variables.

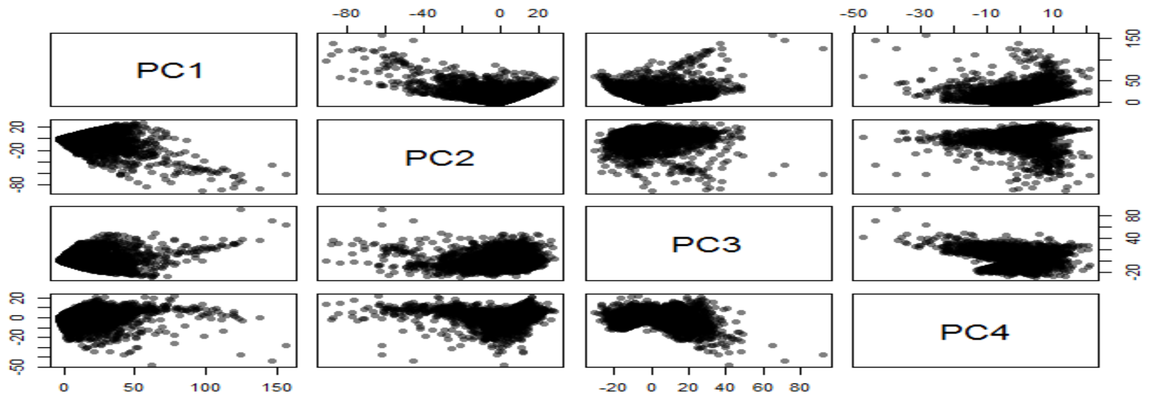


Fig 8: 2-D projections of data which are in a 4-D space.

- Now applying K-means to this reduced dimensional data and plotting the results.

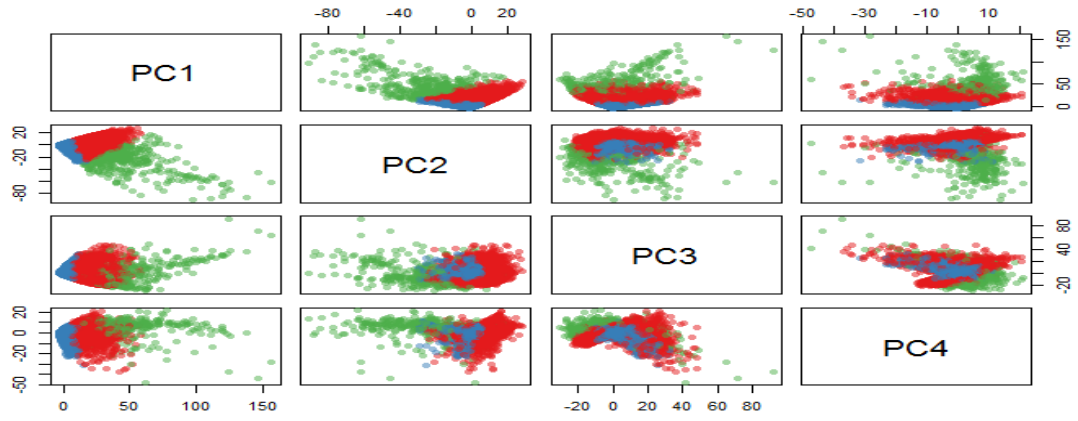


Fig 9: 2-D projections of data which are in a 4-D space after K-means.

Observations

- From above, we conclude that we can divide businesses in Yelp mainly in three clusters based on just check-in information. (yelp_academic_dataset_checkin)
- This implies, some of the categories in yelp are highly correlated.
- The idea being that two breakfast restaurants will get most of their checkins from the morning till noon, while two bars will get most of their checkins during the evening and at night.
- Hence, we expect the correlation between two businesses of the same type to be quite high, while different types of businesses (i.e. a breakfast restaurant and a bar) will result in little or no correlation.

- **What kind of words are part of two extreme star Categories (1 star and 5 star)**

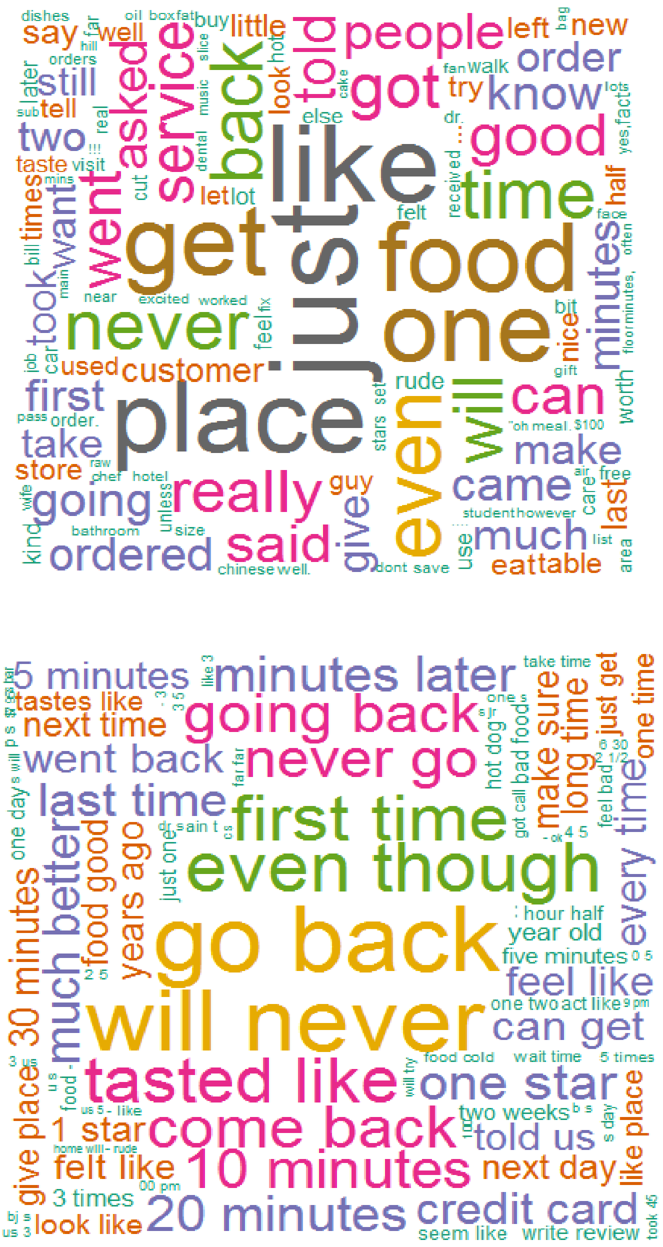


Fig 10: Word clouds for unigram and bi-trigrams for 1 star reviews.

Observations

- We can clearly so many negative sentiment words in these images in just a glance.
- People are clearly not happy with the restaurants where waiting time for food is high, which comes as most prominent in the above figures.

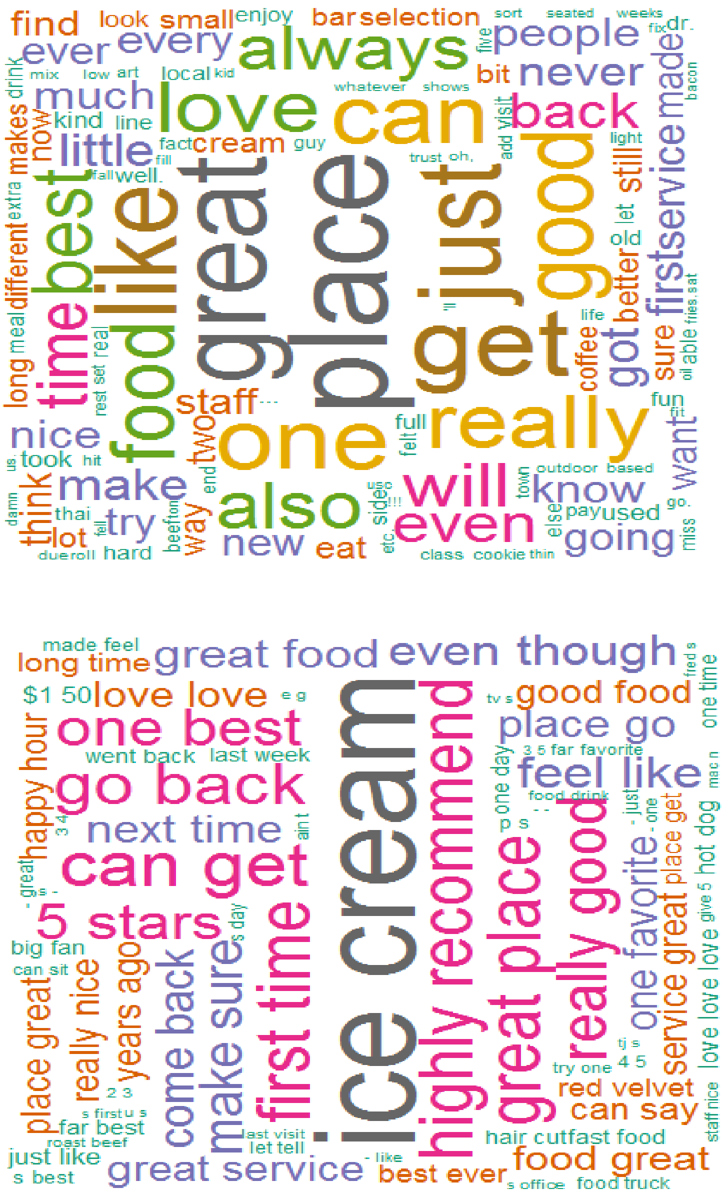


Fig 11: Word clouds for unigram and bi-trigrams for 5 star reviews.

Observations

- In the unigram word cloud, there are so many positive sentiment words like good, best, nice, enjoy, love etc.
- People really want good service and ice cream. As these two just pop out in the bi-tri gram word cloud.

Conclusion

- Naïve Bayes Classifier for predicting primary category of businesses gives 55% Precision.
- We can cluster the categories of businesses in three clusters based on the check-in information using K-means.
- We can observer some interesting patterns in the words used in review text for one and five star reviews.

References

[1] http://www.r-statistics.com/2013/08/k-means-clustering-from-r-in-action/

[2] http://datablend.be/?p=308

[3] https://www.yelp.com/developers/documentation/v2/all_category_list