Part 1:

a) Collected user mentions over a period of 10 minutes from twitter stream on 10/13/2012 at 4:31 p.m.
Total number of nodes: 38919
Total number of edges: 24272

Source Code: UserMentions.java
Outputfile: morementionsnew.gv
Raw Graph file: morementionsnewsorted.gv
Graphs: topconnectedcomps.png, connectedcomps.png
Steps:
  • First sorted the file and removed uniq terms using shell command.
    sort morementionsnewsorted.gv | uniq > morementionsnewsorted.gv
  • Ran Ccomps command to generate the first two visualizations:
    ccomps -zX#0-1000 morementionsnewsorted.gv | grep "-
    cat <(echo "digraph mentions "{) - <(echo ")") | sfdp -Gbgcolor=black -Ncolor=red -Ecolor=white -
    Nwidth=0.02 -Nheight=0.02 -Nfixedsize=true -Nlabel=" -Earrowsize=0.4 -Gsize=75 -
    Gratio=fill -Tpng > connectedcomps.png
  • Ran Ccomps command to generate the first two visualizations:
    ccomps -zX#0-7 morementionsnewsorted.gv | grep "-
    cat <(echo "digraph mentions "{) - <(echo ")") | sfdp -Gbgcolor=black -Ncolor=red -Ecolor=white -Nwidth=0.02 -
    Nheight=0.02 -Nfixedsize=true -Nlabel=" -Earrowsize=0.4 -Gsize=75 -Gratio=fill -
    Tpng > topconnectedcomps.png
  • Used morementionsnewsorted.gv file in Gephi to plot rest of the graphs.
Plotted the graph for 1000 connected components using graphviz.

In the above graph we can clearly the following:

- Large clusters in the middle. These depict closely knit user communities where people have mentioned a specific person a number of times e.g. Youtube, Justin bieber (these have been explained in detail in the later sections)
- There are multiple self loops also, which means quite a few people mentioned themselves in the status (but such graphs are not useful for analysis)
- We can also see on an average one user has mentioned about 3-4 people. This is evident from the star shaped graphs with one central node and edges pointing to multiple nodes in the shape of a star.
Top 8 connected components:

b) Enhanced the graph by re-plotting them using Gephi.
   - Labeling the nodes
   - Showing large components separately.
It was interesting to note that despite the election season Obama, Romney did not find a place in mentions from Saturday.

justinbieber, youtube managed to be the top connect components. Also UnionWorld J and X Factor appeared in the mentions graph since they had appeared on X factor the previous day.

Graphs have been displayed below:

Highest connected component: justinbieber and the one direction band members. This community seems to consists of pop stars, Justin Bieber, Austin Mahone.

Louis Tomlinson, Zayn Malik, Harry styles, Real Lian Payne are all from one direction band and have been mentioned in the top left part of the graph below. This band was
in news in the last 2 days due to a controversial confession of the band members on X Factor show last on Saturday evening.

To enhance visualization the node labels have been sized based on the in-degree (number of mentions)
Second community: YouTube
Third community: Union J, world and District 3 music were favorites to leave X-factor last night and hence have been mentioned so many times in yesterday's tweets.
c) To enhance processing of the graph, the graph was first loaded in Gephi and the algorithm to calculate modularity class was run. This algorithm, divided the network into communities and then top 6 communities were taken and colored differently for better visualization of the graph.

The size of the node label was also set according to the in-degree value.

To enhance processing of the tweets, mentions were picked up directly using UserMentionEntities function. The tweets were not parsed separately from the status text. Also tweets in other languages, which come as question marks were removed. The label sizes have been adjusted based on the in-degree and we can see that justonbieber, UnionJWorld and YouTube has the highest in-degree for the mentions graphs.
PART2:

a) Ran the script for Interest graph and found the following top suggestions based on that. The recommendations provided were very relevant, a lot of the people I follow are from the Ischool hence the top recommendation to follow is UCBTweeter. I would also definitely be interested in following MartiHearst, twitterapi and nytimes which clearly a lot of my friends are already following. I would also be interested in following google, which 6 of my friends are following. This recommendation purely goes by the number of friends following a particular person and does not take other relevant factors like hashtags into account.

12 of your friends are already following UCBTweeter
9 of your friends are already following timoreilly
9 of your friends are already following Oprah
8 of your friends are already following jack
8 of your friends are already following MartiHearst
8 of your friends are already following MarsCuriosity
6 of your friends are already following twitterapi
6 of your friends are already following rusransv
6 of your friends are already following nytimes
6 of your friends are already following neiltyson
6 of your friends are already following marissamayer
6 of your friends are already following google
6 of your friends are already following ericschmidt
6 of your friends are already following SteveMartinToGo
6 of your friends are already following Joi
6 of your friends are already following FakeDorsey
6 of your friends are already following BerkeleyISchool
5 of your friends are already following tedtalks
5 of your friends are already following sacca
5 of your friends are already following mkapor
5 of your friends are already following karaswisher
5 of your friends are already following johnbattelle
5 of your friends are already following cshirky
5 of your friends are already following aweigend
5 of your friends are already following andersoncooper
5 of your friends are already following ZooeyDeschanel
5 of your friends are already following YouTube
5 of your friends are already following UCBerkeley
5 of your friends are already following Gladwell
5 of your friends are already following BorowitzReport
5 of your friends are already following ActuallyNPH

b) Graph Metrics:

i) **Clustering coefficient** for a node is the ratio of the link between neighboring vertices and the max possible number of link.

The graph was converted into an undirected graph and Cluster Coefficient was calculated using Gephi inbuilt algorithm and the result are as shown below:
Average Clustering Coefficient: 0.026

Total triangles: 852
The clustering coefficient value is not too high clearly indicating and triadic closure is not full filled to a greater extent in this network.

iii) Kcore analysis:

On producing a 3 core network I found that there were 5 close knit communities. This is based on calculating modularity class[ link ]The communities are clearly visible. The red community belongs to I school people and is closely linked to each other. The one in pink looks like that of Twitter engineer’s community and the green colored belongs to the entertainment industry. We can see clear demarcation in communities here.

ii) Eigenvector centrality is a measure of the importance of a node in a network. It assigns
relative scores to all nodes in the network based on the principle that connections to high-scoring nodes contribute more to the score of the node in question than equal connections to low-scoring nodes. Google's PageRank is a variant of the Eigenvector centrality measure. Based on the above analysis and on my list of friends, UCBTweeter followed by MartiHearst the most important node in the network. This is a logical conclusion since UCBTweeter and MartHearst are in my list of top recommended people to follow.

c) Visualization of my interest graph. I took the interest graph and ran a modularity analysis on it. Modularity is a measure of how well the network decomposes modular communities. I found out that there were 32 modular communities. Below is the table showing the analysis.
Below is the plot of top 4 communities for a 3 core network.

We can see that a lot of my recommendations e.g. UCBTweeter, tim o’reilly, Marti Hearst came from the Berkeley community(marked light green) and the Twitter Engineers community (marked red)
Source Code: question2.py and TweetRecs.py
Outputfile: twitter_network_edgelist_new.csv
Steps:
• Ran question2.py to get a list of recommendations and obtained twitter_network_edgelist_new.csv from the same code
• Used twitter_network_edgelist_new.csv in Gephi and removed all the edges that pointed to {}. (Note: the csv file generated had an edge from user to {} with the weight = total friends or 100 whichever is lower. This was removed because assigning weight to the edges was irrelevant here).
• Used the clean file to plot other graphs using Gephi
PART 3:

a) For this part I try to analyze the central users of the friends graph and compare their mentions graphs to see whether the centrally connected users also have a good conversation pattern amongst each other or their conversations are more wide spread are not tightly confined with the friends.

b) I start with the friends graph obtained in Part2 and find out the central users of my friend’s graph. I interpret centrality in terms of K core. Higher the K core, more tightly are the users connected to each other in the network.

I took 6 core network and found the following central users:

![Diagram of central users with connections]

A) CoreyHyllested
B) myy_precious
C) UCBTweeter
D) shreyas
E) gilad
F) aneesl
G) Alpa
H) deliprao
I) rruslan
J) billgramham
K) larsonite
L) lintool
M) kostas
N) FakeDorsey

I ran a java program friendsMentions.java to get the first 200 mentions for each of the users in the graph above (that is the maximum number of mentions allowed per page, 200 was also the optimum number of mentions because pattern has to be analyzed for a specific group over the last 3 weeks or so).

d) From the mentions graph I take the in-degree parameter to find out which users have been mentioned the most number of times and arrange the nodes in decreasing order (here rank refers to the in-degree):
<table>
<thead>
<tr>
<th>Rank</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>jco</td>
</tr>
<tr>
<td>6</td>
<td>gilad</td>
</tr>
<tr>
<td>6</td>
<td>kevinweil</td>
</tr>
<tr>
<td>6</td>
<td>UCBTweeter</td>
</tr>
<tr>
<td>6</td>
<td>twitter</td>
</tr>
<tr>
<td>5</td>
<td>pankaj</td>
</tr>
<tr>
<td>5</td>
<td>squarecog</td>
</tr>
<tr>
<td>5</td>
<td>zhenghuali</td>
</tr>
<tr>
<td>4</td>
<td>TwitterOSS</td>
</tr>
<tr>
<td>4</td>
<td>aneeshs</td>
</tr>
<tr>
<td>4</td>
<td>argyris</td>
</tr>
<tr>
<td>4</td>
<td>frost</td>
</tr>
<tr>
<td>4</td>
<td>greenberg</td>
</tr>
<tr>
<td>4</td>
<td>jaykreps</td>
</tr>
<tr>
<td>4</td>
<td>mashable</td>
</tr>
<tr>
<td>4</td>
<td>posco</td>
</tr>
<tr>
<td>4</td>
<td>kostas</td>
</tr>
<tr>
<td>4</td>
<td>billgraham</td>
</tr>
<tr>
<td>4</td>
<td>deliprao</td>
</tr>
<tr>
<td>4</td>
<td>larsonite</td>
</tr>
<tr>
<td>4</td>
<td>myy_precious</td>
</tr>
<tr>
<td>4</td>
<td>othman</td>
</tr>
<tr>
<td>4</td>
<td>raffi</td>
</tr>
<tr>
<td>4</td>
<td>sm</td>
</tr>
<tr>
<td>4</td>
<td>sam</td>
</tr>
<tr>
<td>3</td>
<td>Alpa</td>
</tr>
<tr>
<td>3</td>
<td>JShoe</td>
</tr>
<tr>
<td>3</td>
<td>avibryant</td>
</tr>
<tr>
<td>3</td>
<td>sampullara</td>
</tr>
<tr>
<td>3</td>
<td>wangtian</td>
</tr>
<tr>
<td>3</td>
<td>CoreyHyllested</td>
</tr>
<tr>
<td>3</td>
<td>BarackObama</td>
</tr>
<tr>
<td>3</td>
<td>isaach</td>
</tr>
<tr>
<td>3</td>
<td>seekshreyas</td>
</tr>
<tr>
<td>3</td>
<td>dickc</td>
</tr>
<tr>
<td>3</td>
<td>marissamayer</td>
</tr>
<tr>
<td>3</td>
<td>rion</td>
</tr>
<tr>
<td>3</td>
<td>slideshare</td>
</tr>
<tr>
<td>3</td>
<td>timoreilly</td>
</tr>
<tr>
<td>3</td>
<td>hoverbird</td>
</tr>
<tr>
<td>3</td>
<td>lintool</td>
</tr>
<tr>
<td>3</td>
<td>pbrane</td>
</tr>
<tr>
<td>3</td>
<td>sagemintblue</td>
</tr>
<tr>
<td>3</td>
<td>michibusch</td>
</tr>
<tr>
<td>3</td>
<td>TwitterEng</td>
</tr>
<tr>
<td>3</td>
<td>abdur</td>
</tr>
<tr>
<td>3</td>
<td>alethomas</td>
</tr>
<tr>
<td>3</td>
<td>ej</td>
</tr>
<tr>
<td>3</td>
<td>rulsansv</td>
</tr>
<tr>
<td>3</td>
<td>thinkingfish</td>
</tr>
<tr>
<td>2</td>
<td>FakeDorsey</td>
</tr>
</tbody>
</table>
e) I resized the node labels and their color based on the in-degree and plotted the graph of mentions (refer the table above for colors):

f) From the friends and mentions graphs I draw the following conclusions:
   a. The first observation is that going by the mentions, we see that the conversation within this group are more about people outside the central users identified in the interest graph of part2. The most mentioned person is “jco” Jonathan Coveney, Kevin Weil, twitter, square cog, pankaj, zenguali, ottman, rafi which do not form a part of the 6 core friend’s network graph.

UCB Twitter and gilad are the only people that have high number of mentions from within central users.
In this example, it goes to show that, even though a specific group of people might belong to a closely connected network but their conversations might be about completely different things. They do not mention each other that commonly in their conversations.

b. Another very important observation here is that “jco” has the highest in-degree of all the users above, but “jco” was not present in the list of suggestions given to the original seed user (sonali_sh) – which means that
even though “jco” forms a vital part of the conversations within the central users of the friends graph, but it does not appear in the seed user’s interest graph. This vital information can further be used to improve the recommendations provided by the interest graph. Some more users which should be provided as recommended users “twitter”, kevin Weil because they have been mentioned several times in the conversation graph. By including the number of mentions parameter in the interest graph, recommendations on interest can be further improved.

**Source Code**: friendsMentions.java
**Outputfile**: question3newmentions.gv