Name – Anshul Roonwal
Course – CSE 590 Data Science Fundamentals
Submission Date – 31th October 2015 (used 2 slip days)

SBU ID – 110554783
Stony Brook University

Introduction to Yelp Data Set

Yelp offers users high-quality local search and reviews for businesses, in particular restaurants and retail outlets. Users gets involved with the website for searching, writing reviews, rating businesses, connecting with other users and checking in at businesses. This Yelp data set consists of data for the Phoenix, AZ area. This is highly very well structured data with almost no missing values. Amongst the two choices of datasets available, I have chosen the academic dataset having data for reviews, users and business together in a single JSON file. The count of each is given below →

#Reviews	#Business	#Users who reviewed
3,30,071	13,490	1,30,873

Source of data

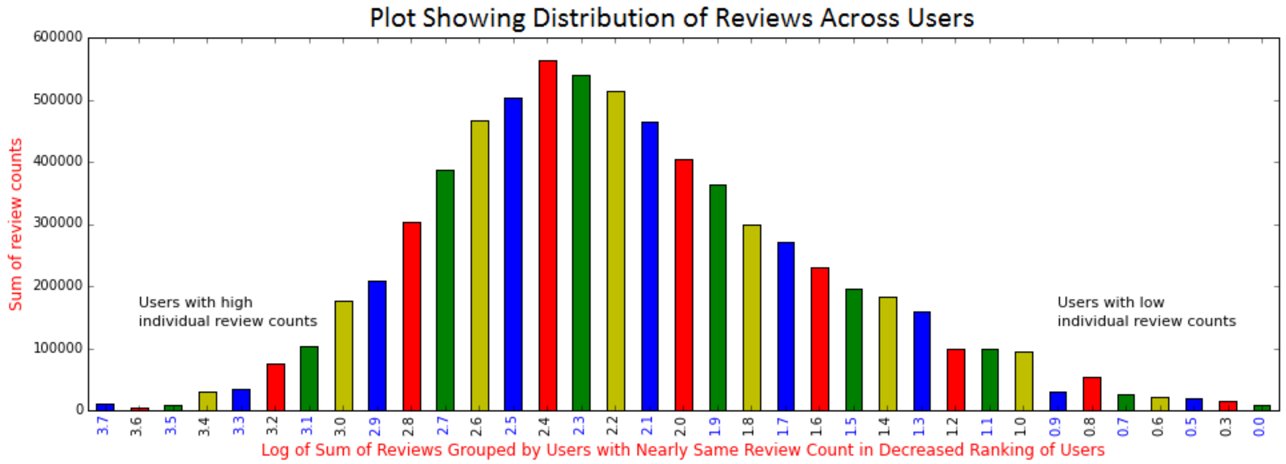
The Yelp data set is made available for academic purposes at-
https://www.yelp.com/academic_dataset

Key Objectives

1) To analyze the distribution of reviews across users and compare it with the distribution of world's richest 1% people.
2) To find what poorly rated restaurants can do better after finding out what their reviewers are taking mostly about using most frequently occurring words.
3) To deduce if Yelp were to omit showing the actual total count of reviews a business got, then can any of the three votes be used to solve the same purpose?
4) To analyze how the rating fluctuates with respect to the length of the review ie; the count of the words in that review and to perform 10 fold cross validation for that model.

1) Analyzing the distribution of reviews across users

Motivation behind this analysis - A very well established fact is wealthiest 1% of the total population holds more than the rest of 99% of the world's economy. Following the same line, I thought of finding out if users who make more reviews (higher in the ranking on x-axis) actually drives the total review count on Yelp.



How did I come up with this plot –

1) Sorted the users by the number of reviews they made on Yelp and ranked them from highest to lowest.
2) Since there are 1, 30,873 users and each user has made different number of reviews ranging between 0 and 5401, here the review count of each users have been represented using log scale on x-axis.
3) To determine which community of users are contributing effectively in the overall review count on Yelp, I first rounded up the log values to one decimal and computed the sum of the review counts of users having same log values (rounded up). Consider the table above for detailed view.

Rank on the basis of review count	User name	Review count of user (R)	Log_{10} R	Round (Log_{10} R)
1	kim n.	5401	3.732474177	3.7
2	Victor G.	5085	3.706290957	3.7
3	Anita L.	3625	3.559308011	3.6
4	Jennifer K.	2975	3.47348697	3.5
5	Tina C.	2883	3.459844642	3.5
6	Rob C.	2752	3.43964843	3.4

Conclusion

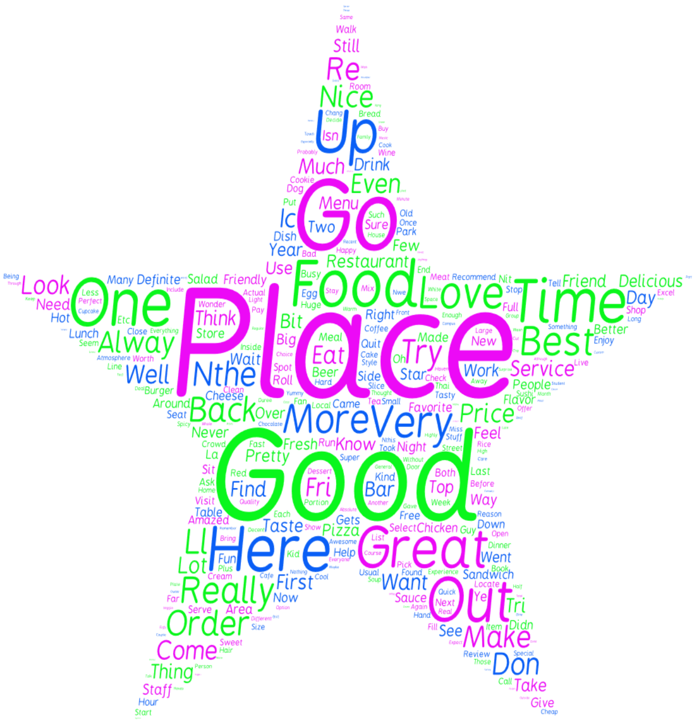
1) The plot clearly shows that what is true for the economy is not applicable here. Here, the users with highest number of reviews are not the ones who accounts for most of the reviews on Yelp. In fact, more contribution of reviews comes from users who are somewhere in the middle of the ranking.
2) The distribution of users, contributing to the review count on Yelp follows the characteristics of a Normal Distribution (if not perfectly Normal) but does not follow a Power law (which is expected in case of economy).

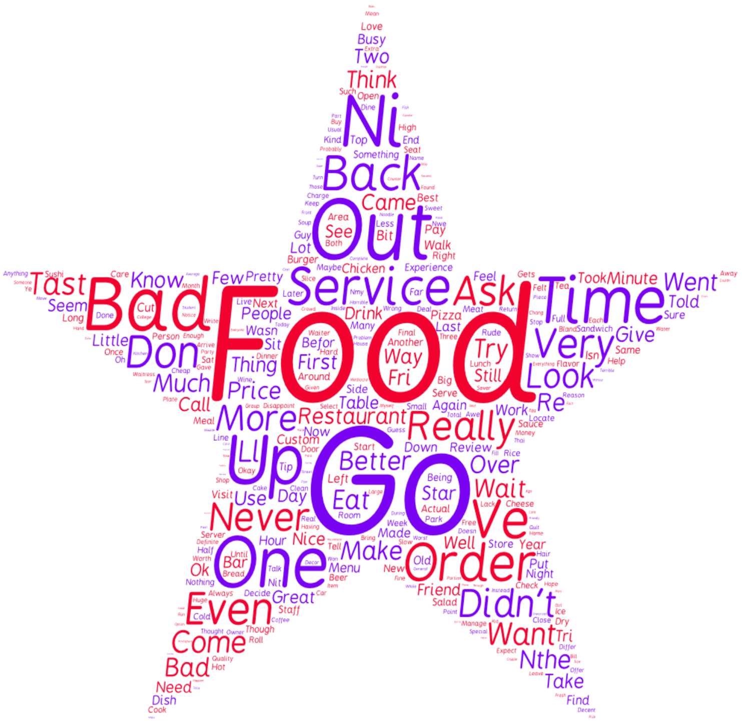
Reference - http://fortune.com/2015/01/19/the-1-will-own-more-than-the-99-by-2016-report-says/

2) Most frequently occurring words in different ratings

Good Rated Businesses – Fig 1 on the left is the word cloud for businesses with **4 or 5** ratings

Badly Rated Businesses – Fig 2 on the right is the word cloud for businesses with **1 or 2** ratings





From the word cloud above, words like Places, Good, More, Great are the ones which are more frequent. On the other hand, from the 1-2 starred word cloud on the right, words like Food, Service, Bad, time, Out, Back are more frequent. If we were to advice the poorly rated businesses to improve their chances of success, a naïve way to do that is to tell them what their customers are caring about are their services, their food and time to service. Improving these factors can probably help them gain more success.

Reference -

https://tagul.com/

3) Replace #Review Count with #Useful, #Funny Or #Cool count?

Briefing Introduction – After a review has been written for a business, other users may vote for that review as *useful*, *funny* and/or *cool.* This new voting concept is yet another example of a feedback mechanisms to businesses. With start of this new voting concept, now a user gets yet another feature on which he can decides on choosing a particular business.

Objective – We are interested in finding ‘how the total count of reviews received by a business, relates to the total count of useful/funny/cool votes received by the total reviews for that business’.

Example – To better understand the objective, here is the examples of top 3 businesses (ranked after the #Reviews).

Note:- It is not surprising to see *‘LA- Westwood’* getting more votes under ‘*useful*’ than #Reviews. They two are different features. So is true for #funny, #cool.

Business_name	***#Reviews***	***#useful***	***#funny***	***#cool***
Diddy Riese Cookies	2634	2106	1751	1954
Sprinkles Cupcakes	1672	1699	1373	1414
Top Dog	1178	912	781	826
	: :	: :	: :	: :
LA - Westwood	265	423	341	341
	: :	: :	: :	: :

Understanding the application of the objective –

Suppose, Yelp wants to omit showing the actual total count of reviews (#Review) a business got and alternatively show these other votes instead, which one to choose as replacement then?. First, why Yelp might want to omit displaying the total review count? This is because providing less data for the user to read without compromising on the purpose enhances the user experience.

Now which of these feature out of these social signal votes should Yelp use as the replacement of #Review? Looking at the values in the table, one can intuitively conclude that #useful displays strong correlation with #Review and should be used as replacement of #useful. But we need to figure that out statistically.

Pearson correlation coefficient –

In order to calculate correlation between two series (data sets), we use Pearson Correlation Coefficient which is given as

$$r = r_{xy} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\ \sqrt{n\sum y_i^2 - (\sum y_i)^2}}.$$

Where, x_1: {x_1,...,x_n} is dataset 1 containing *n* values and $y_{1:}$ {y_1,...,y_n} another dataset containing *n* values. ’r’ then denotes the correlation coefficient. It can range from 0 to 1.

In our case, there are 4 data sets and we want to find one out of [2], [3] or [4] which highly correlates to [1]–

1) **#Reviews** 3) **#Funny**
2) **#Useful** 4) **#Cool**

Results –

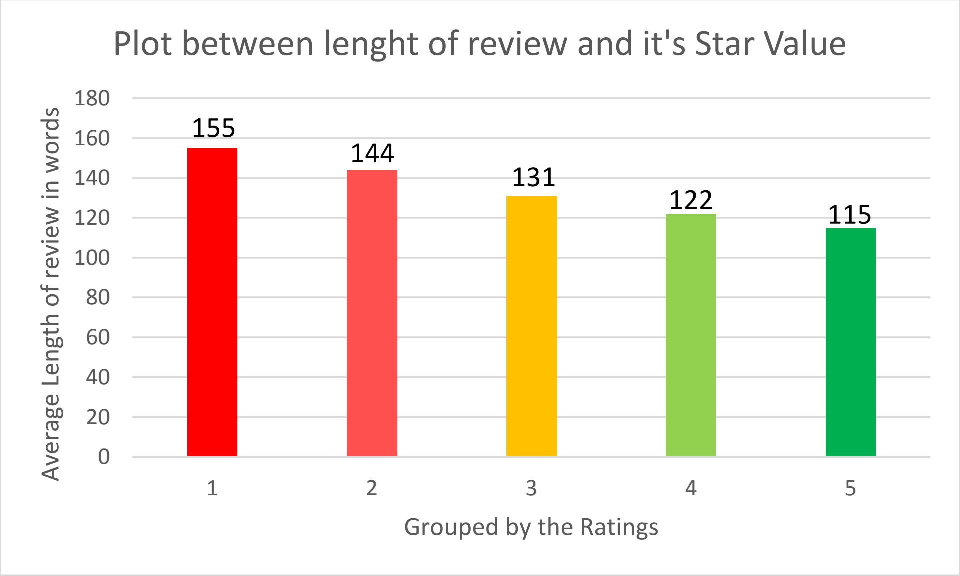
Correlation (#Review, #Useful)	0.92903464
Correlation (#Review, #Funny)	0.943963181
Correlation (#Review, #Cool)	0.962185827

Conclusion –

While intuitively, #Useful looks like an obvious strong candidate for replacement of #Review as the absolute difference between #Review and #Useful is least among the three, the Pearson Correlation Coefficient has revealed that it is actually #Cool that correlates with #Review more. Hence if Yelp were to replace #Review with any of the other votes count, then #Cool should be the choice to be considered first, followed by #Funny and lastly #Useful.

4) When does a reviewer write more? For good ratings or for bad?

Since sentiment analysis is out of scope for this mini project 3, I thought of using the Review 'Text' to determine when it that the user is likely to write long reviews. Is it when he is very satisfied with the business (and provides high ratings) or when he is highly unsatisfied with the business or is there no such pattern?



Inference – Can we jump to the conclusion that users tend to right more verbose (more in length) reviews when they are unsatisfied with the business and thus rated it less?

10 Fold Cross Validation Test –

The above results were taken after taking the whole data set of review at once. How do we test if is not overfitting to the larger dataset that I have not considered (remember I am working on the academic dataset, not the bigger challenge dataset)? We can run 10 fold cross validation test to ensure that the mean square error is minimized in each test.

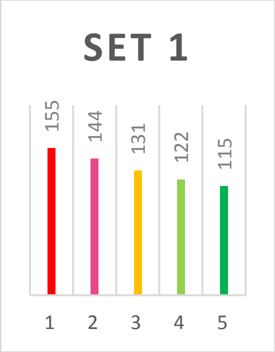
#Reviews = 3,30,071,   We have chosen K =10 for the cross validation test

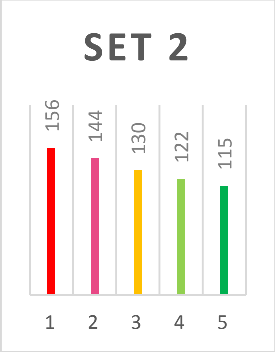
#Reviews in each training set = #Reviews – 10% of data taken at 10% intervals.

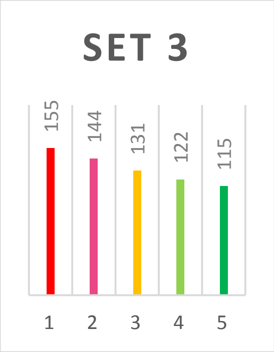
So, 1st Training data set will consider Reviews from → 3,30,071 – Reviews in range (0 to 33,000) Set 1
2nd Training data set will consider Reviews from → 3,30,071 – Reviews in range (33,000 to 66,000) Set 2
3rd Training data set will consider Reviews from → 3,30,071 – Reviews in range (66,000 to 99,000) Set 3

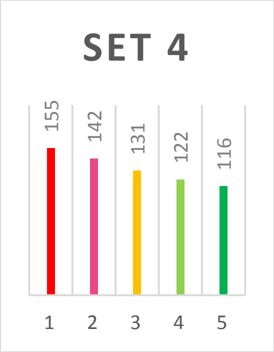
& so on…

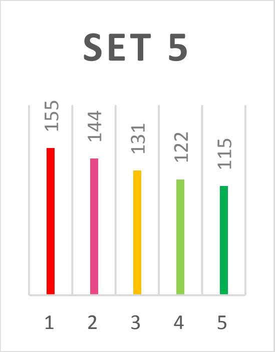
Here are the results for all those 10 test

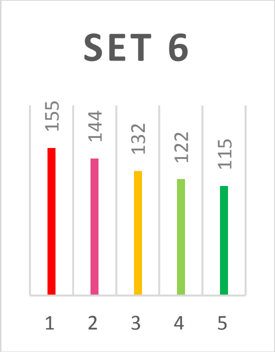


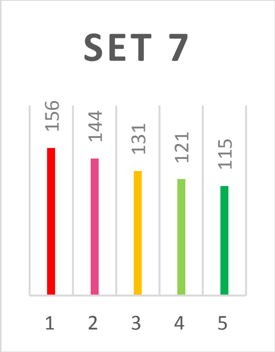


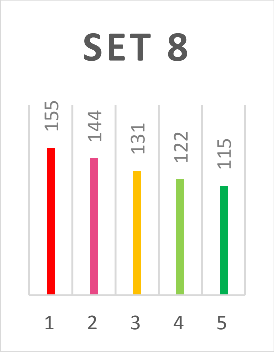


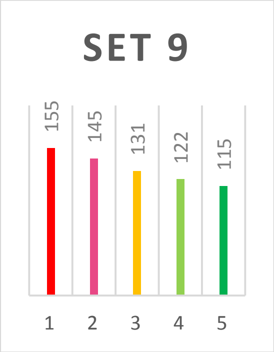


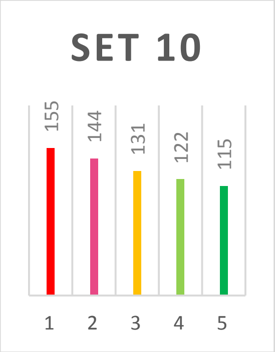












Conclusion

Looking at the plots for the result of 10-fold cross validation test, we can say with confidence that there is not much variation in the test results. Thus our inference from above that- 'users tend to right more verbose reviews when they are unsatisfied with the business and thus rated it less' will still hold true. We have cross verified it by taking 10 different samples from the main training set. Also, we can infer that there wouldn't be a problem of overfitting and that this conclusion will stay true for the main challenge dataset as well.