Yuxin Jin

CS 199

Miodrag Potkonjak

June 1, 2014

**Finding Influential Users from Movie Reviews**

The goal of this project is to spot influential users whose reviews have strong impact on other reviewers. The data set came from Amazon movie and TV reviews. Analysis focused on scores users rated movies and helpfulness of their reviews voted by other users. Scores and helpfulness are compared injunction with similarity between pairs of users. The hypothesis is that if a user is influential, its reviews should have high helpfulness; other users reviewing the same movies user should generally have small differences in scores from those of the influential user.

The original data set has 7,850,072 reviews. Each entry included information of product ID, product title, price, user ID, user ID, user profile name, review helpfulness, review scores, review time, review summary and actual content. Here is a sample of the reviews:

product/productId: B0001Z3TLQ

product/title: By the Sea [VHS]

product/price: unknown

review/userId: A3421LTBSWSPXK

review/profileName: KML

review/helpfulness: 5/6

review/score: 4.0

review/time: 1089417600

review/summary: A romantic zen baseball comedy

review/text: When you hear folks say that they don't make 'em like that anymore, they might be talking about “BY THE SEA”; This is a very cool story about a young Cuban girl searching for idenity who stumbles into a coastal resort kitchen gig with a zen motorcycle maintenance man, three hysterical Italian chefs and a Latino fireballing right handed pitcher who plays on the team sponsored by the resort's owner. As is often the case she 'finds' herself through honest, often comical but always emotional, interaction with this sizzling roster of players. With the perfect mix of special effects, that salsa sound and flashbacks, BY THE SEA, gets 4 BIG stars from me!

Since only users with large number of reviews and movies with plenty of feedbacks are interesting for the purpose of this project. The criteria for narrowing down the perimeter were a user should have at least reviews and a product should have at least 500 feedbacks. Only reviews that satisfy both thresholds were selected. The data used in the analysis has 2,036,011 review entries by 997 users on 121,632 movies. MATLAB is a primary tool for plotting and analysis in this project. Thus, in preprocessed .csv file, only product ID, user ID, review helpfulness, review score, and timestamp were included. Product ID and user ID were mapped to pure integers.

The similarity metrics of two users were based on Jaccard indexes. The formula is

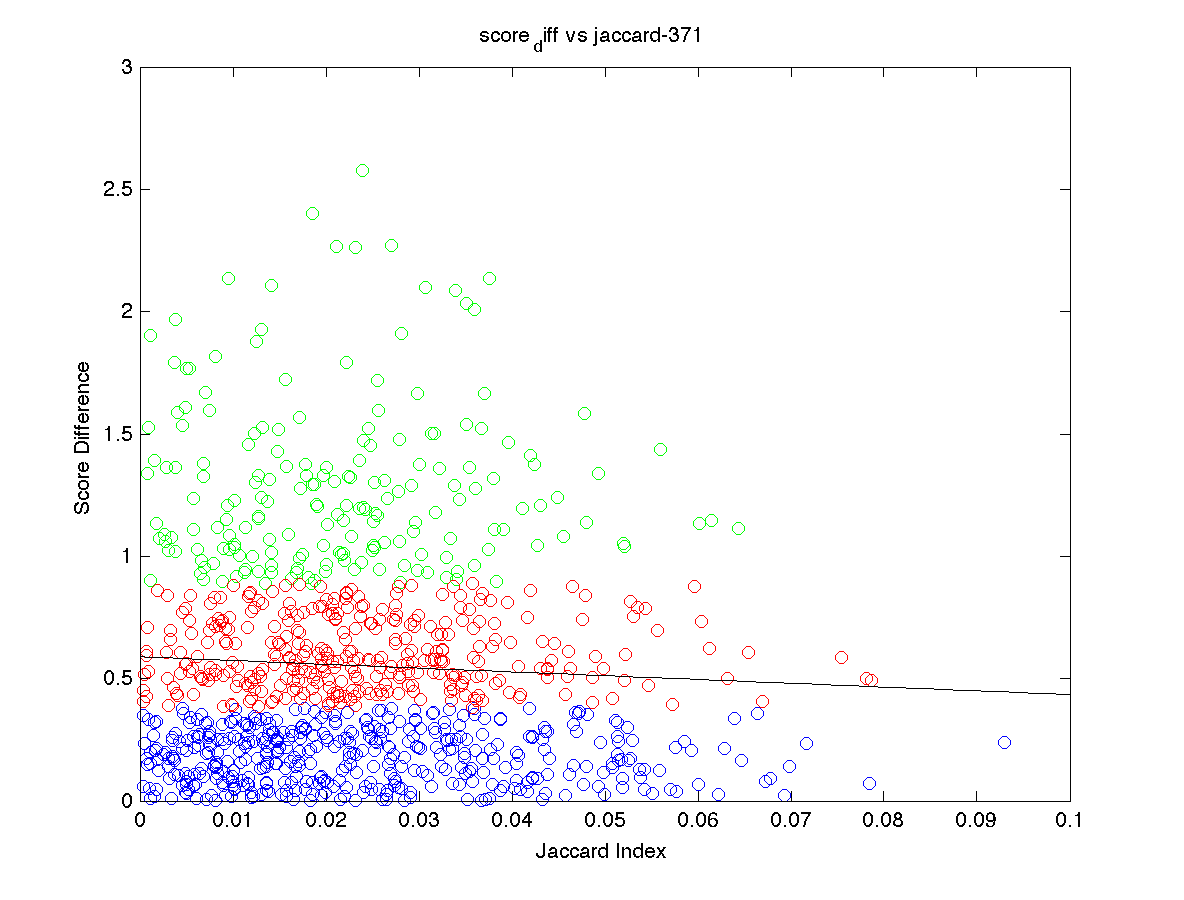
In this case, set A is the movies watched by one user and set B by another. Take the intersection and union set A and B. Divide the number of same movies the pair of users reviewed by the total number of unique movies the pair reviewed. The range of jaccard index would be between 0 and 1 inclusive. However, jaccard index equals 1 only if two users are the same person in this data set. It is highly unlikely that two users reviewed the exact same movies for 400 or more times. Jaccard index of 0 is discarded since there is so similarity between the two users and that produces no useful information about influence. The next step is to select a small group of users that are potentially influential. Later, compute jaccard index of the rest of the users to individuals in the chosen group. Plot score difference, average helpfulness, and other variables against similarity represented by jaccard index to figure out if the individual user is influential.

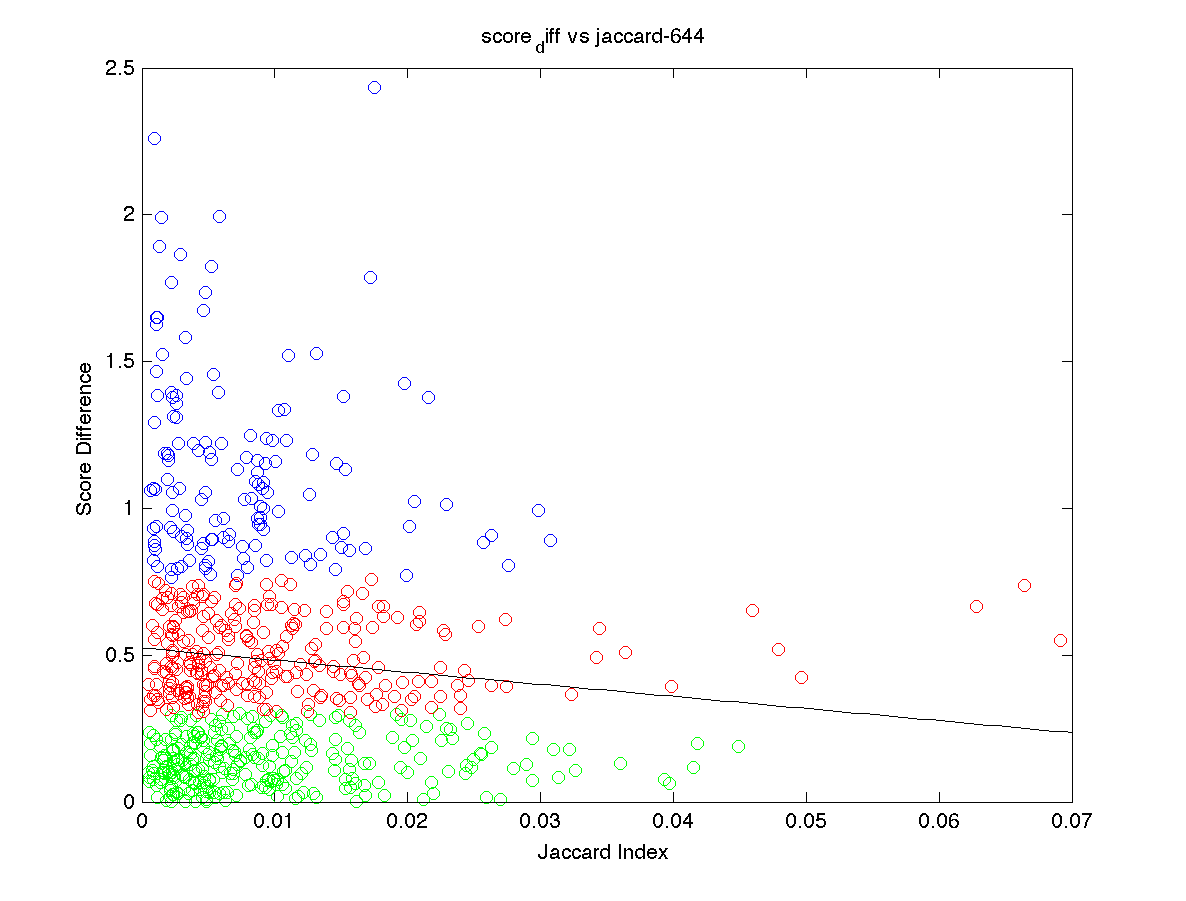
First, need to collect some basic statistics on the data to determine the set of criteria for narrowing potential influential users. Majority of users reviewed between 400 to 1,000 movies. The user that reviewed most number of movies has 1,052,286 reviews with an average helpfulness above 0.5, not including reviews that had no helpfulness rating. However, the average helpfulness of all reviews that have helpfulness rating is 0.6144, higher than that of the user with the most reviews. The average score given by all reviews to movies is 4.0251. There are 462,176 out of 2,036,011 reviews not rated by any other users. That is 22% of the reviews. Thus, the marks used to select potential impactful users are

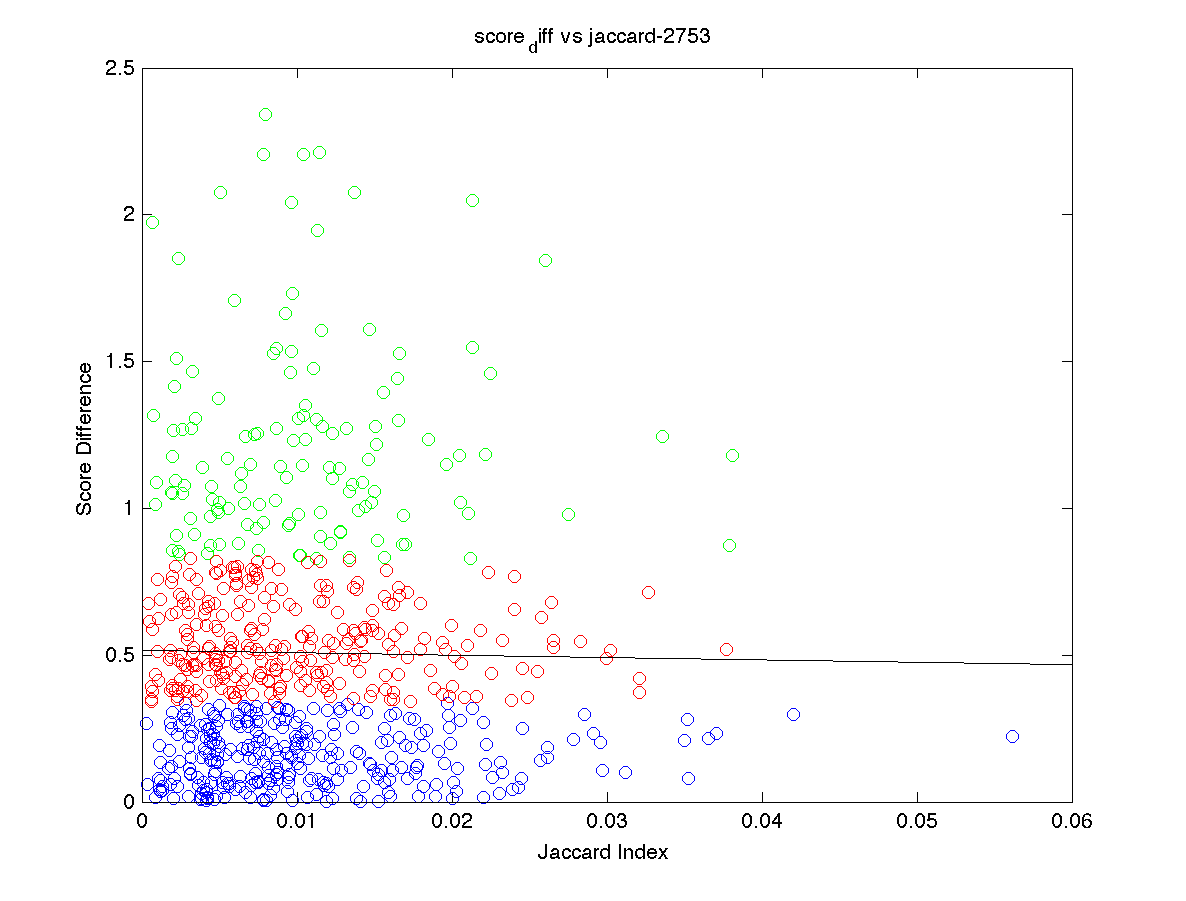
* More than 95% of its reviews had helpfulness feedbacks;
* Average helpfulness of its reviews that had 5 or more feedbacks is larger than 0.9.

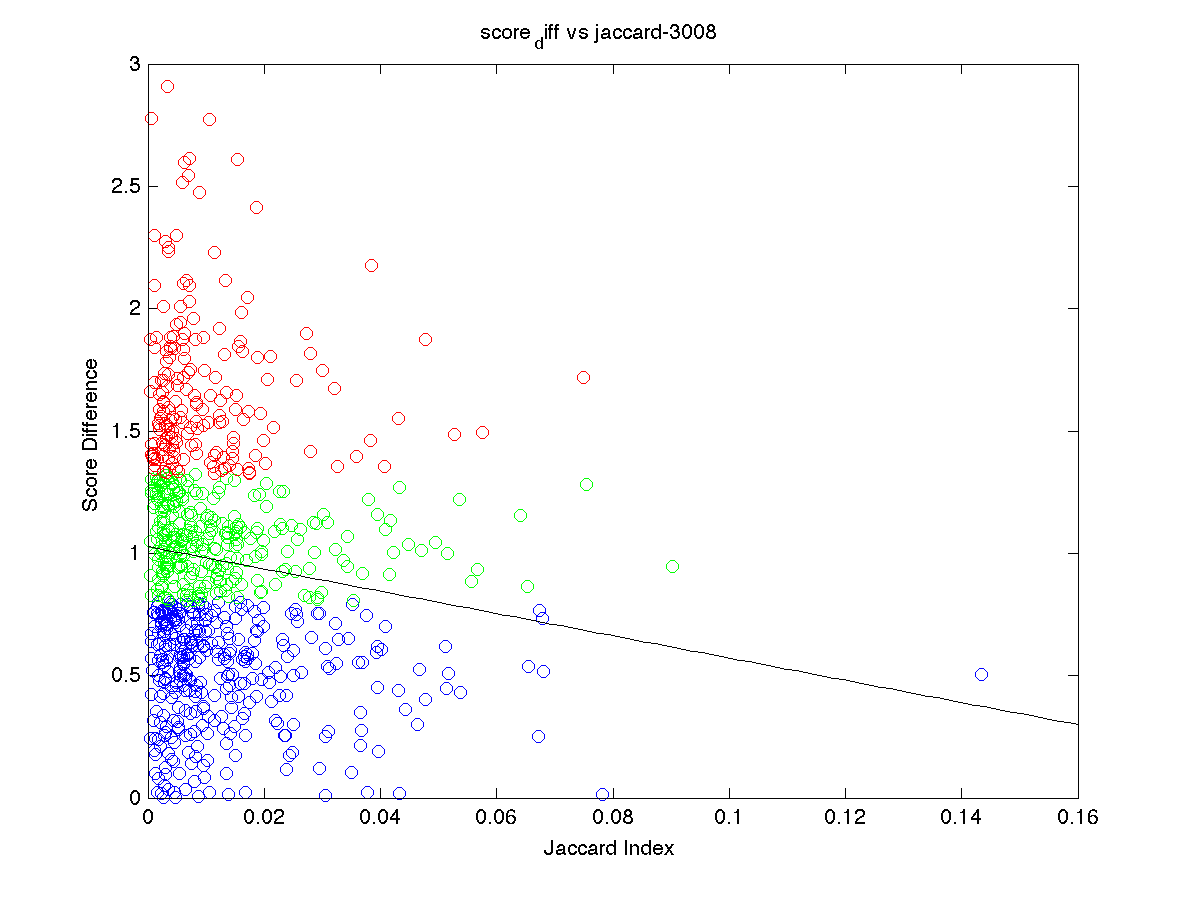
The first criterion shows that many other users read the reviews. The second condition ensures that most reviews are high quality.

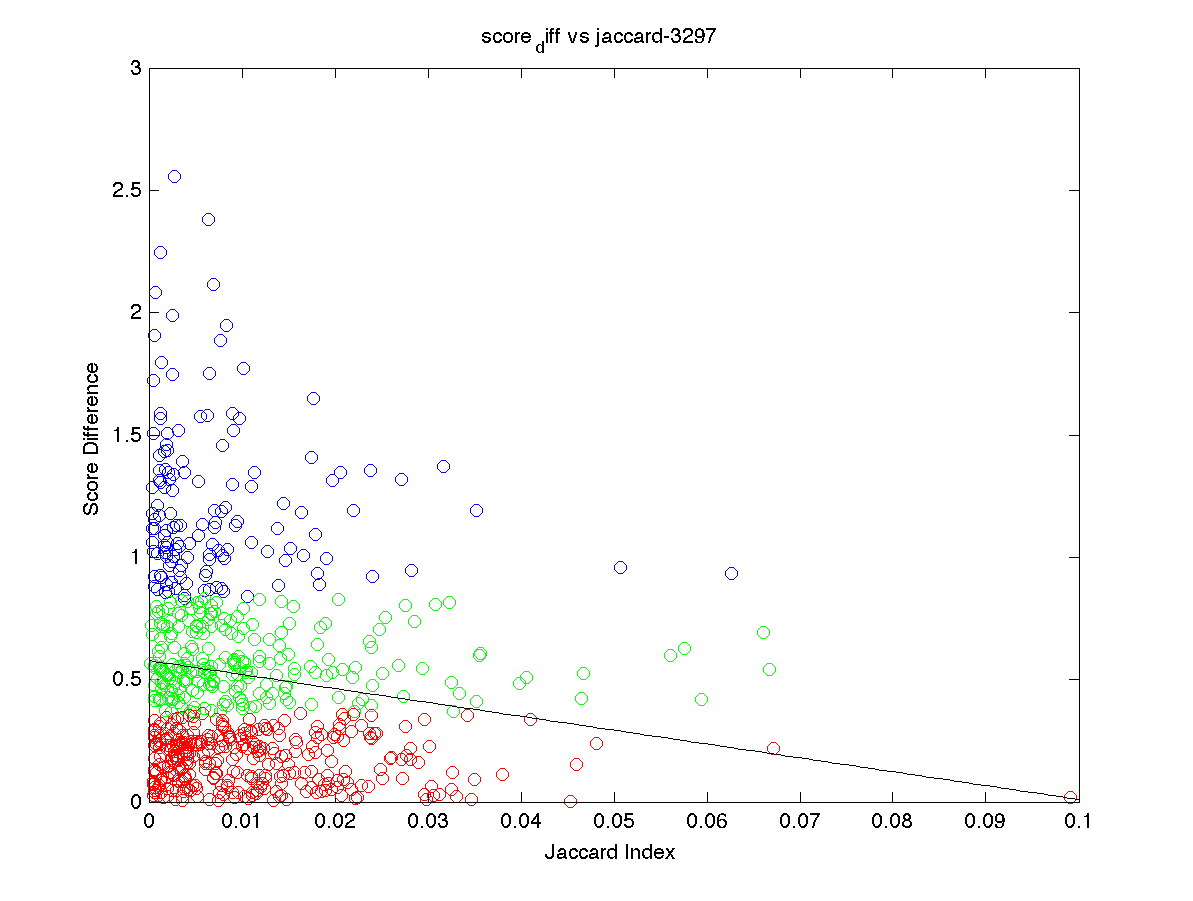
After processing with those restrictions, 8 candidates emerged. Their mapped UIDs are 371, 644, 2753, 3008, 3297, 8188, 19306, 43949. Plot score differences between the chosen user and the rest of the users vs. their jaccard indexes for each individual chosen users. There are obtained graphs:

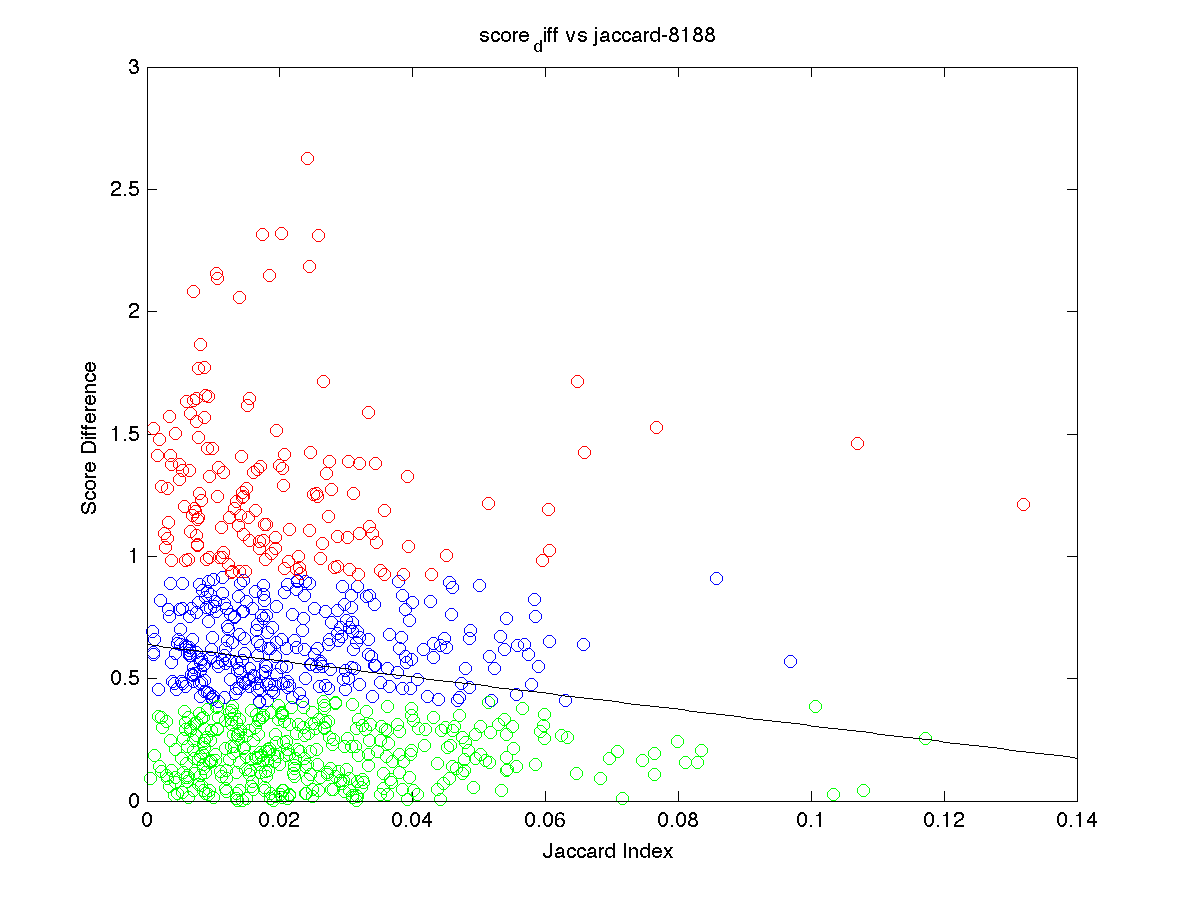


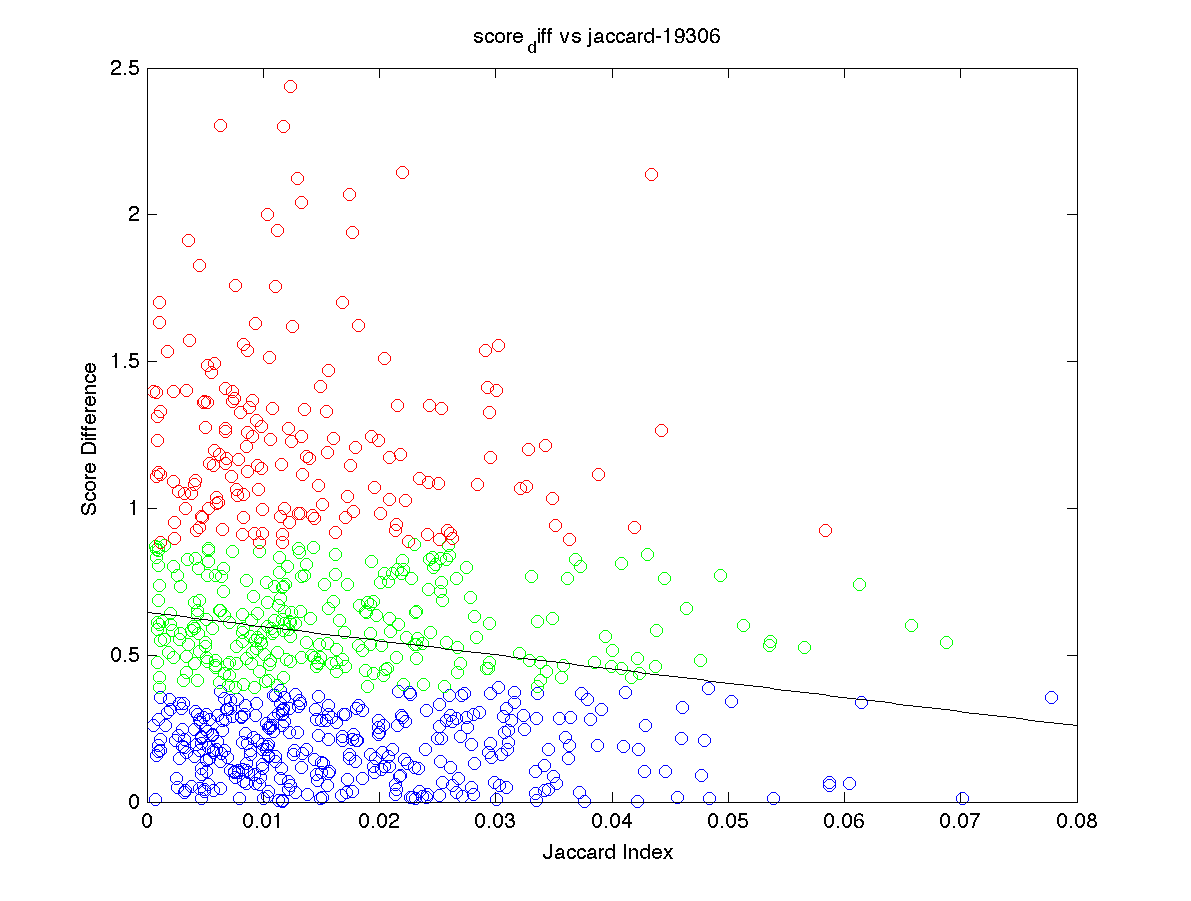


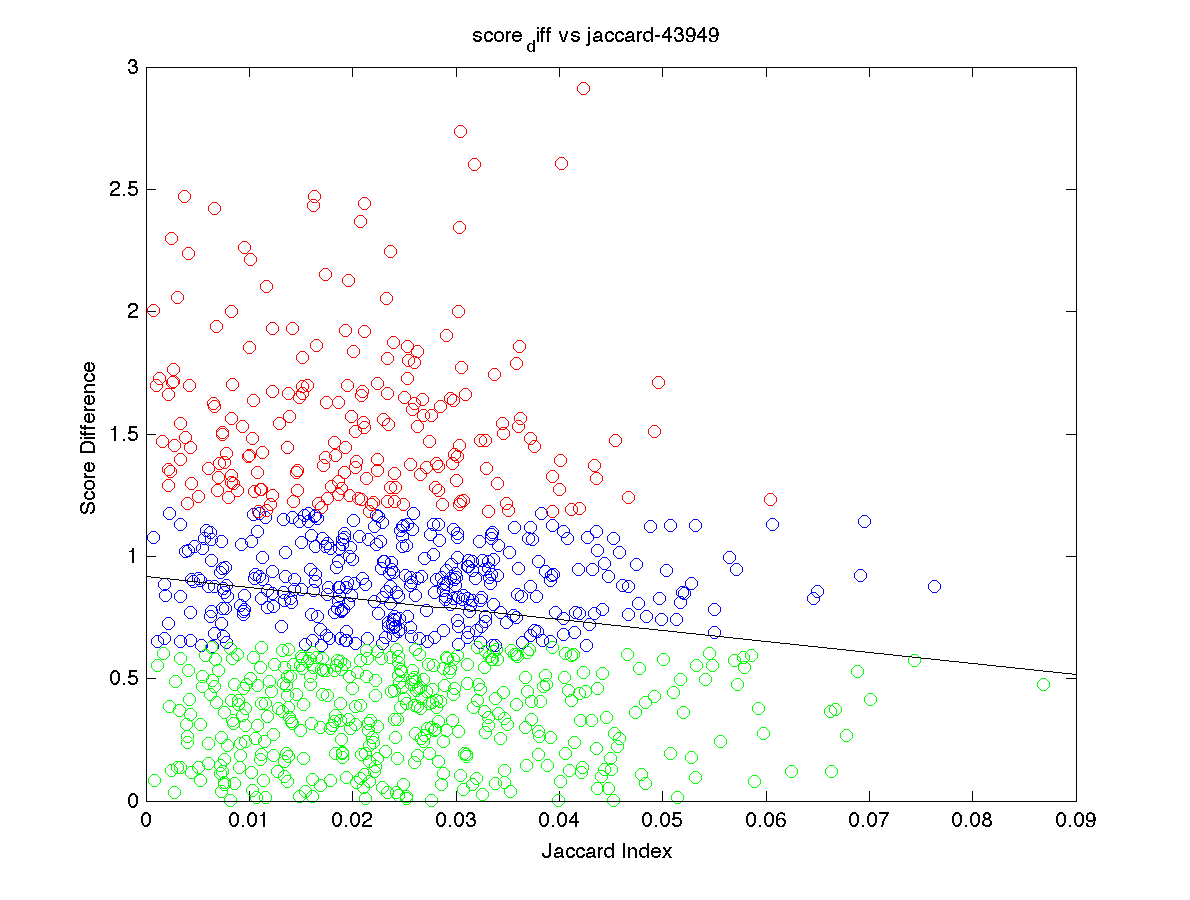




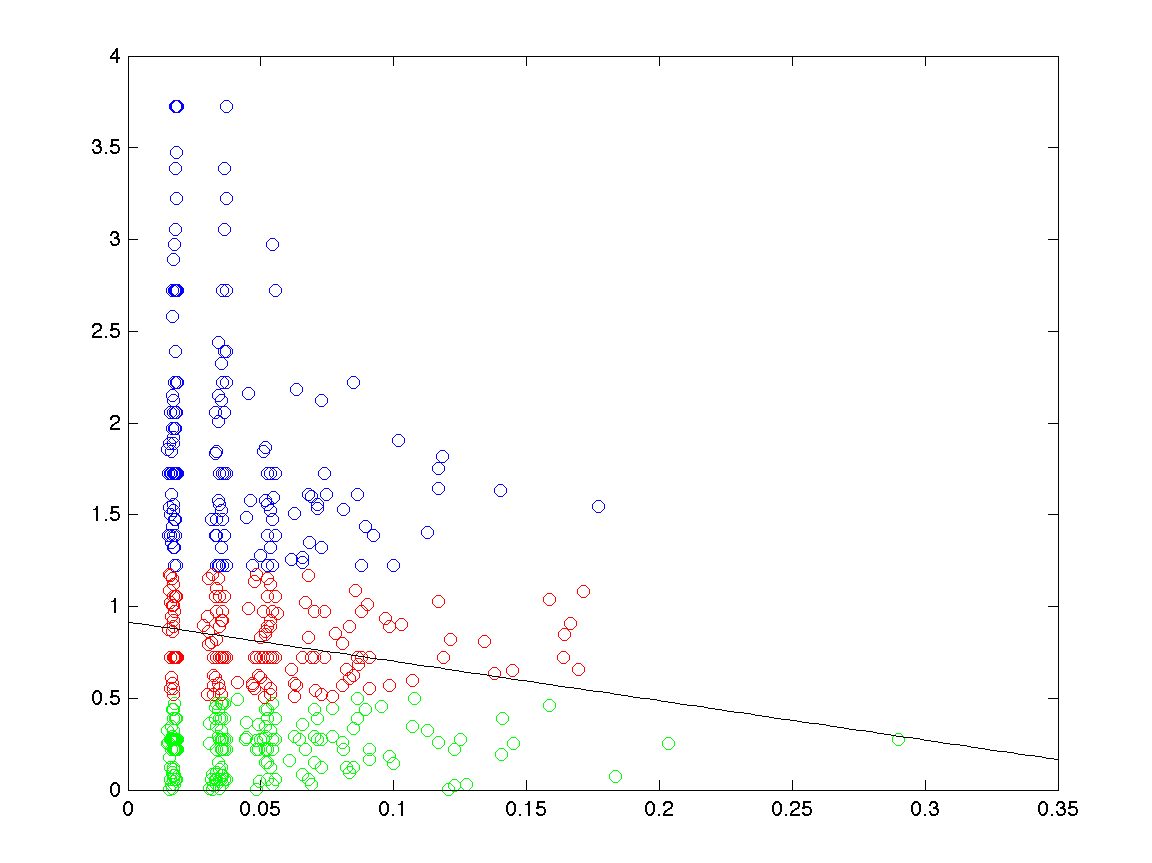
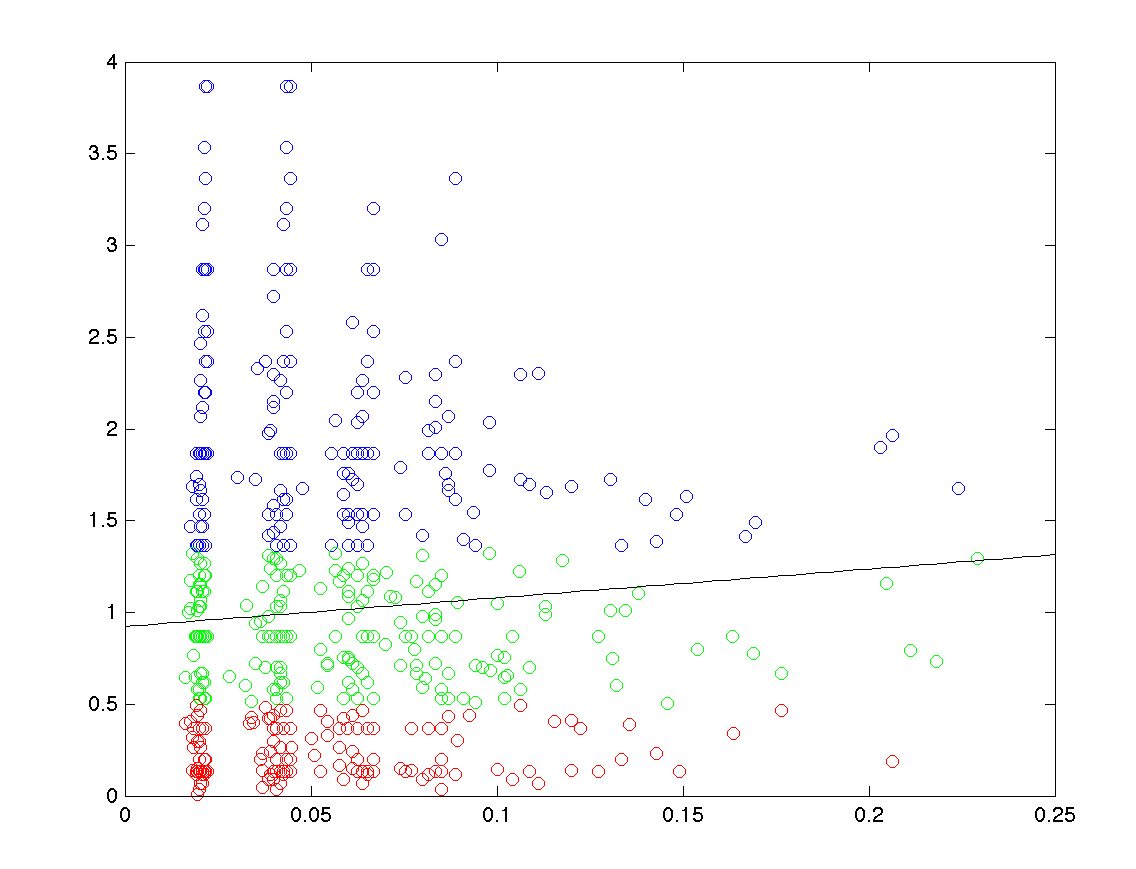








The max score difference is between 2.5-3.0. The slopes are unanimously negative. This supports the hypothesis that for an influential user, the average score difference between that user and other users decreases as similarity between them increases. It also demonstrates that how majority of other users do not deviate far from influential users. When the slope is more negative, the score difference decline faster as jaccard index, or similarity between the pair of users. The smaller the y intercept, the smaller the overall score difference. It means that there are many points with y value, or score difference, close to zero. To verify the results are unique to influential users. I plotted the same graph for a few other users that just made a lot of reviews.



One graph had a positive slope and other has a bunch of points with score difference higher than 3.0 and even 3.5. These evidences agree that the graph pattern exhibited by potential influential users is not pure coincidence.

Together, a steeper negative slope and a lower y intercept reflect stronger influence of the reviewer. Upon such observations, users 644, 3297, 8188, and 19306 seem to be the best results among the candidates chosen. Their scoring patterns are very closely followed by the rest of the data. Their own statistics also imply that many other users read their reviews and found the reviews helpful.

The graphs and observations support the hypothesis. The next step should be to look at the actual content of those reviews and summarize characteristics shared by them. Ultimately, we want to find out the recipe of writing an influential review. One interesting result from the study is that an overwhelming majority of the reviews gave the highest score, 5.0, to the movies, and especially among those who only reviewed once. There could be two possible explanations. One, movie reviewers on Amazon are generally nice, and two, movies on Amazon are typically good movies. This fact, to a certain degree, blurs the distinction between different users and reviews based on pure numbers. A user who rates high scores to movies could have similar score difference graph as influential users. However, to truly influential, the reviewer also has to have the two criteria used to select the potential candidates as well: 1) most of its reviews are read and rated by others, and 2) people find the reviews helpful. It would be helpful to have access to the sales history data. Correlating the time of each review and the sales record in a period of time after the review should produce interesting observations.