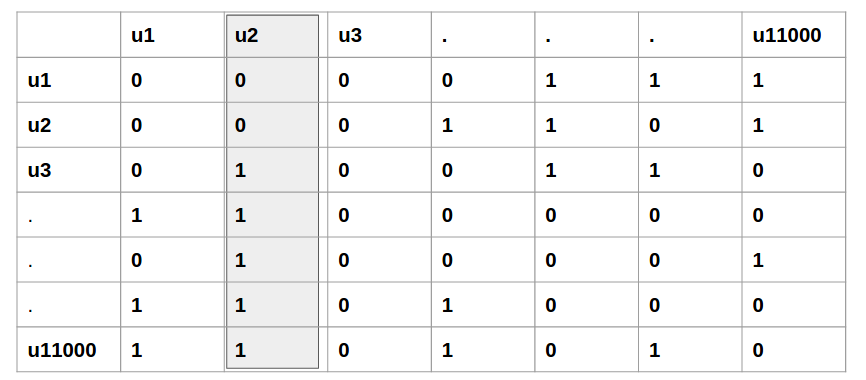
Celebs Followed by Filtered vs Non filtered

Next we wanted to figure out two things like what is the difference (if any) in the kind people who are popular in filtered users sample set ( approximately 11k users ) and the people who are most famous in the Filtered Vs Non Filtered set.( Twitter’s friend concept).

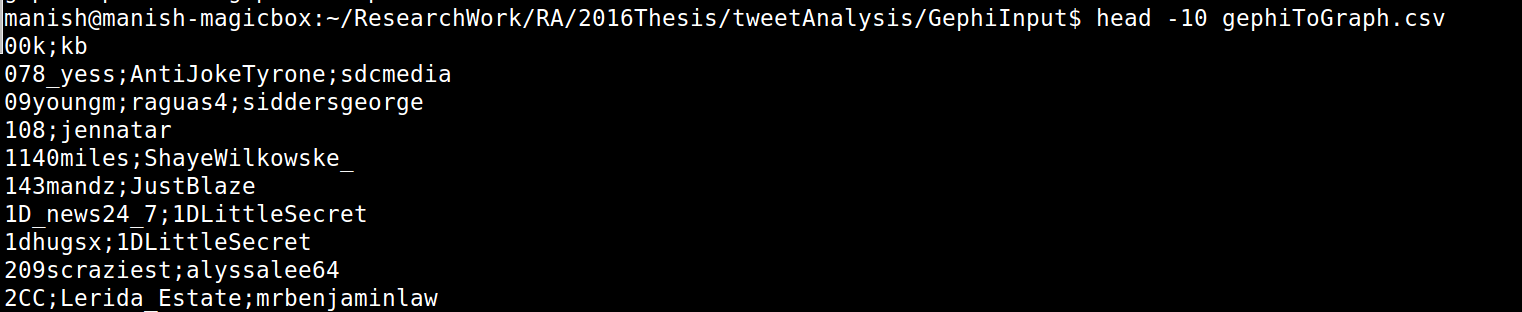
To do this we used aws 5 parallel instances to download set of friends for the filtered users.

Later we created two type of metrics

One of the metrics was square metrics with and .



Which effectively was converted to this format before feeding it to visualization tool **Gephi.**



So, the methodology followed for creating this list was at first using AWS infra friends of all 11k users were downloaded. It took close to 11 days to get all the data. Some issues were twitter restriction on no of query we could push to server and 15 min block time when ever the rate limit exceeded. Our practical observation was after every 5k connection, twitter used to reset the connection for next 15 mins. Hence we came up with parallel system which used AWS machines to download data set using 5 different bots.

Here is how the metrics was filled :

Say for user X



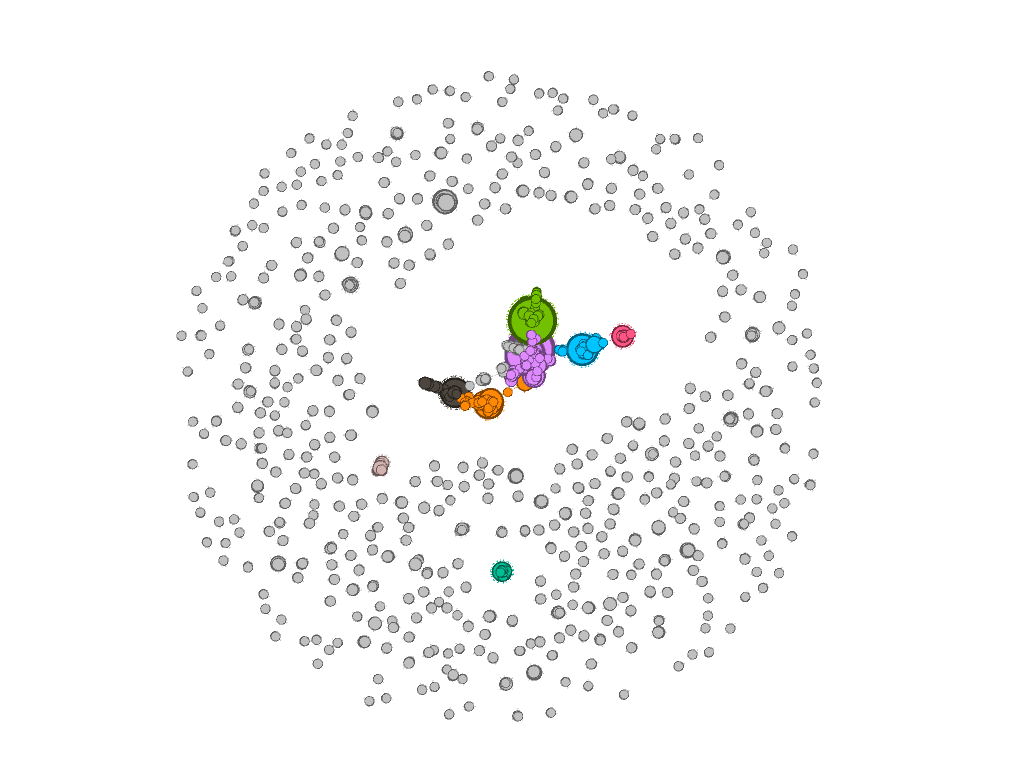
|  |  |  |  |
| --- | --- | --- | --- |
|  | u1 | u2 | u3 |
| u1 | 0 | 0 | 1 |
| u2 | 0 | 0 | 0 |
| u3 | 0 | 0 | 0 |

in Gephi it will represented as



out of 11k user in consideration, only 1000 had at least outdegree in the given 11k user set.

When we put all the 1000 users based on their outdegree in the same group, the graph looked like



The graph at first sight looked like this. This kind of graph can be read as, the users among themselves weren’t strongly connected. However some of the comedy channel, few bots, and some users with very contrast sentiments. Which seems about right.

A more closer look at the high indegree looks like following graph, where size represents the indegree strength. i.e size  count(indegree).



Figure 1Zoomed vie of user To User as friend analysis

So we wanted to study the users who were so popular. We had this notion that may be these users share a lot of negative content/ sexual jokes. Which was making them so popular. So we downloaded last 3200 tweets of these individuals whose indegree was higher and ran python NLTK sentiment analysis on top of them. However, this package could only give us final polarity of a word, but not break in negative and positive distribution. We decided not to miss on that feature as because of volume of data it was possible that the overall sentiment was getting balanced . Hence we wrote our own dictionary based learning where we use Google provided negative

We were also interested in looking at the tweet frequency, if the same users retweet were also popular? Was there any correlation between correctness (Language correctness) and sentiment of sentence as such?

We also computed the avg Negative of Filtered set across tweets across user as well as for non filtered set as well

The values were

Filtered Set User: 0.39

Non Filtered User Set : 0.24

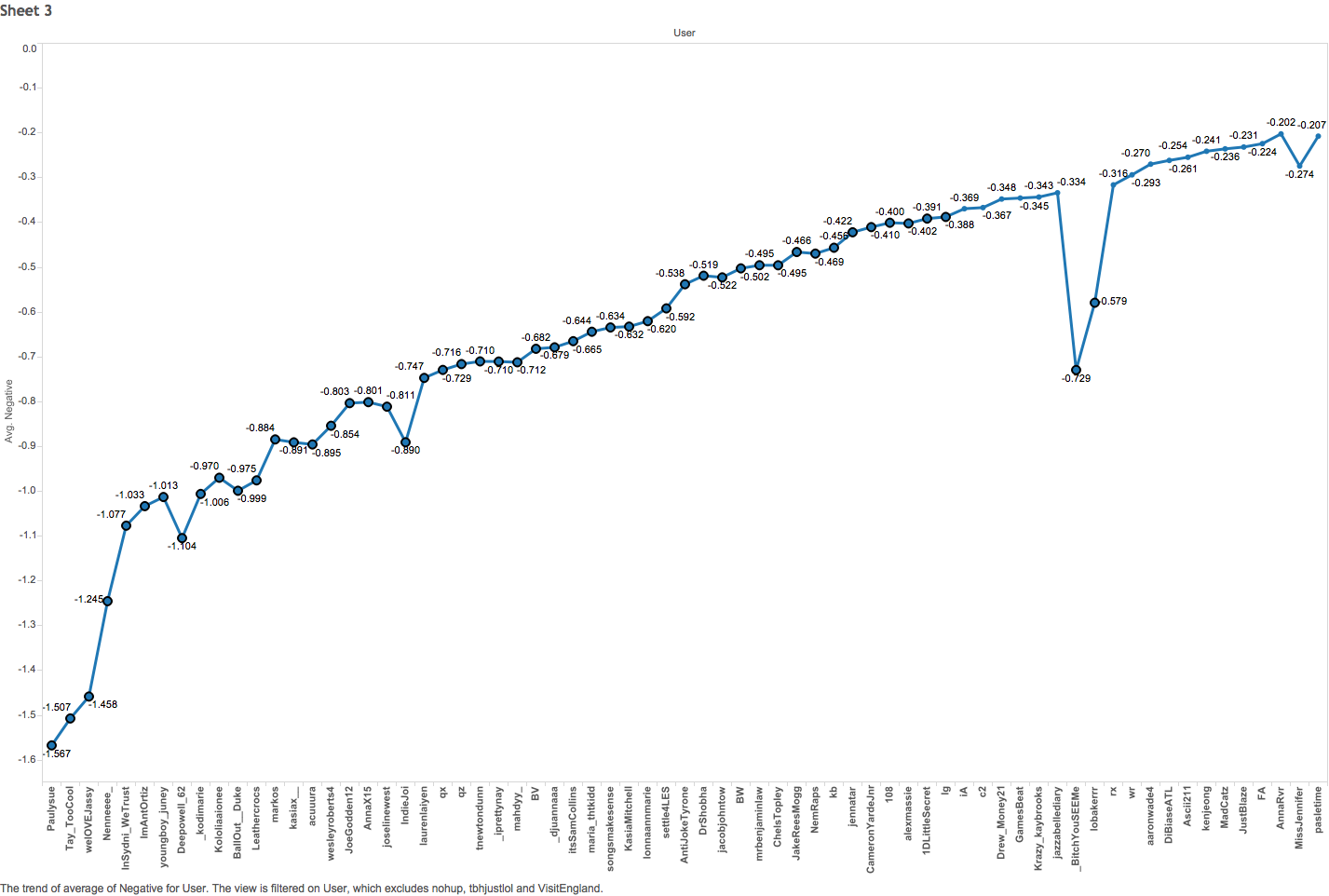


Figure Top 10 % , sorted by negative average

This property holds a good meaning that, because the top 10% users who are most followed tweet on average more negative content than any other user in the set of 1000 users, the users are popular.

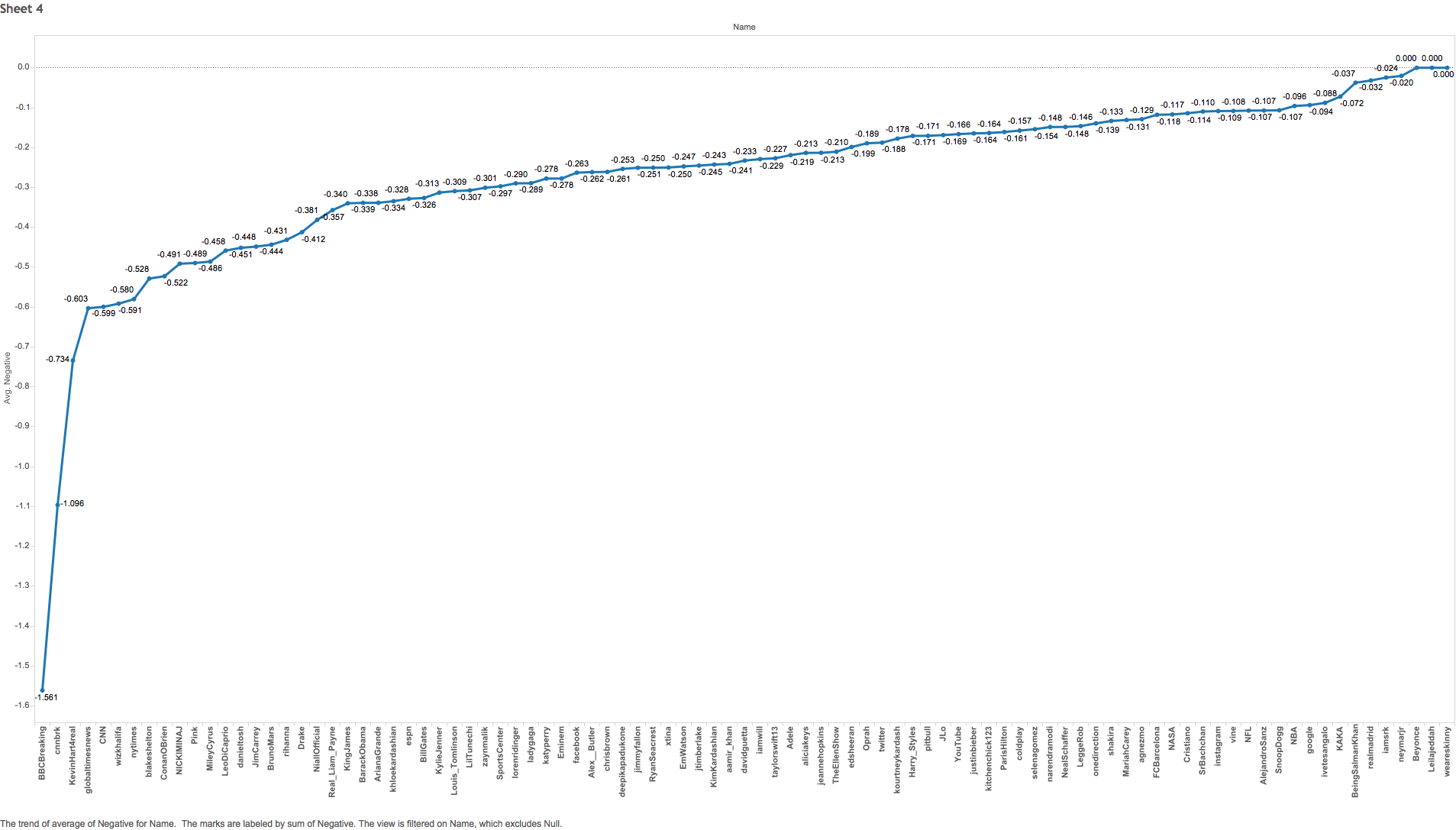


Figure top 100 user NF in comparison with our NF users

The relative comparison shows that, “The users who are popular in filtered category, are possibly famous because many of them tweet more than average negative sentiment across group.”

The other NF graph suggests that, unlike the Filtered popular set, the other random celebrity tweet way less in no as far as negative tweet content is concerned.

One possible Result: Users in filtered category can be given weight based on their indegree value. And top 10 % user with highest indegree value can be “labeled” aggressive in negative tweets.