Influence Propagation in Location Based Social Networks

Design Lab Report

Yetesh Chaudhary

under supervision of

Prof. Bivas Mitra

Dept of Computer Science and Engineering,

Indian Institute of Technology, Kharagpur,

Nov 18, 2014

Table of Contents

Influence Propagation in Location Based Social Networks

(Yetesh Chaudhary under guidance of Prof. Bivas Mitra)

1) Yelp Dataset:

This dataset contains tips and reviews posted for 10 years from 2005 to 2014.

# of users	70817
# of users with 5 average stars	11933
# of tips	113993
# of reviews	335022
# of locations	15580
# pair of friends	303032

Using gender prediction from name of a user [1], following statistics is also obtained.

#male users	36843
#female users	33974

2) Experiments:

A) Temporal Analysis

A1) Temporal Analysis for same location

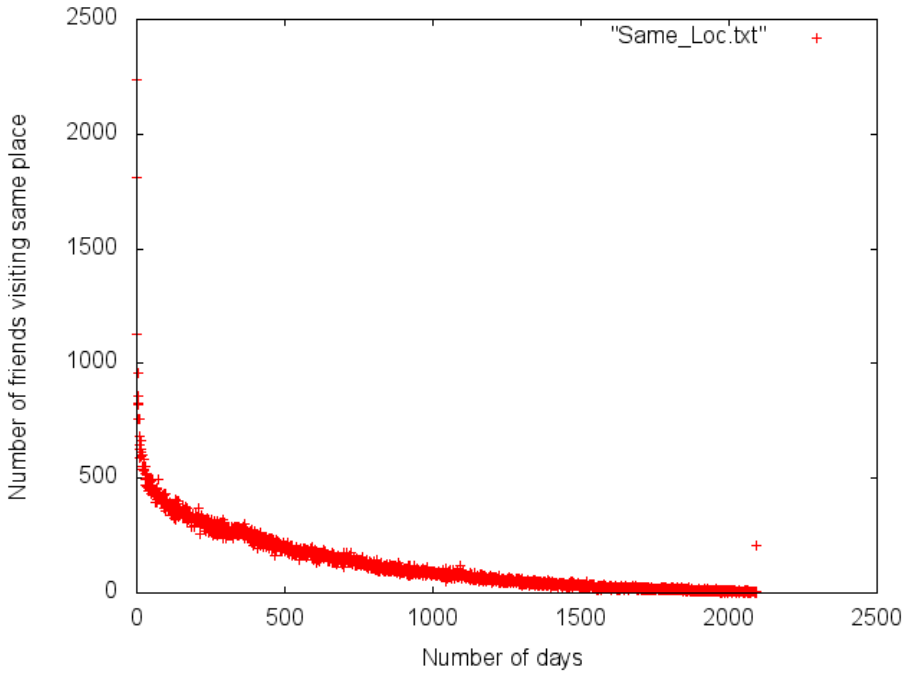
A1.1) Description of Experiment:

I got user information of 70817 users from Yelp dataset in the json format . From there, I formed a dictionary of friends list of each user having unique user id. I also had information of 335022 reviews and 113993 tips posted for 15580 different locations. Each tip and review has a user id and a location id (called business id in Yelp). So, a dictionary of list of (user,date) tuples for each location was formed where date is the date when user posted tip or review for that location. If user posted more than once for the same location, then the most recent post was considered. Then, all pairs of users

who are friends and posted tip or review for the same location are taken into consideration for temporal analysis.

A1.2) Results:

The timings of tips and reviews of 303032 pairs of friends in 15580 different locations checked in through Yelp are analyzed. The plot of **count of pairs of friends visiting same location** versus **the time difference between their tips/reviews postings** after their visits is shown here.



A1.3) Observations:

It is observed that most of the friends post tips and reviews **within the same day** for the same location. So, a high degree of influence of friendship is observed here.

A2) Temporal Analysis for cluster of nearby location

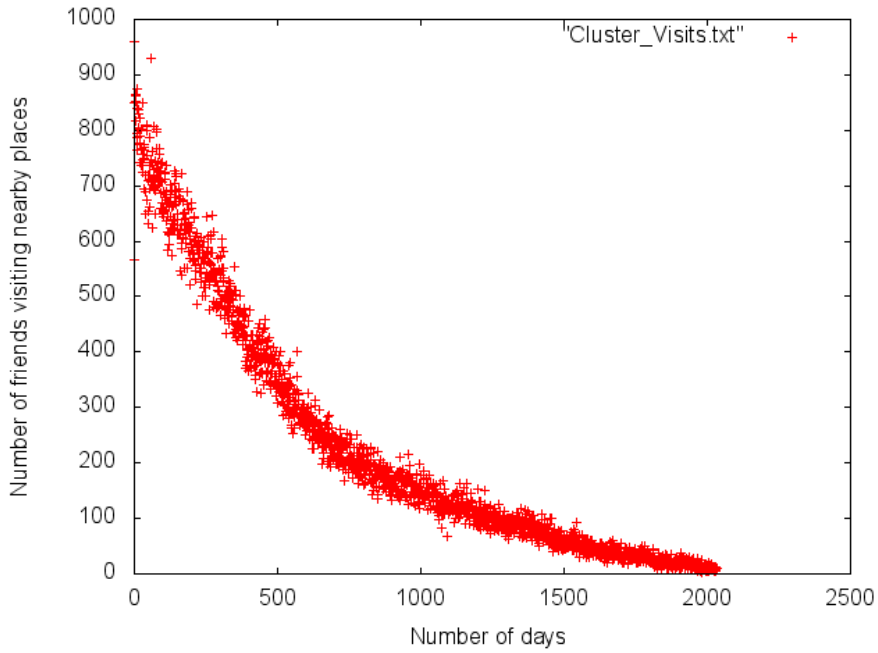
A2.1) Description of Experiment:

The locations are then clustered based on their latitude and longitude in 30 clusters. The clustering algorithm used is k means unsupervised learning through RapidMiner[2].

Then, all pairs of users who are friends and posted tip or review for the **same location cluster** are taken into consideration for temporal analysis.

A2.2) Results:

The timings of tips/reviews of pair of friends visiting the same clusters are analyzed as shown in the given plot.



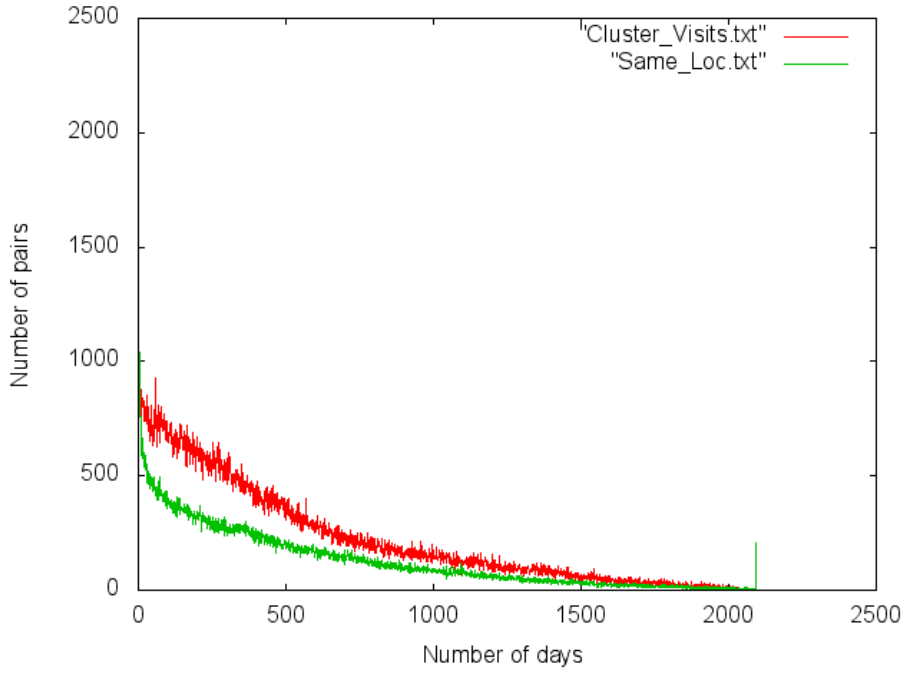
A2.3) Observations:

High degree of influence of friendship is also observed in a cluster of nearby locations. Friends post tips/reviews for nearby locations that are very likely to be visited within the same day.

A3) The comparison of clusters of nearby locations and same location

A3.1) Observations:

Friends tends to visit and post tips/reviews for nearby locations more than for the same location. This also indicates the close geographical distribution of friends.



A4) Gender distribution of pairs of friends

A4.1) Description of Experiment:

The gender was predicted from name of the user using naïve bayes classifier as described in [1]. The count of various gender combinations for pairs of friends who visit the same location and post tips/reviews within the same day was then found out.

A4.2) Results:

Type of pair	Count	Percentage
Male-Male	546	24.39%
Female-Female	557	24.88%
Male-Female	1135	50.71%

A4.3) Observations:

Out of pairs of friends who visit the same location and post tips/reviews within the same day, about 51% pairs were found to be consisting of opposite gender. This clearly indicates the high probability of influence of opposite gender.

A5) Similarity between friends visiting same location

A5.1) Description of Experiment:

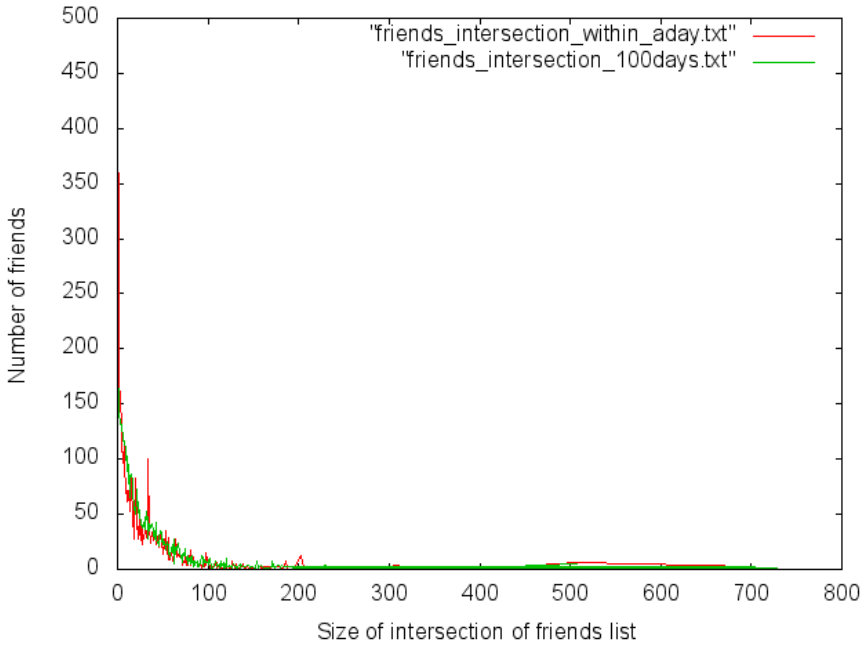
To measure the similarity of friends posting tips/reviews for a location within a day, the average **Jaccard coefficient** for friend pairs posting tips/reviews for same location was found out. The average was defined for the given time difference between the postings of tips/reviews of pair of friends for the same location.

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}.$$

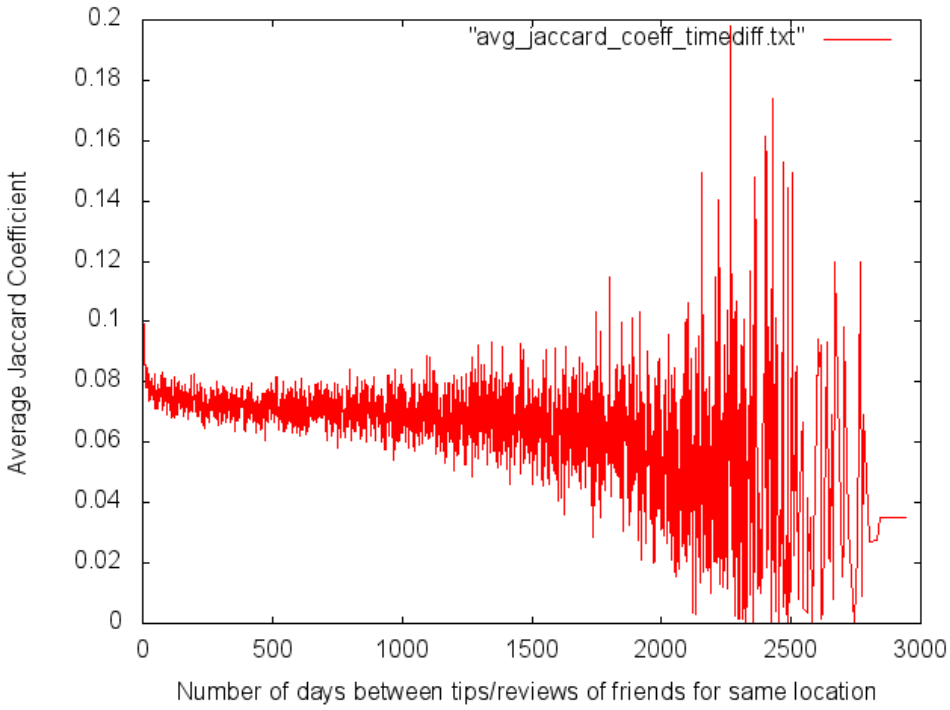
Here, the set A and B are the list of friends of two friends in the pair who post tips/ reviews for same location. The Jaccard coefficient here indicates the extent of mutual friendship of two friends.

A5.2) Results:

1) The given plot shows the count of pairs of friends who post tips/reviews for same location within a day and within 100 days vs the number of their mutual friends.



2) The plot of average jaccard coefficient with the the time difference between tips/ reviews postings of friend pairs for the same location is shown here:



A5.3) Observations:

1) Firstly, the number of mutual friends follows a approximate power law distribution. That is, there are very less friends having a large number of mutual friends.

2) The average Jaccard coefficient is found to decrease with increase in number of days of posting of tips/reviews between two friends for the same location. This indicates a large degree of ties between friends who post tips/reviews for the same location within few days.

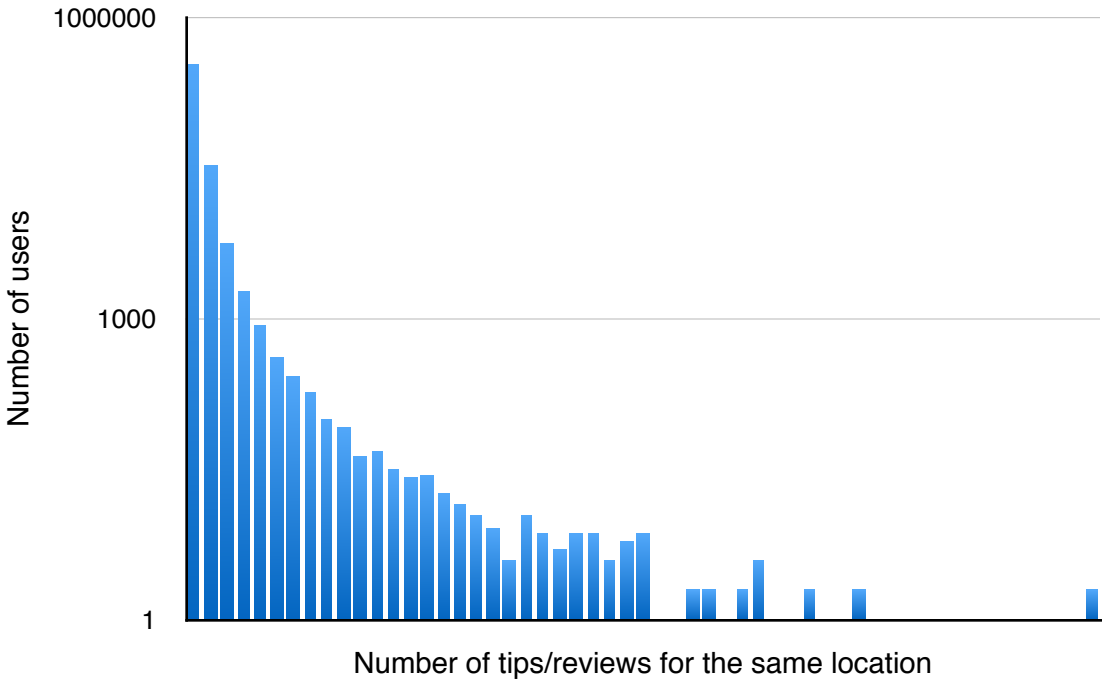
A6) Frequency distribution of user statistics for more than one posting

A6.1) Description of Experiment:

For each location, the number of tips/reviews posted by the same user along with the number of such users was stored in a dictionary.

A6.2) Results:

The plot of number of users posting tips/reviews Vs the number of postings for the same location is shown here.



A6.3) Observations:

89% of the users post tips/reviews once for the same location. About 10% of users post tips/reviews more than once but less than 28 times for the same location.

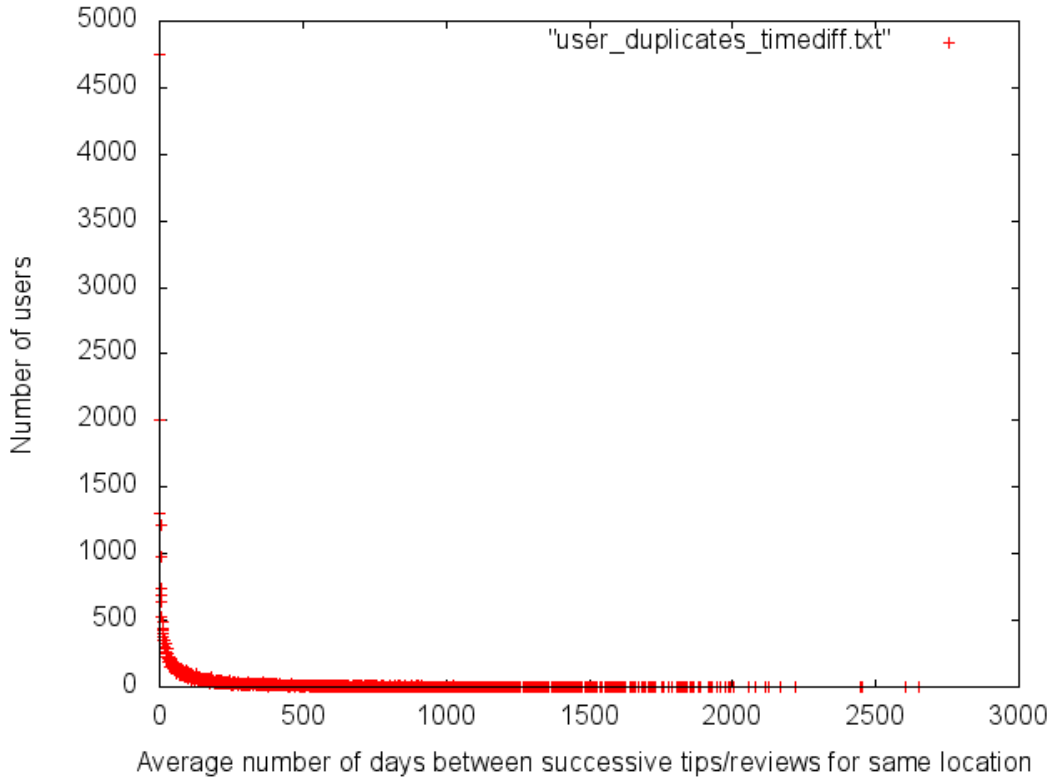
A7) Time distribution of user statistics for more than one posting

A7.1) Description of Experiment:

For each location, the average time difference between successive tips/reviews posted by the same user along with the number of such users was stored in a dictionary. The average is considered over all successive tips/reviews posted by the same user for the same location.

A7.2) Results:

The plot of number of users Vs average time difference between posting of tips/reviews for the same location is shown here.



A7.3) Observations:

The average time distribution between successive posts of tips/reviews of a user for the same location follows power law. If users post tips/reviews of same location more than once, it is found that most users post that within 4-5 days.

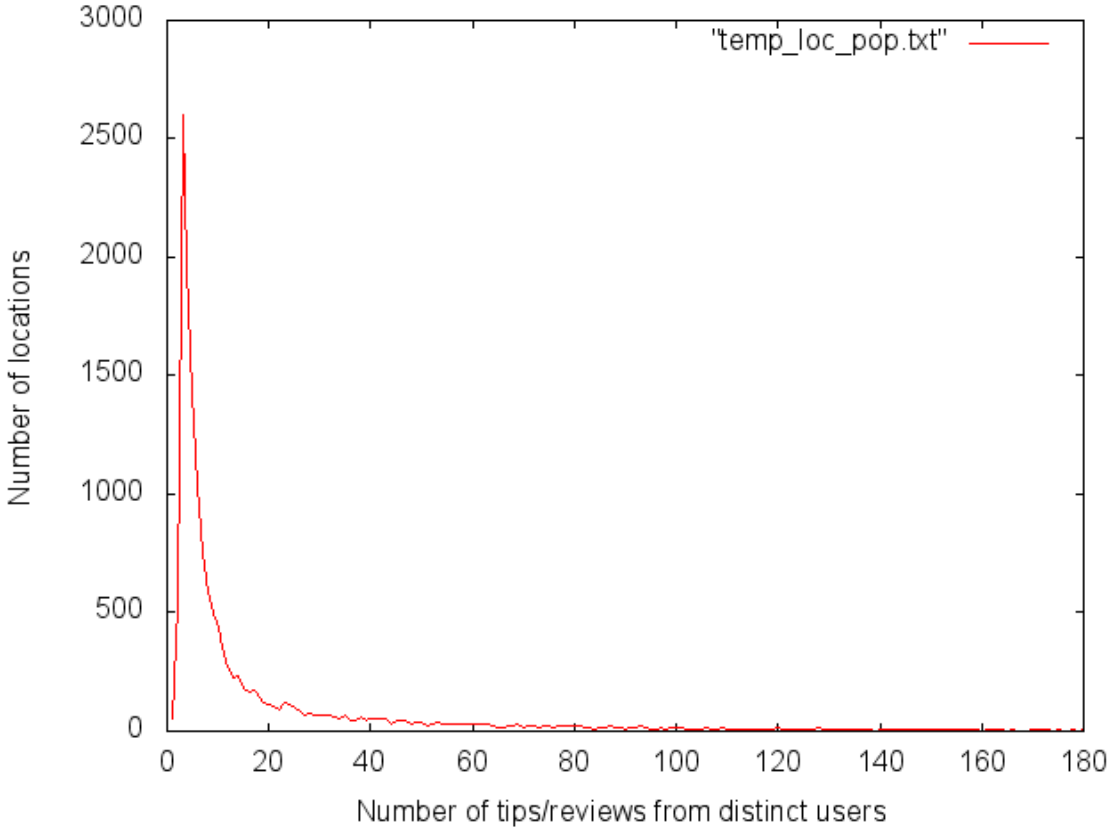
A8) Location popularity

A8.1) Description of Experiment:

We define popularity of a location by the total number of tips/reviews it has from distinct users. For each location, the number of tips/reviews from distinct users was stored in a dictionary. Then number of locations having a particular number of tips/reviews from distinct users was computed.

A8.2) Results:

The plot of number of locations vs number of tips/reviews from distinct users is shown here.



A8.3) Observations:

Most of the locations have 3 tips/reviews from different users. The most popular location has 1533 tips/reviews from distinct users.

A9) Tips Vs Reviews

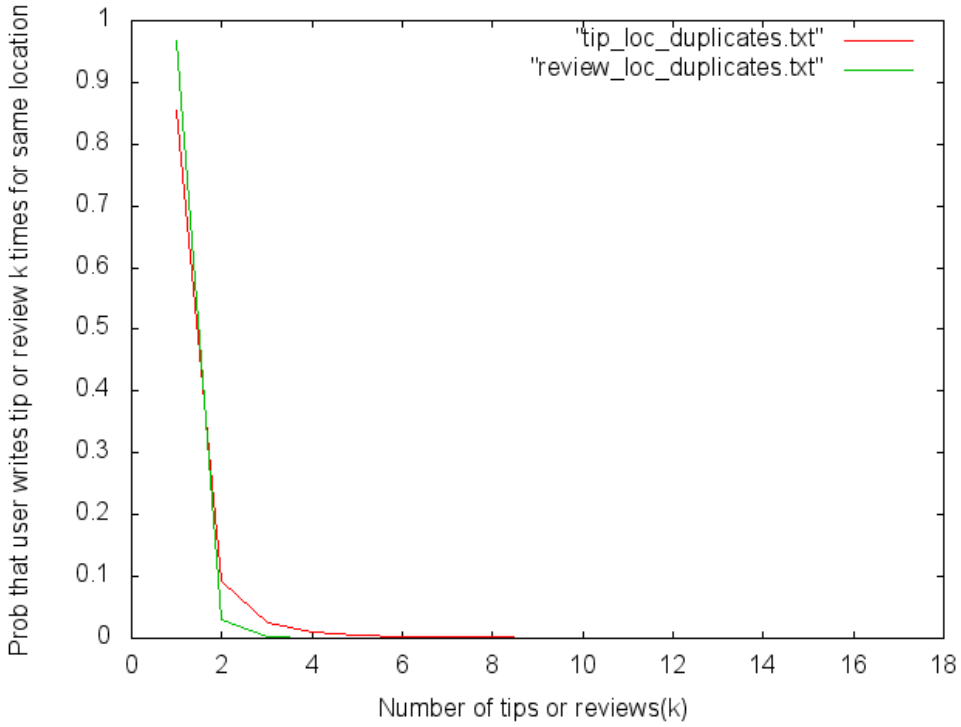
A9.1) Frequency distribution of user statistics for more than one tip vs review

A9.1.1) Description of Experiment:

A6) was performed separately for tips and reviews. In order to arrive at a meaningful comparison, counts was normalized and probabilities was considered.

A9.1.2) Results:

The plot of probability that a user posts tip vs review more than once vs the number of respective posts is shown here.



A9.1.3) Observations:

Overall users post reviews more than tips.
Probability that a user posts tip more than once is higher than probability that a user posts review more than once for the same location.

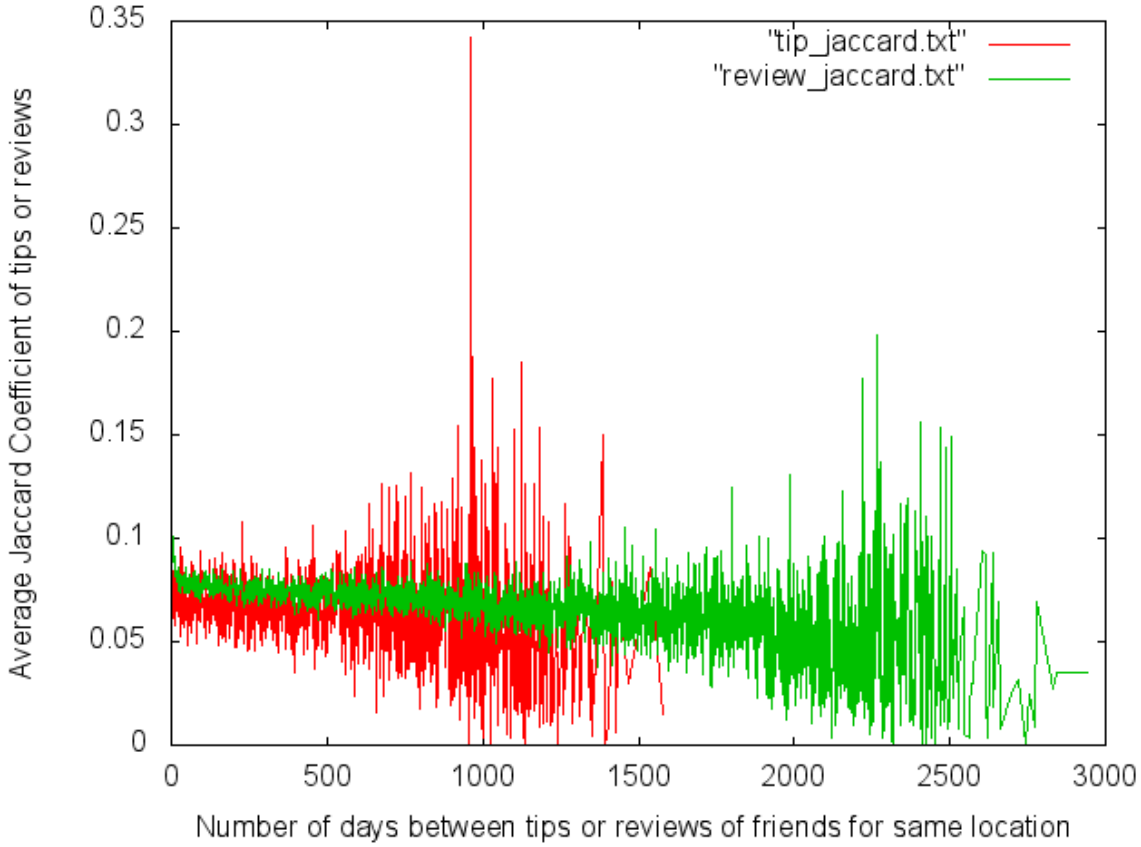
A9.2) Similarity of friends for tips Vs reviews

A9.2.1) Description of Experiment:

As explained in A5), the average Jaccard coefficient between two friends for a particular time difference between their tips vs reviews postings for the same location is obtained. The average is taken over all friend pairs for that time difference and location.

A9.2.2) Result:

The plot of average Jaccard coefficient of friend pairs for tips vs review for time differences between postings is shown here.



A9.2.3) Observation:

The decrease in Jaccard coefficient in case of friends posting reviews is found to be higher than friends posting tips.

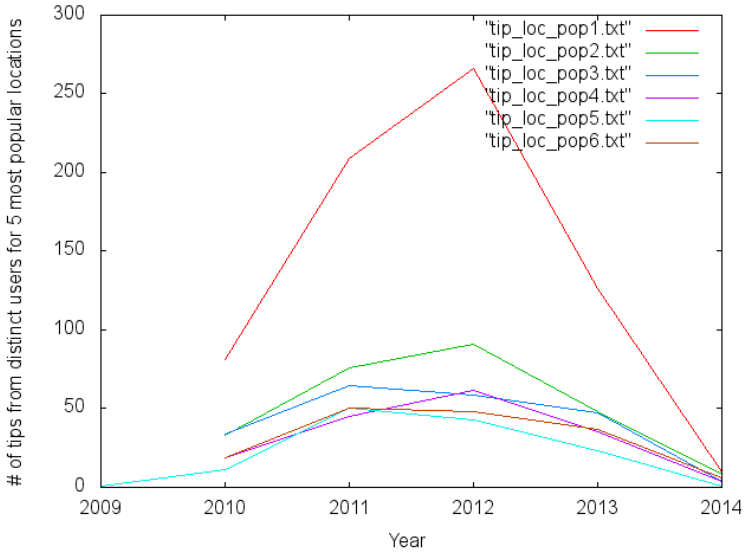
A9.3) Popularity trend of tips and reviews

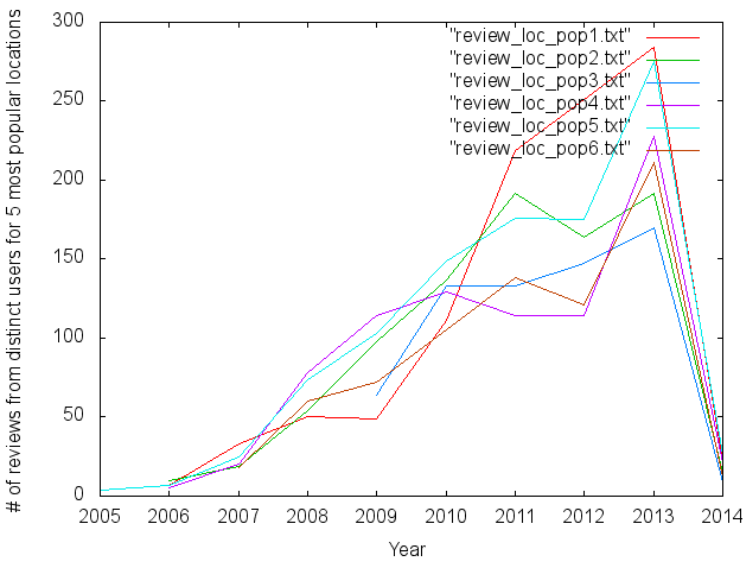
A9.3.1) Description of Experiment:

The top 30 most popular locations in terms of tips and in terms of reviews were separately taken. Then the locations common to both tip and review popularity were taken. They come out be total 21 in number. Out of these 21 locations, five locations with highest (tip count)*(review count) were taken.

A9.3.2) Results:

The plot of tip and review popularity yearly is shown here.





A9.3.3) Observations:

There is sharp increase in number of reviews yearly for 5 most popular locations. i.e. review popularity increases significantly. The tip popularity also increases but at a much lower rate.

B) Semantic Analysis

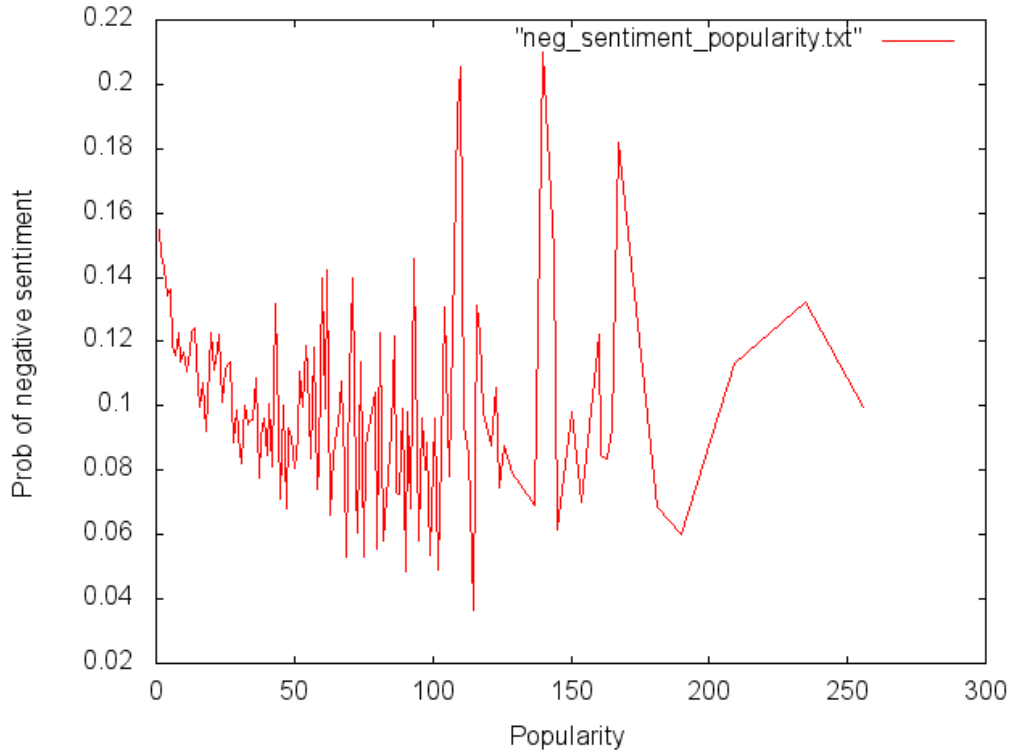
B1) Sentiment correlation with tip popularity

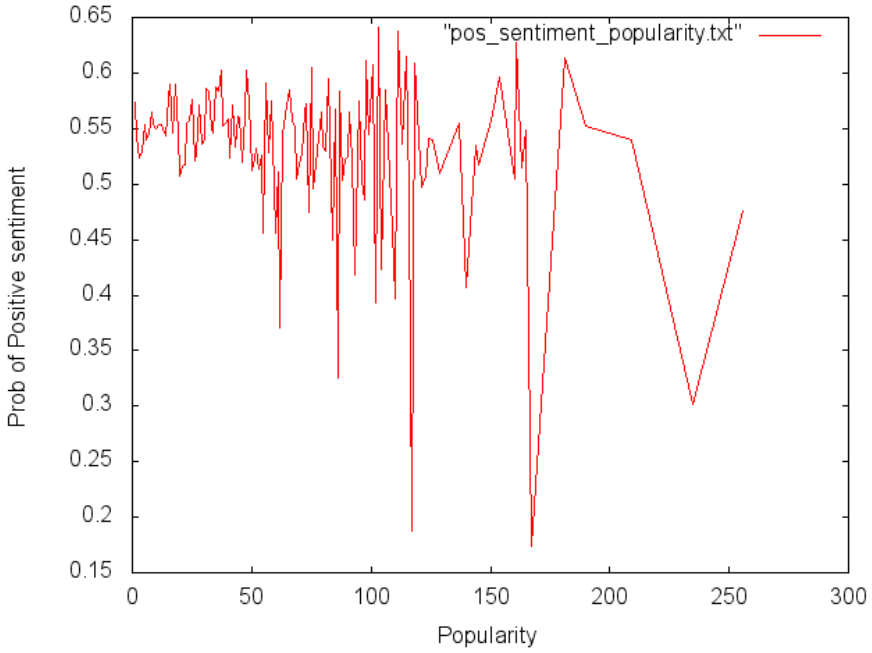
B1.1) Description of Experiment:

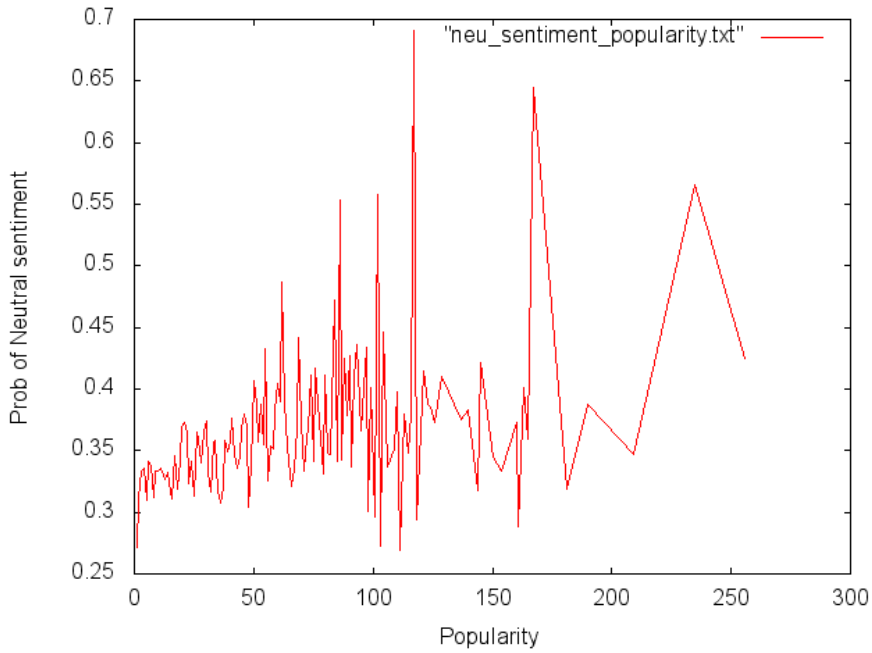
For each location corresponding to a particular tip count, probability of a negative, positive and neutral sentiment tip was calculated. Then average probability over all locations corresponding to that tip count was computed. The sentiment of a tip was calculated using its text as given in [3]. The score is between [-1,1]. If score < 0, then a negative sentiment was considered. If score > 0, then a positive sentiment was considered. Otherwise, if score = 0, then a neutral sentiment was considered.The probability of a negative sentiment tip for a location is number of negative sentiment tips for that location divided by total number of tips for that location. Similarly, for positive and neutral sentiment tips.

B1.2) Results:

The variation of average probability of negative sentiment, positive sentiment and neutral sentiment with location popularity is shown here.







B1.3) Observations:

1) The probability that a location has negative sentiment decreases with increase in its popularity.
2) The probability that a location has positive sentiment remains somewhat stable with increase in its popularity.
3) The probability that a location has neutral sentiment increases with increase in its popularity.

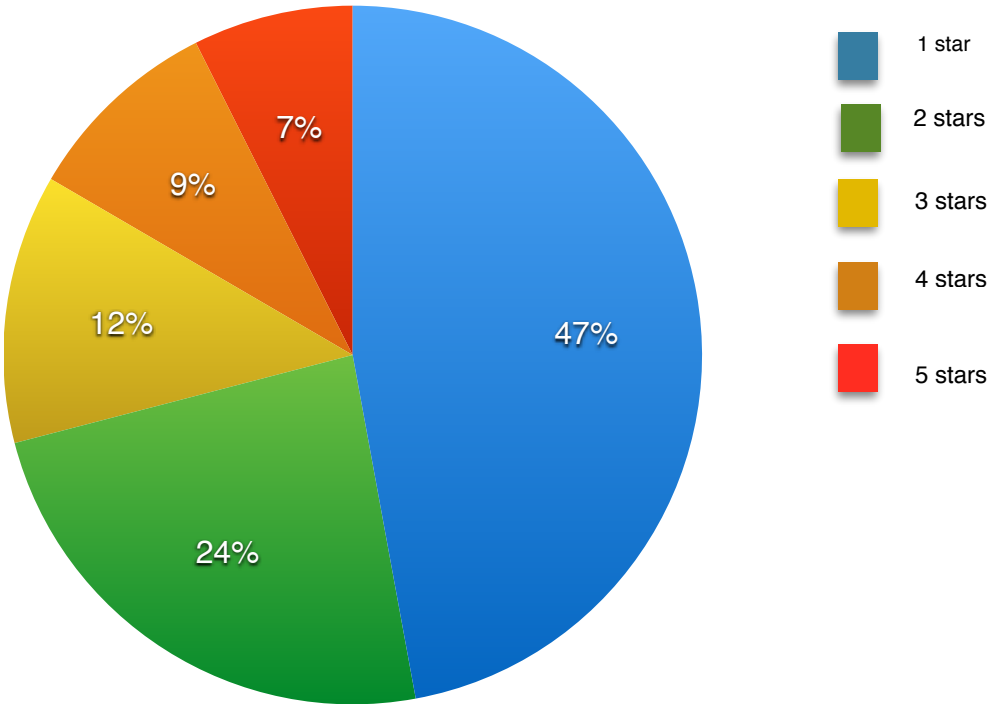
B2) Relation between Sentiment and rating of reviews

B2.1) Description of Experiment:

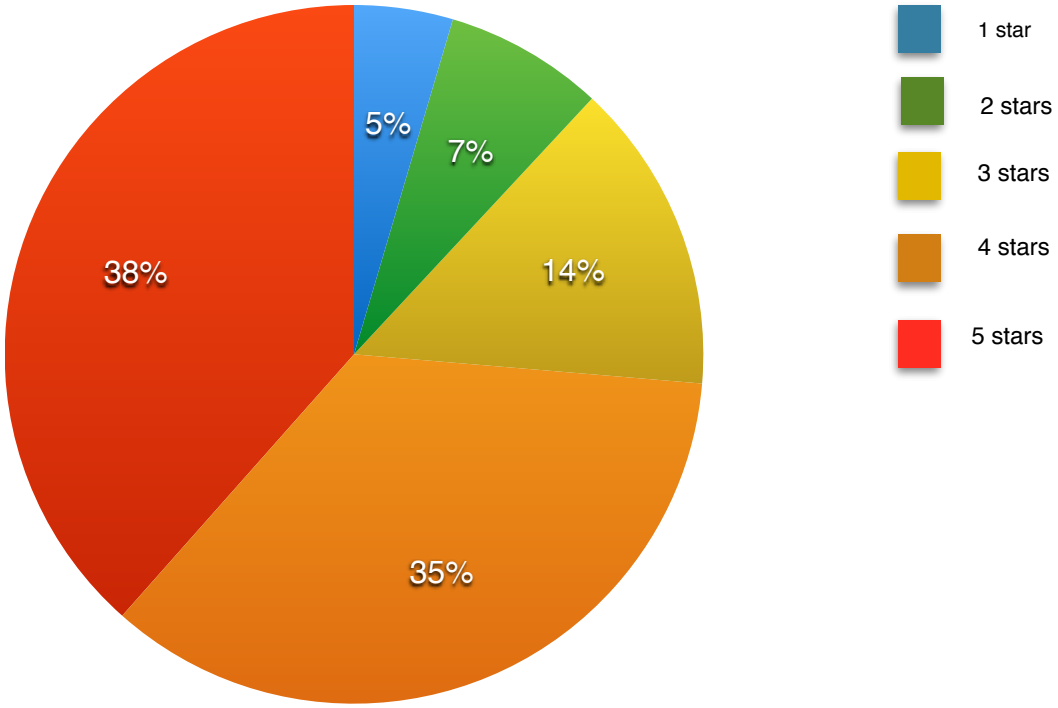
1) The count of reviews with different ratings(from 1 to 5) and negative sentiment was computed using text of review as given in [3].
2) The count of reviews with different ratings(from 1 to 5) and positive sentiment was computed using text of review as given in [3].
3) The count of reviews with different ratings(from 1 to 5) and neutral sentiment was computed using text of review as given in [3].

B2.2) Results:

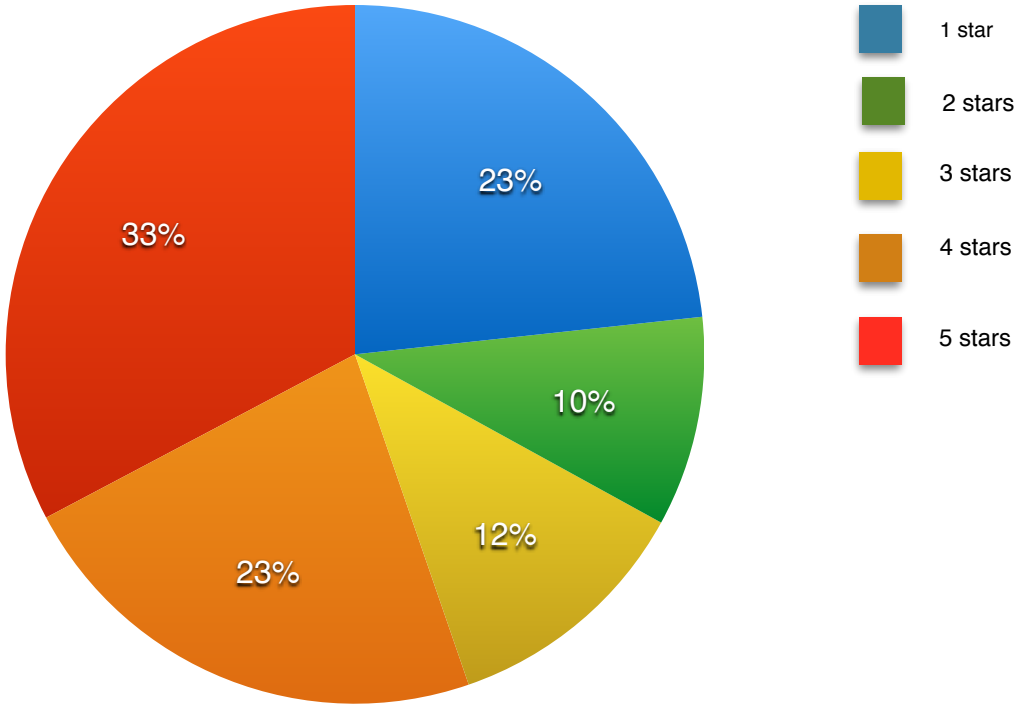
The pie chart distribution for negative, positive and neutral sentiments of reviews along with ratings is shown here.



Negative sentiment



Positive sentiment



Neutral sentiment

B2.3) Observations:

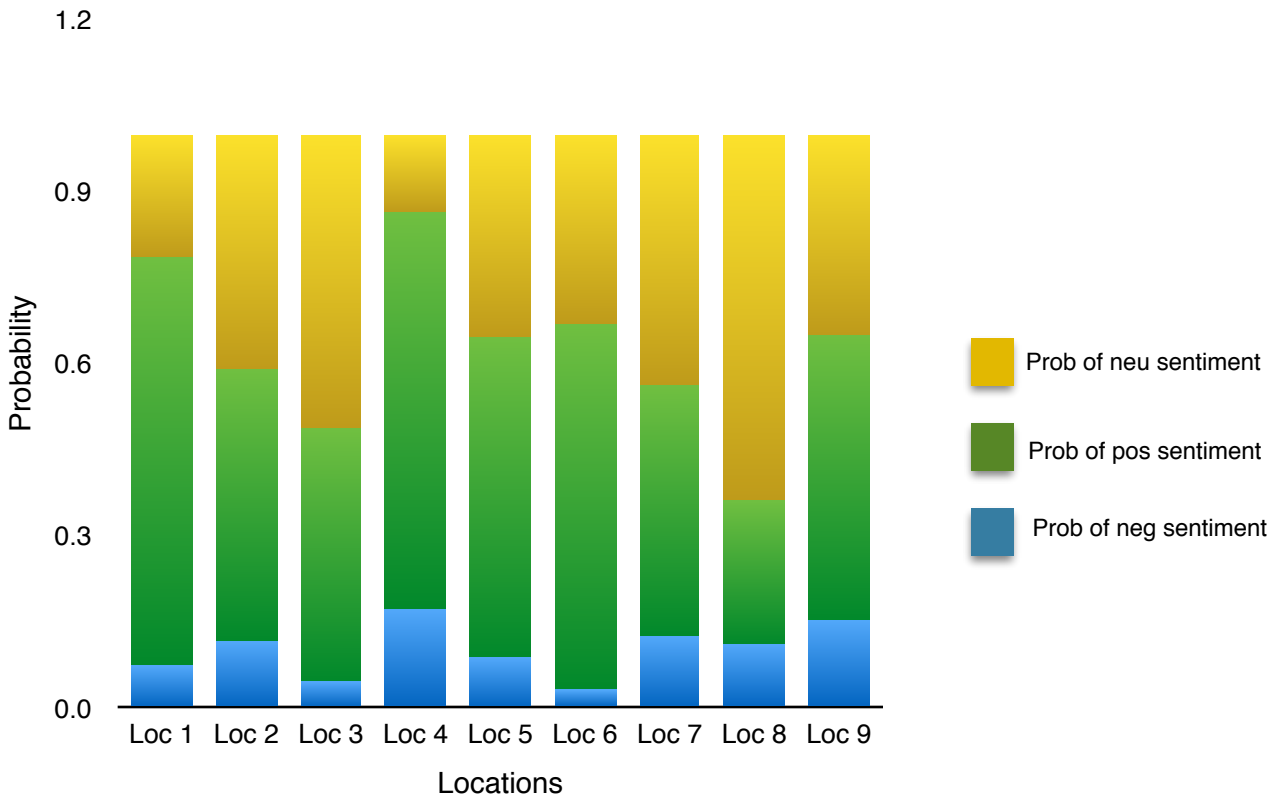
1) 71 % of negative sentiment reviews have 1-2 ratings.
2) 73 % of positive sentiment reviews have 4-5 ratings.
3) Neutral sentiment reviews are mostly rated 5 stars. But significant percentage of neutral sentiment reviews are also rated low.

B3) Tip sentiment of locations with same popularity

B3.1) Description of Experiment:

9 locations were selected which have same popularity (or number of tips/reviews). They are having very high popularity i.e. 127 number of tips/reviews. The probability of negative, positive and neutral sentiment tips from these 9 locations were then compared.

B3.2) Results:



B3.3) Observations:

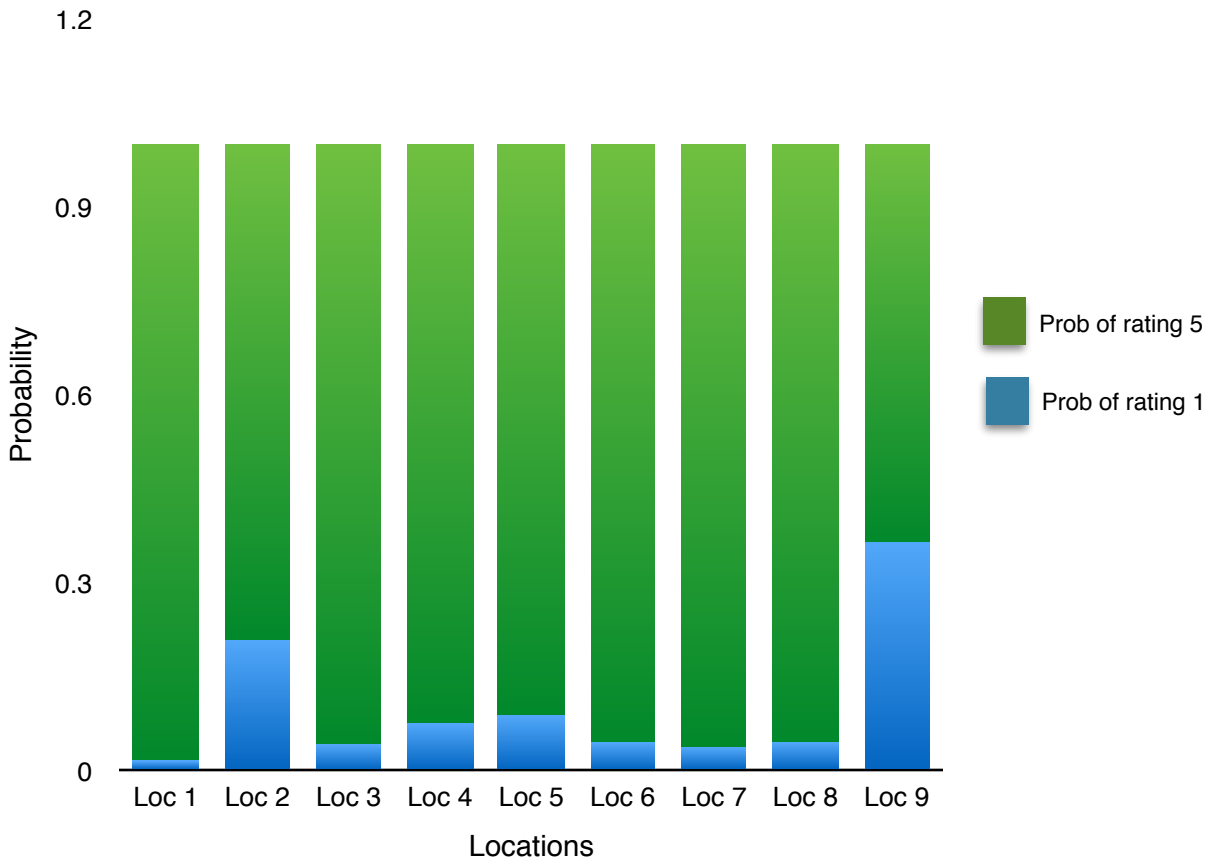
Overall, probability of positive sentiment is higher for these locations because these locations have high popularity. Moreover, the location 8 has 0.25 probability of a positive sentiment tip being posted while location 1 has 0.71 probability of a positive sentiment tip being posted. On an average, all locations have 0.5 probability of a positive sentiment tip being posted.

B4) Review rating of locations with same popularity

B4.1) Description of Experiment:

9 locations were selected which have same popularity (or number of tips/reviews). They are having very high popularity i.e. 127 number of tips/reviews. The probability of rating 1 and rating 5 reviews from these 9 locations were then compared.

B4.2) Results:



B4.3) Observations:

Overall, probability of rating 5 reviews is higher for these locations because these locations have high popularity. Moreover, the location 9 has 0.63 probability of a rating 5 review being posted while location 1 has 0.98 probability of a rating 5 review being posted. On an average, all locations have 0.9 probability of a rating 5 review being posted.

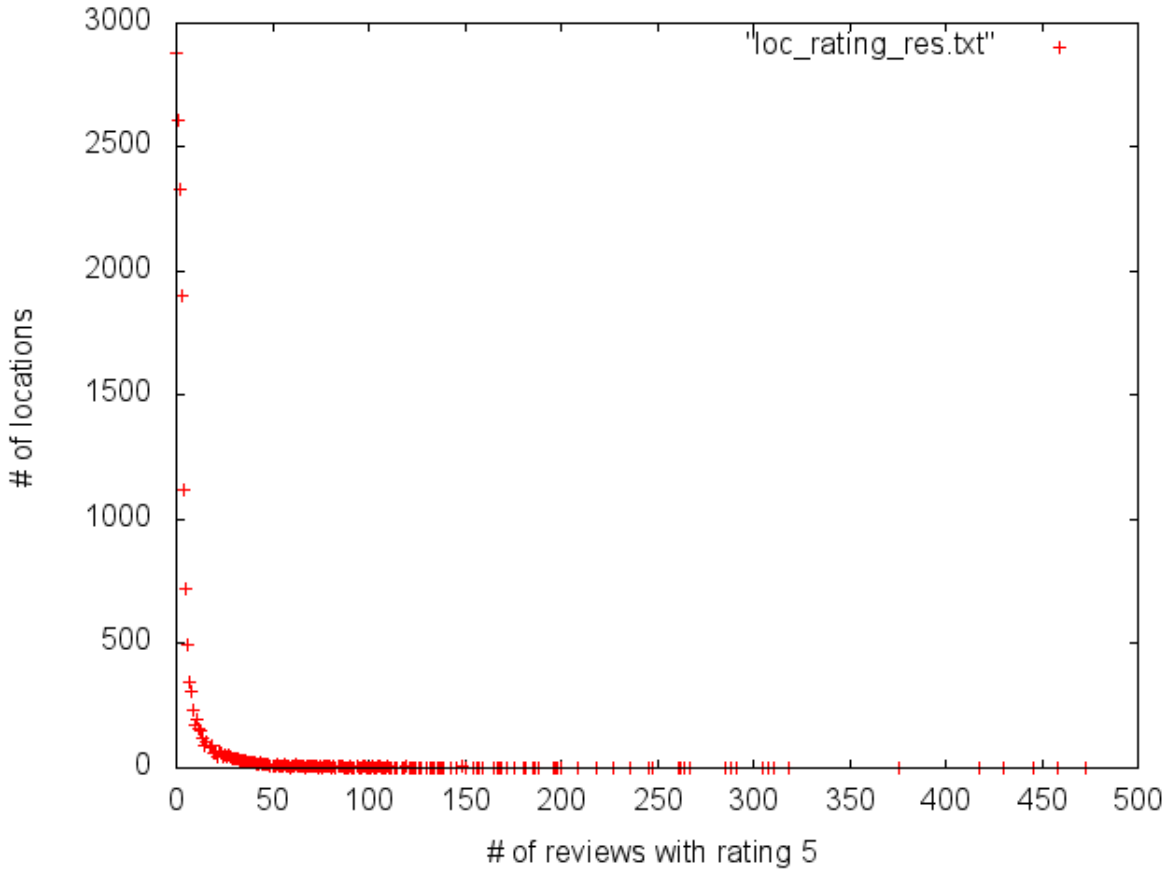
B5) Distribution of highly rated reviews

B5.1) Description of Experiment:

For each location, number of reviews of various ratings was stored. Then number of locations having a particular number of reviews with rating 5 was computed.

B5.2) Results:

The plot of number of locations with number of reviews of rating 5 is shown here.



B5.3) Observations:

The number of locations having number of reviews with 5 rating follows power law distribution. There are very less locations having highly rated reviews.

B6) Multiclass Supervised Classification of reviews based on rating

B6.1) Description of Experiment:

- Learned a model to predict rating of a new review.
- Converted all review texts into tf-idf vectors.
- Training : Test set = 70:30.
- Used SVM, logistic regression and naïve bayes from [4].

B6.2) Results:

SVM with linear kernel:
Test accuracy =58.372% for 5 class classification
Average Precison=55.332%
Average Recall=53.084%

	Assigned 1	Assigned 2	Assigned 3	Assigned 4	Assigned 5
True 1	6057	1159	403	455	517
True 2	1818	2791	2185	1438	571
True 3	559	1102	4664	6525	1349
True 4	224	202	1734	19623	10957
True 5	209	60	260	7918	27220

Confusion Matrix

Logistic regression:
Test accuracy =60.355% for 5 class classification
Average Precison=58.573%
Average Recall=54.261%

	Assigned 1	Assigned 2	Assigned 3	Assigned 4	Assigned 5
True 1	6005	1282	433	426	445
True 2	2023	2883	2094	1288	515
True 3	652	1530	4578	5961	1478
True 4	291	454	2260	18459	11276
True 5	269	154	480	8317	26447

Confusion Matrix

Multinomial Naive Bayes:
Test accuracy = 48.619% for 5 class classification
Average Precison=52.223%
Average Recall=30.570%

	Assigned 1	Assigned 2	Assigned 3	Assigned 4	Assigned 5
True 1	1240	17	19	4468	2847
True 2	124	19	49	7005	1606
True 3	34	3	36	11767	2359
True 4	12	0	7	21002	11719
True 5	17	4	1	9323	26322

Confusion Matrix

B6.3) Observations:

Logistic regression performed best in classification of reviews based on ratings.

3) References:

[1] http://stephenholiday.com/articles/2011/gender-prediction-with-python/
[2] https://rapidminer.com/
[3] https://pypi.python.org/pypi/textblob
[4] http://scikit-learn.org/stable/