Info 290t: Data Mining

Homework 3 | Question 8 | Research Write Up

Steve Lee | Seongtaek Lim | Bradford Malt | Michael Sloan-Rossiter

Research Question

We hypothesize that increased user-to-user engagement may induce users to post more frequently with more verbose reviews. Specifically, we ask what clusters of 'user types' exist are associated with different levels of user-to-user engagement and the level of user activity?

Data Set

Our research utilizes the Yelp Review dataset of ~45,000 Yelp reviewers in the Phoenix, AZ market. This research uses the yelp_reviewers.txt file, in which each reviewer (user) is represented as a tuple (row).

Specifically, we evaluated the following data fields:

- Q3, Total number of reviews: provided data field which indicates the number of reviews created by the reviewer
- Q4/Q3, Cool votes per review: a constructed data field which takes the total number of cool votes for the reviewer and divides by number of reviews
- Q5/Q3, Funny votes per review: a constructed data field which takes the total number of funny votes for the reviewer and divides by number of reviews
- Q6/Q3, Useful votes per review: a constructed data field which takes the total number of useful votes for the reviewer and divides by number of reviews
- Q18_group7, Average number of days between two reviews: a constructed data field

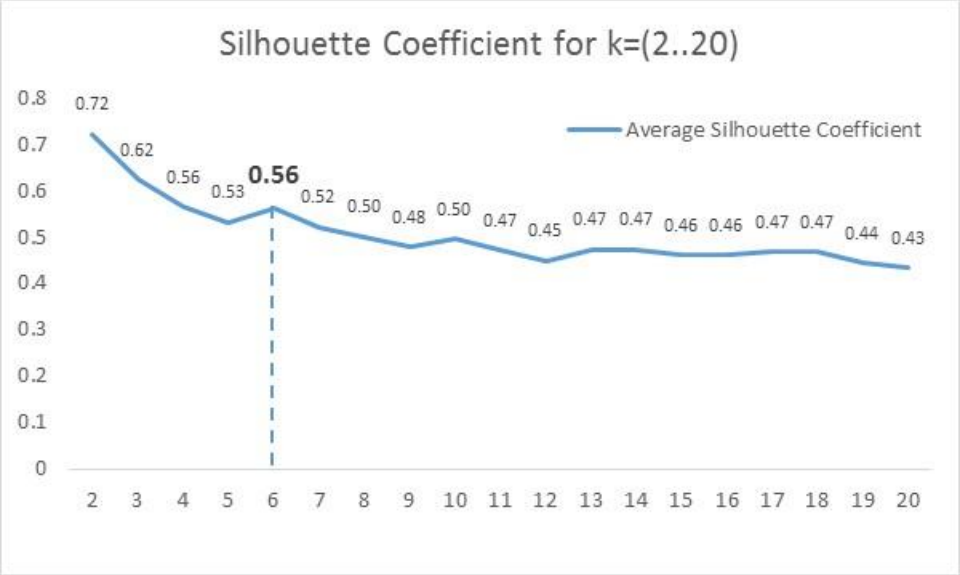
Approach

We hypothesize that increased user-to-user engagement may induce users to post more frequently with more verbose reviews.

To evaluate this hypothesis, we take a k-means clustering approach on our dataset, which covers 44,993 unique reviewers across 6 data fields.

First, to select the optimal k value (number of clusters) we iteratively performed K-means for values of k between 2 and 20. We ran the k-means function provided by the Sklearn Python library. We selected the optimal k value based on evaluation of the average silhouette coefficient under each clustering approach. To calculate average silhouette coefficient at each value of k, we chose to include a sample of 4,000 data points. As shown below, k=6 is a strong candidate for optimal k value because it locally maximizes the silhouette coefficient before the marginal benefit of adding k falls dramatically. This

result (local maximum at k=6) was consistent across multiple simulations and alternative sample sets of the data.



Results

At k=6, the k-means simulation returns six clusters of reviewers. The six cluster centroids are described in the table below:

Cluster	Total reviews	Cool votes per review	Funny votes per review	Useful votes per review	Average characters per review	Days between reviews
1	7.60	0.34	0.26	0.74	121.21	35.11
2	1.10	0.42	0.36	1.01	986.85	9.51
3	1.00	0.83	0.92	2.20	3524.03	0.00
4	1.37	0.33	0.25	0.79	464.89	12.23
5	2.53	0.36	0.28	0.84	247.79	511.76
6	1.02	0.59	0.54	1.36	1848.48	5.38

Cluster 1 is dominated by the number of reviews, however shows low average statistics for cool votes, funny votes and useful votes. It is also on the low range of characters per review and on the high end of days between reviews. This cluster generally seems to be Yelp customers who have been around a long time and write a decent number of times on yelp, but write short reviews that are not universally thought of as useful, funny, or cool. These customers are helpful in driving up the absolute number of reviews for Yelp, but do not add much substance.

The rest of the clusters show low total reviews, however vary in how they are clusters based off of the remaining features. Cluster 3 are the one time reviewers, as they have no days between reviews,

however first time reviewers receive the highest useful votes. This could show that people tend to put a larger amount of effort into their first reviews. This is also shown through the length of their first post, which is almost double the second cluster's average character per review (3524 vs 1848).

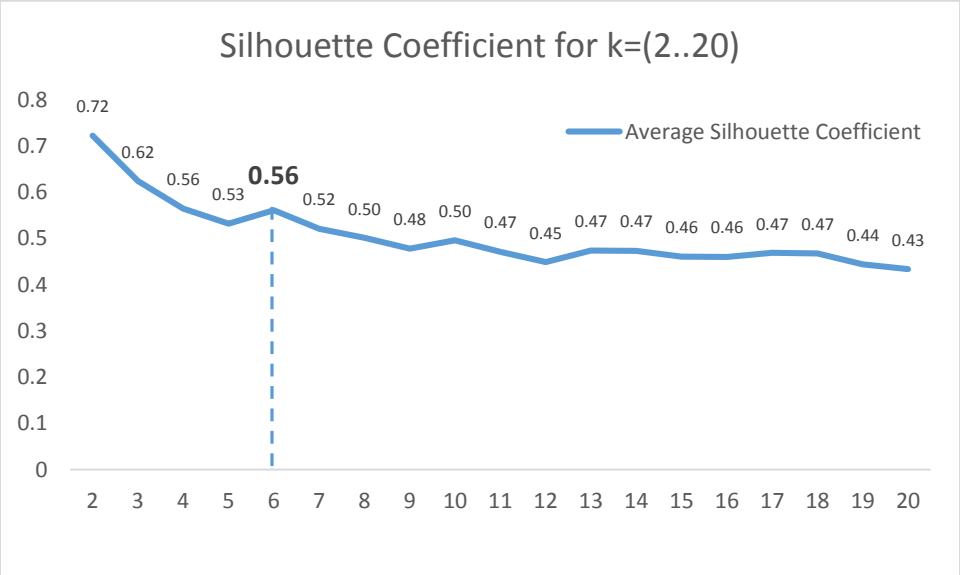
Cluster 2 and 4 are relatively similar, however cluster 2 seems to be slightly better customers for Yelp. They review slightly more often, their reviews are about twice as long, and they receive slightly higher useful votes per review. It seems between all the clusters that there is a relationship between the length of the review and how many useful votes it receives.

Cluster 5 are the long time but infrequent users. They have been with Yelp for an average of over 3.5 years, however are not power users of the system. They receive low scores across all categories and write short reviews. Something must provoke these customers to go on and write a review every few years.

Cluster 6 is most likely newer customers that have reviewed recently and write long reviews. As we noted, there appears to be a relationship with length of reviews and the number of useful votes a review receives, so cluster 6 has higher than average useful votes. This cluster is most likely close to cluster 3. Cluster 3 reviewers (first time reviewers) could fall into Cluster 6 as they mature as Yelp users.

Exhibits

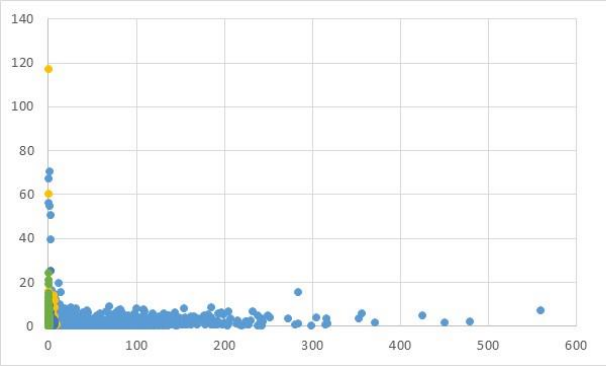
Exhibit A. Average silhouette coefficient with at values of k, elbow method indicates that k=6 is 'best'



- Ran for k=2 through k=20
- Sample size = 4,000
- Ran multiple times and consistently shows k=6 is improvement over 5 and 7 but the trend after k=7 is of fairly predictable degradation (little marginal benefit to increasing k)

Exhibit B. Representation of clusters in 2D space: total reviews vs. average funny votes per review

Cool votes per review on y-axis, total reviews on x-axis. Second graph restricts x-axis to 0-20 range to highlight differences between clusters among lower frequency reviewers.



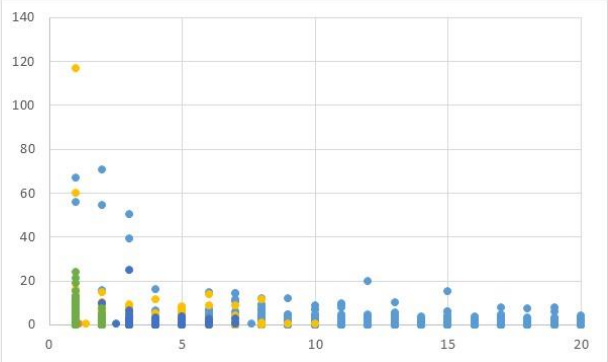


Exhibit C. Representation of clusters in 2D space: days between reviews vs. average useful votes per review

Days between reviews on y-axis, useful votes per review on x-axis. Second graph restricts x-axis to 0-80 and y-axis to 0-600 ranges to highlight differences between clusters among lower frequency reviewers.