Predicting Success of a Business By Identifying Influential User Groups Using Yelp Data

Vishrant Vasavada (s1672896)

Another Team Member:
David Hodges (s1603605)

School of Informatics
University of Edinburgh

Instructor: Prof. Rik Sarkar

Introduction

B2C businesses have to rely a lot on the evaluation and prediction of customers' feedback for their success. Yelp has a "star rating" system for businesses as well as review text, which generates the humongous amount of user data using which a lot of research questions could be answered. In this project, we use the Yelp Dataset to identify the influential users to which the emerging businesses should pay more attention since we believe that it is the major factor which define the rating trajectory for the business.

Problem Formulation, Statement and Overview of Results

The focus in this project was laid on identifying possible influential user groups as "elite" users, socially well-connected users and those who write reviews which reflect what many find useful. For the last category, we do the analysis of the review texts to identify important keywords, which should reflect important factors in business quality (e.g. food, services, atmosphere). Social Network analysis of Yelp users is done by creating a "friend network" and comparing the review sentiments of the related users to check for the similarities. The two users are related if they are friends, or they both have reviewed the same business. We believe that this study could prove useful to any business to know which reviews to pay more attention to depending on the users writing them.

In the following sections, we briefly explain what all experiments we performed, the results we got and the conclusions. Not all results that we got were as expected. For example, in our analysis, we noted that no sharp relation is seen in the plot of the number of fans and average review votes for the users. Also, in our scatter plot between average review votes and the percentage of important keywords in the review, we didn't get non-decreasing curve as expected. The scatter plot shows that there are many reviews using a high percentage of important keywords but still fail to receive high votes. The scatter plot looked similar even when we considered only the elite users. Although, we observe that the elite users have a high number of fans and also has a high number of average votes per review. This leads us to a conclusion that users using important keywords in their reviews are not necessarily influential users but the elite users are. While socially well-connected normal user's review would be less viewed and given importance than that of the elite user, their opinions could still affect the network because of "small world" phenomenon observed in Yelp friend network. This is because two friends are likely to have same sentiments for same restaurant.

Dataset and related work

We used Yelp Dataset Challenge data as our primary dataset. We used business, user and review data from the dataset for restaurants in Edinburgh. We only use restaurants because they have far more reviews on Yelp than other types of business. We also prune the data to consider the restaurants only in the Edinburgh city since the total data is too large to work with because of the time constraints. On initial analysis of data, we observe that the dataset follows a Power Law degree distribution quite well (figure 1). This is expected because, like any other social media, we would expect the vote count, friend count, and review count to exhibit preferential attachment behavior.

The effects of influential Yelp users on new businesses has previously been studied[1] and the rating trajectory modeled[2]. We expand on this work by considering additional data (eg: review text) when identifying sets of influential users.

[1] *Mathur, Harmelen, Gupta, Discovering Emerging Businesses.*
[2] *Jaisinghani, Todi, Liu, Modeling Growth and Decline of Businesses in Yelp Network.*

Approach and Results

We used Python's RAKE library for the keyword extraction from the reviews. We sorted the list of keywords we got based on their frequency and took the first half of the list of high-frequency or popular keywords. The top 10 keywords in this list were: *food, place, good, time, back, edinburgh, service, menu, friendly, great*. For all the reviews using these high-frequency keywords, we checked the votes they got. We expected a non-decreasing plot since we expected reviews using important keywords to be more useful to the users.

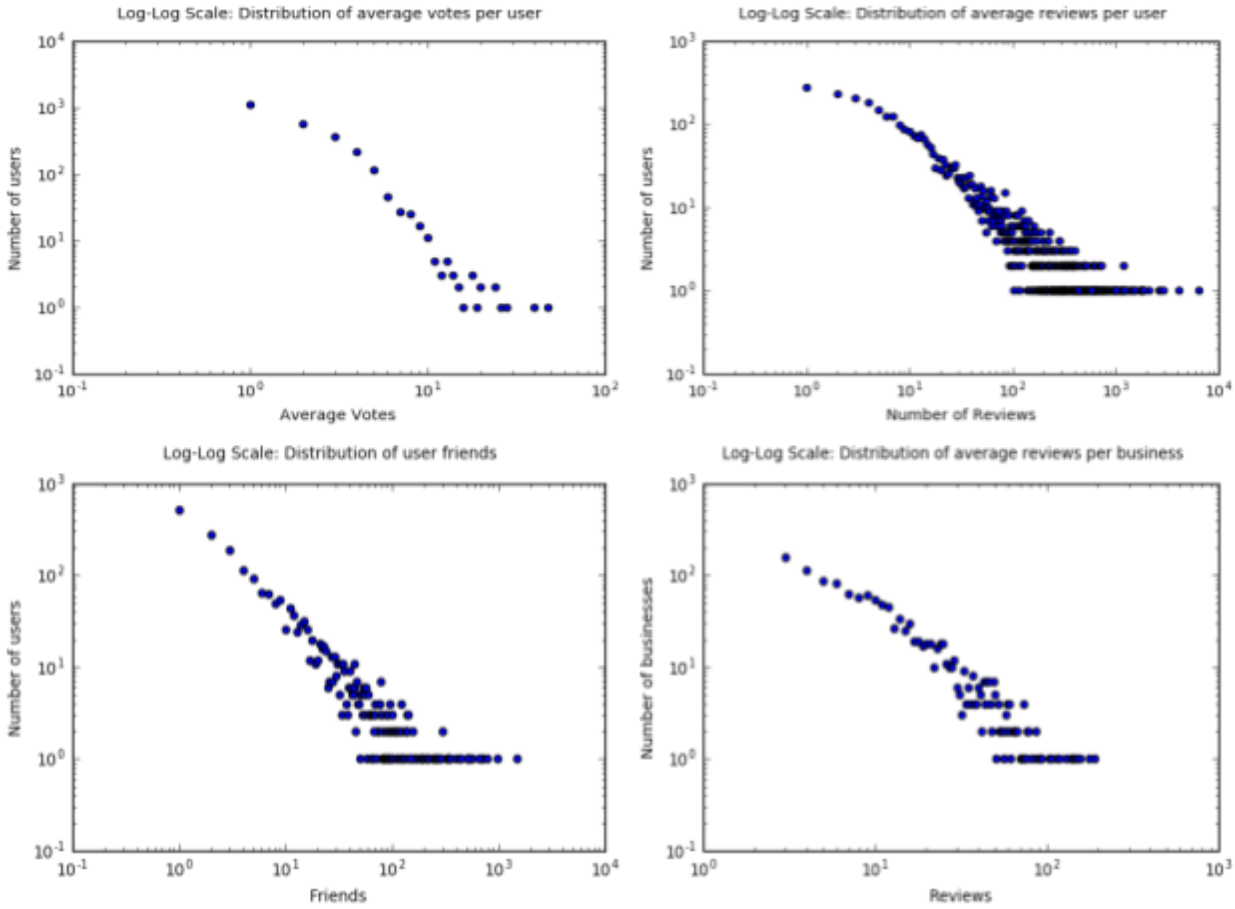

Figure 1. Power Law Distribution of Yelp Dataset

However, the real outcome was very different and as shown in the figure 2. A similar plot was obtained when we considered only the elite users which led us to believe that the user using high percentage frequency of these keywords in their review text doesn't really create a strong influence on users.

Our next step was to study how impactful the elite and socially well-connected users are. To study socially well-connected users, we created a "friend network". Nodes in this network represented users and the edges between two nodes represented the friendship between users. Below we summarize the basic statistics of this network (table 1).

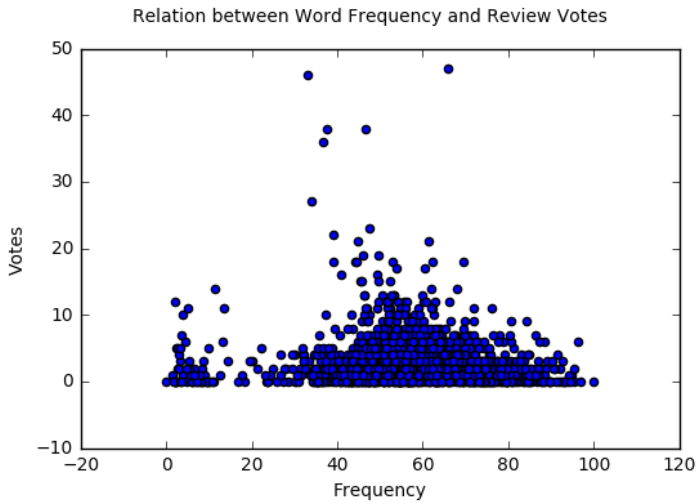

Figure 2. Review Votes and Percentage Frequency of Important Keywords in Review Text

Nodes	3828
Edges	7109
Avg. Clustering Coefficient	0.597
Avg. Path Path	2.945
Diameter	8

Table 1. Yelp Friendship Network General Statistics

Using the number of edges in this graph as expected number of edges for the Erdos-Renyi graph, we get average clustering coefficient and network diameter as 0.001 and 14, respectively, which when compared to that of the Yelp Friendship network, leads us to a conclusion that Yelp Friendship Network follows a small world phenomenon.

We then modified our friend network so that the edge now exists between the two user nodes if the users are friends and they have also reviewed the same business. This graph had 169 edges. For all the connected users, we measured the review sentiment polarity. All the co-reviewers turned out to have same sentiments for a restaurant except for a couple. We can thus conclude that the friends tend to have similar tastes and sentiments for the restaurants. Since the friendship network follows small world phenomenon, the opinions about particular restaurant given by a few socially well-connected people would tend to spread fast over the network. This group of people could be believed as an influential group since they will have an indirect impact on the business star ratings.

We again modified our friend network to include only the users who have been elite for at least 1 year. The edge in this graph exists between two elite user nodes if their review sentiments for the same business are different - that is, one has positive polarity while the other has negative polarity. This graph consisted of 123 nodes and 104 edges. The graph density is 0.014 which denotes that it is sparsely distributed. Hence, most of the elite users have the same sentiments for the same business.

In our analysis, we also noticed that a high fan count doesn't necessarily mean high average review votes for the user whether or not they have elite status. However, the users with elite status for many number of years tend to have a high number of fans and also a high number of average votes per review. We believe that this is because the users who have been elite for quite a few years might be more popular in the network than the other users, including the ones who have been elite for just a couple of times. Users tend to give weight to their reviews and vote them more than any other reviews. Hence, the users who have maintained their elite status for many years are the most influential users in the network. The histogram falls sharply for users who have been elite for 11 and 12 years. We believe that the Yelp was very new in 2004/5 with a small user base to get any fan support and review votes. As Yelp got popular, the user base expanded and the users might have got aware about what "elite" status actually is and how significant it is, leading to a sudden rise in fan followers of elite users and also more attention (votes) to their reviews.

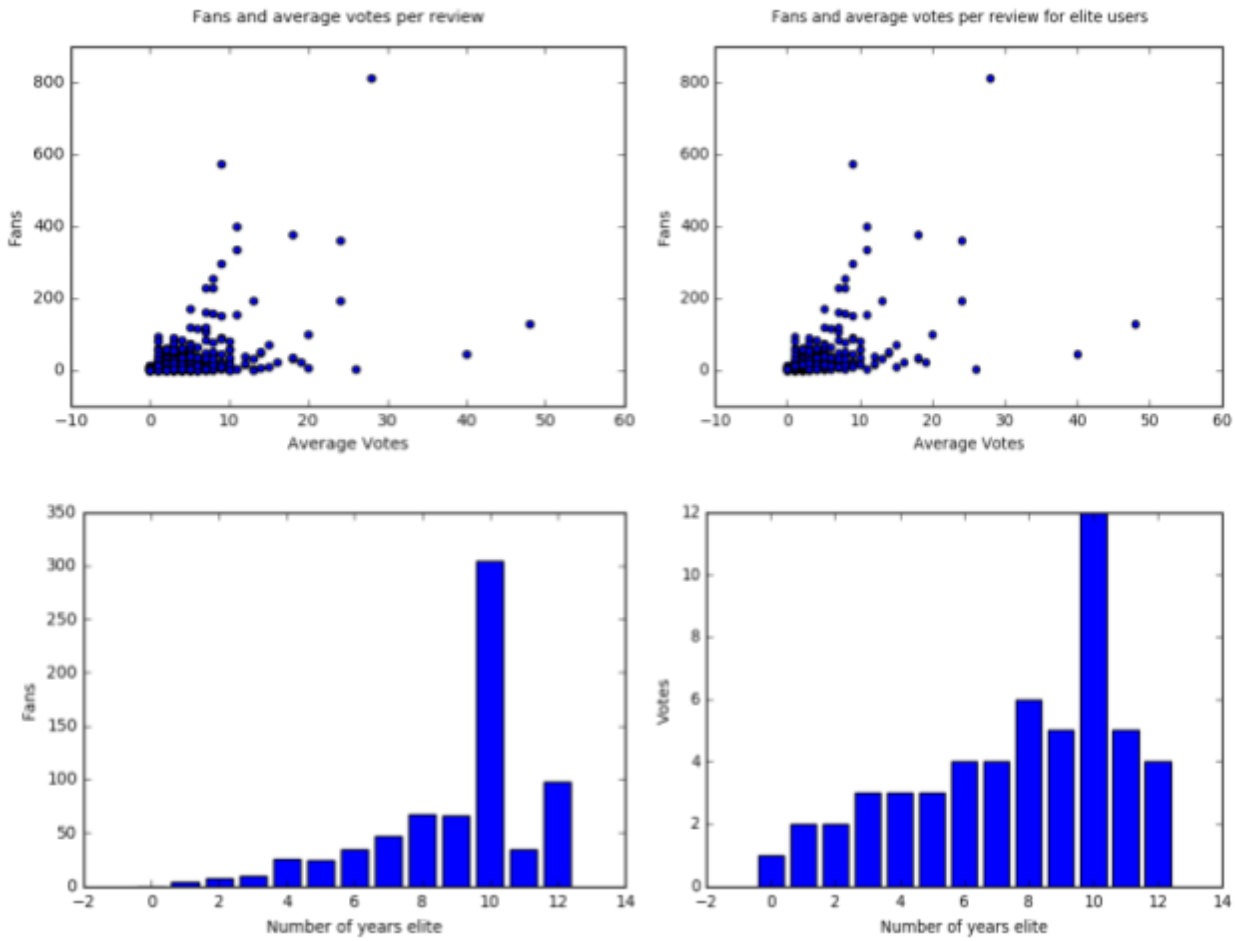

Figure 3. Plots between 1) User fan count and average review votes, 2) Elite user fan count and average review votes, 3) Number of Years the user has been elite and fan count, and 4) Number of Years the user has been elite and average review votes

Conclusions and Future Work

From the analysis of Yelp Dataset, we conclude that the popular users, that is, the users who have been elite for many years are the most influential. Reviews with a high percentage of popular keywords don't necessarily get high votes. It might be easy to measure the sentiment polarity of the review by extracting the important keywords and to predict the star ratings from it, but it is definitely hard to identify the users who write reviews that others find useful. Maybe, considering restaurant categories individually could give us better results because general keyword extraction might miss out few important keywords like "good pizza crust" or "lovely pasta" for Pizzeria or Italian Restaurants.