Report on Hoboken Tweets

Data Preparation

The csv file “Hoboken_tweets.csv” was divided into 5 different files.
This was to divide the data evenly and then perform network analysis on it using Wordij & Gephi. The analysis is done on the 5 files differently and then a combined analysis is done to compare all the 5 files.

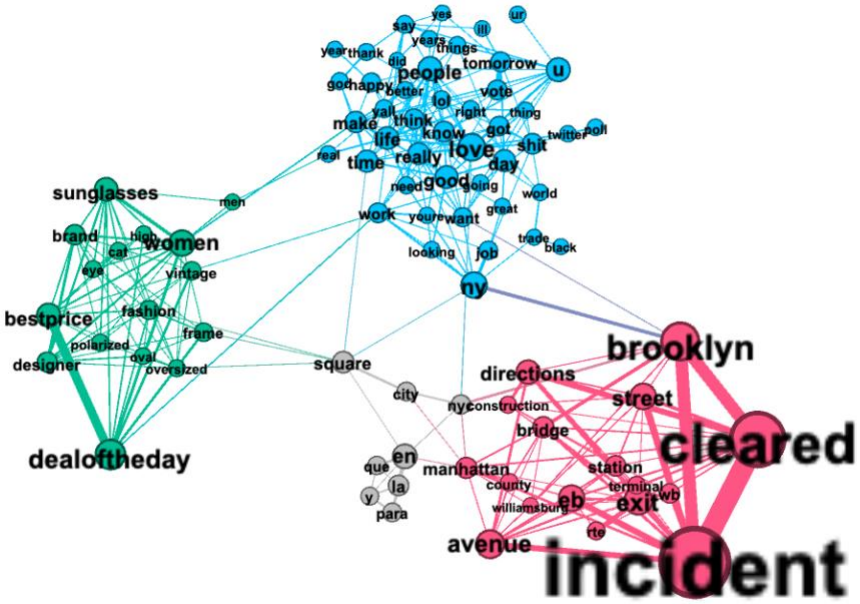
The 5 files were cleaned by using layouts and filters in Gephi to get a discreet network structure.
The commonly used layouts for the network structure are “Force Atlas” and “Force Atlas 2”
The same filter called “Degree Range” was used to remove nodes with lesser degree nodes in order to get a cleaner network structure.
The network structure was partitioned based on Modularity class of each file.
The size of the nodes was ranked according to Weighted degree of each file. Thus giving bigger nodes for high weighted degree and small nodes for low weighted degree.

Analysis

File 1

Structural analysis:

Structure of network is shown below:



The above structure is partitioned with respect to Modularity classes. The size of nodes is ranked according to their "Weighted average."
A "Degree Range" filter is applied before calculating all the statistics metrics.

Interpretation according to the structure above is that, there are 3 main topics that people are talking about here. The topics are as follows:

An incident that took place in New York which probably cleared some avenue and streets in Brooklyn and Manhattan.

Deal of the day for different fashion brands. Where people are talking about the deals and discounts offered on men and women's fashion brands.

The third topic looks a little ambiguous where people are talking about all the positive things like love, work, people and happiness in general.

Statistical Metrics Analysis:

The table shown below depicts the top 5 words based on the highest weighted degree:
Note: The entire excel sheet of statistic metrics is attached in the submission section.

Sr. no	Label	Weighted Degree	Degree	Betweenness Centrality	Modularity Class	Clustering Coefficient
1.	incident	336	12	160.921656	5	0.515152
2.	cleared	254	12	108.894868	5	0.575758
3.	brooklyn	179	9	737.471021	5	0.277778
4.	dealoftheday	106	9	147.623267	2	0.333333
5.	exit	93	12	62.235197	5	0.515152

Weighted degree – Here incident has the highest weighted degree which means it has the maximum number of weighted edges connected to it. This means that most people's tweets are highly related to this word "incident" in this context of analysis.

Betweenness Centrality – It means extent to which a particular node lies on the shortest path between other nodes. Here Brooklyn has the highest betweenness centrality which means it is somehow related to all the topics people are talking about directly or indirectly.

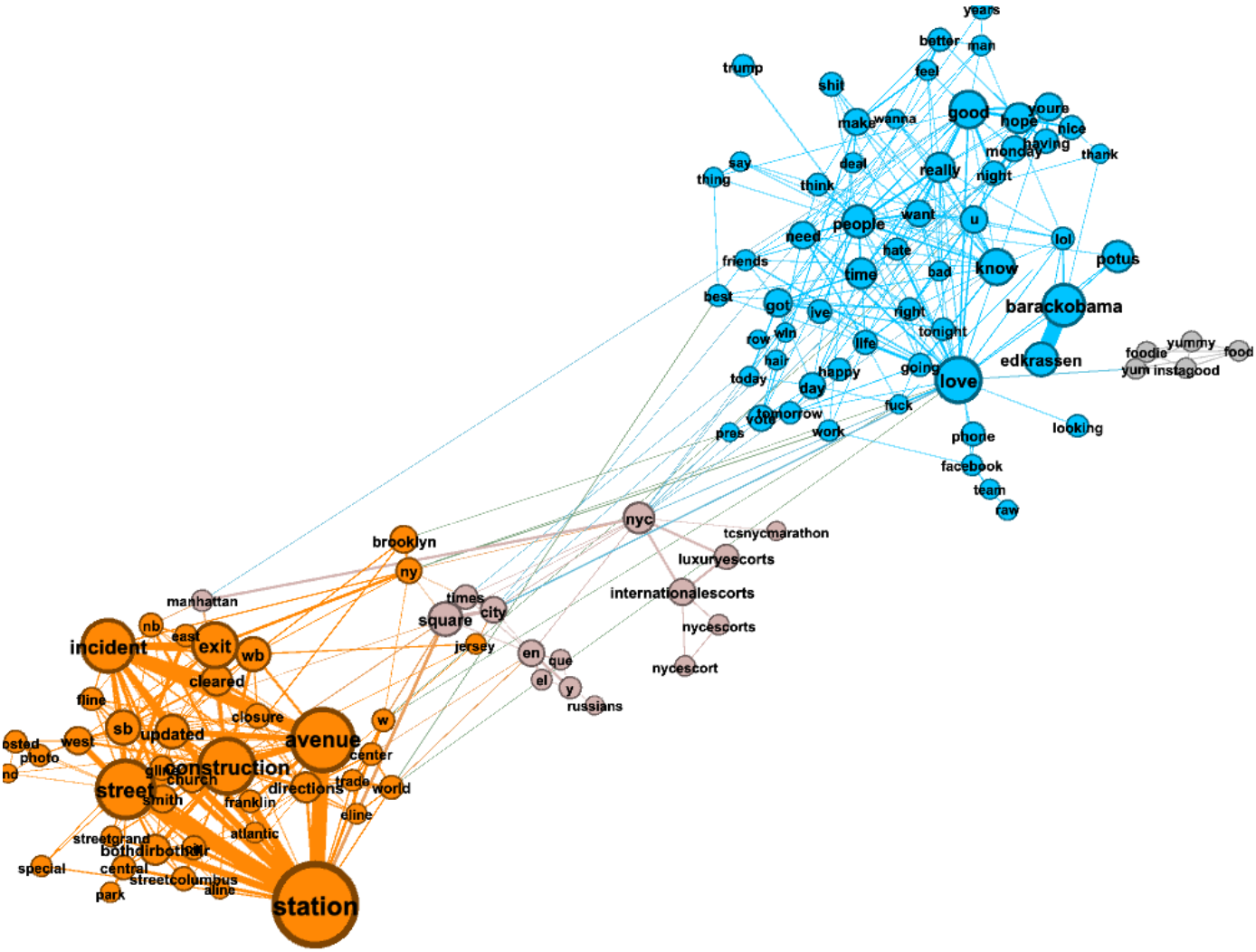
Modularity class – Modularity class are just clusters or groups of different topics(words). This particular file has in total 8 modularity classes.

Clustering coefficient – It means how well are all the other nodes connected to each other. A clustering coefficient of 1 denotes a "small world" phenomenon which means all the nodes are connected to one another in some or the other way and a clustering coefficient of 0 means none of the nodes are connected to one another.

File 2

Structural analysis:

Structure of network is shown below:



The above structure is partitioned with respect to Modularity classes. The size of nodes is ranked according to their “Weighted average.”
A Degree Range filter is applied before calculating all the statistics metrics.

Structure interpretation

Interpretation according to the structure above is that, 2 main topics can be identified that people are talking about here. The topics are as follows:

The first topic is about some station, and again some incident that took place. People are also talking about some "construction"

The second topic is somewhat obscure where people are talking about love, and also about the vote day, where they also talk about Barack Obama whose twitter handle is "Potus 44"

Statistical Metrics Analysis

The table shown below depicts the top 5 words based on the highest weighted degree:

Sr. no	Label	Weighted Degree	Degree	Betweenness Centrality	Modularity Class	Clustering Coefficient
1.	station	35	479	13048.66828	4	0.259259
2.	avenue	27	334	5088.839507	4	0.372549
3.	street	30	304	2740.038657	4	0.25
4.	construction	33	284	11929.97693	4	0.25
5.	incident	20	262	4628.882328	4	0.438095

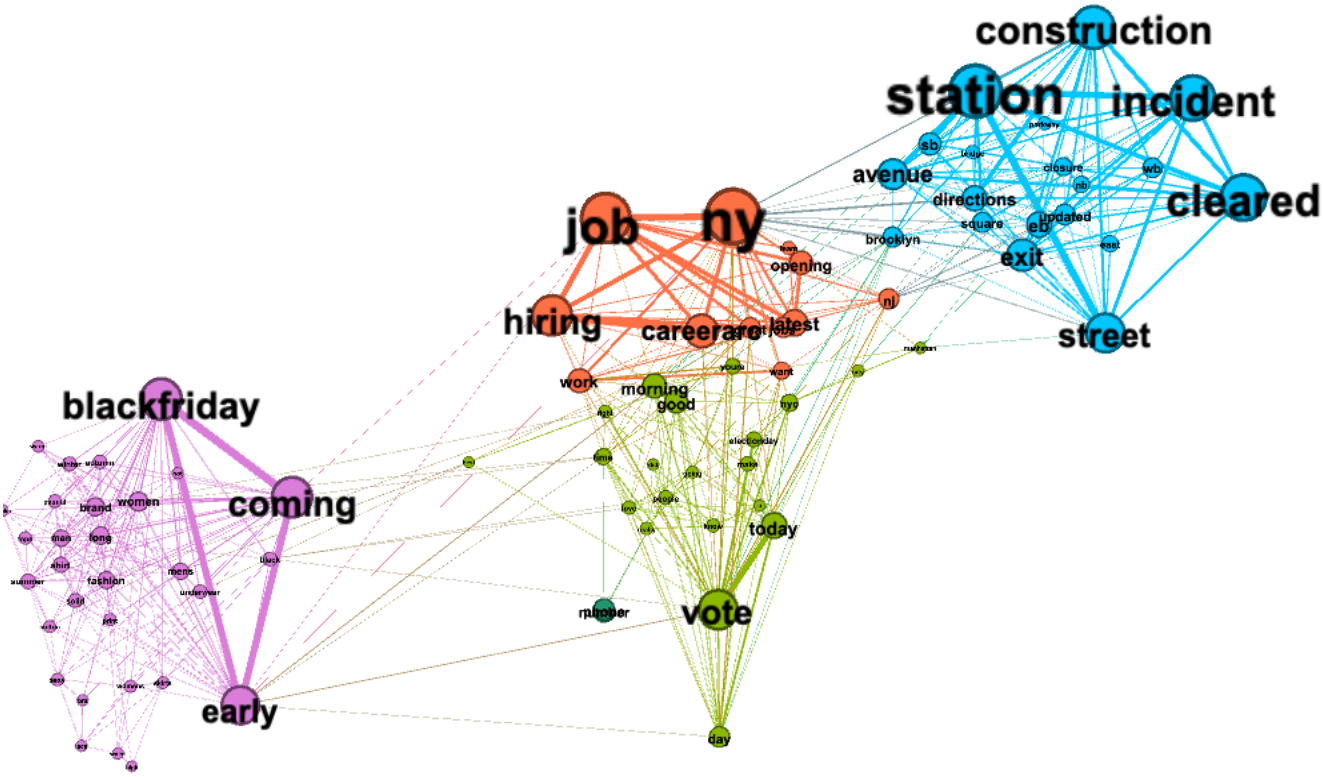
Note: The entire excel sheet of statistic metrics is attached in the submission section.

In the above table the word "station" has the highest weighted degree meaning the maximum tweets by people are related to the word "station."

File 3

Structural analysis:

Structure of network is shown below:



Structural Interpretation:

Interpretation according to the structure above is that, 3 main topics can be identified that people are talking about here. The topics are as follows:

The first topic is about black friday coming early as seen above. This topic might be the evolution to the topic of "deals of the day" from file 1.

The second topic which can be identified is about jobs in NY. This can be clearly seen from the above structure where people are talking about hiring, careerac, work also the "vote" appears here which can be traced back to file 1 again.

The third topic is the same about an incident, station, construction, cleared streets, exits, avenues. A Degree Range filter is applied before calculating all the statistics metrics.

Statistical Metrics Analysis

The table shown below depicts the top 5 words based on the highest weighted degree:

Sr. no	Label	Modularity Class	Degree	Weighted degree	Betweenness Centrality	Clustering Coefficient
1.	ny	4	30	639	749.758548	0.264368
2.	vote	1	27	424	320.030353	0.327635
3.	early	3	27	402	854.565684	0.350427
4.	blackfriday	3	25	451	78.590804	0.433333
5.	coming	3	23	439	173.488383	0.418972

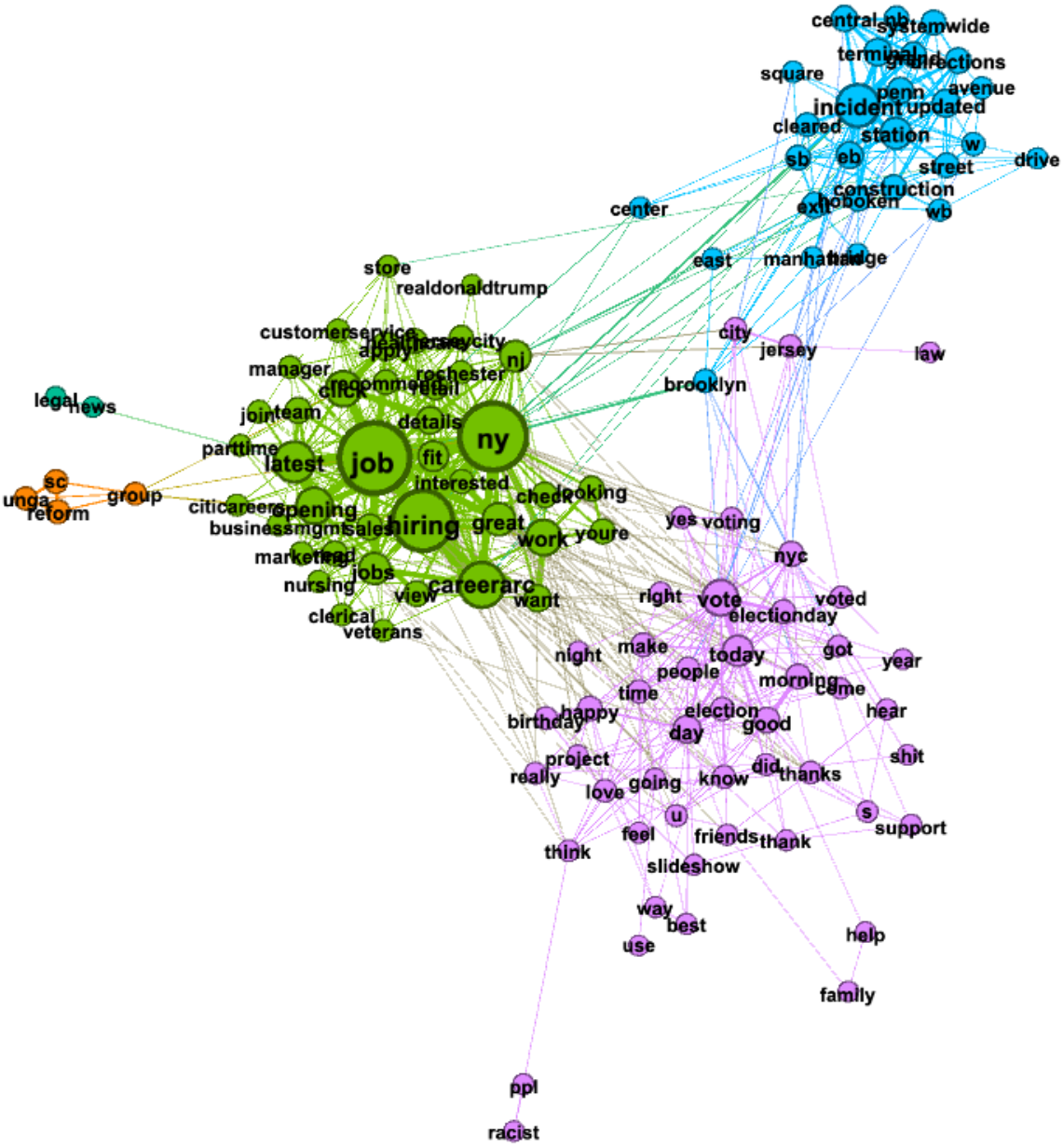
Note: The entire excel sheet of statistic metrics is attached in the submission section.

In the above table the word “ny” has the highest weighted degree meaning the maximum tweets by people are related to the word “ny” for this particular file.

File 4

Structural analysis:

Structure of network is shown below:



Structure Interpretation:

The topics in the above structure are the same as in file 3. People are mostly tweeting about 3 main topics, jobs in NY, incident and the election day.
A Degree Range filter is applied before calculating all the statistics metrics.

Statistical Metrics Analysis

The table shown below depicts the top 5 words based on the highest weighted degree:

Sr. no	Label	Weighted Degree	Degree	Betweenness Centrality	Modularity Class	Clustering Coefficient
1.	job	1317	37	280.652091	1	0.286787
2.	ny	1231	53	2311.63646	1	0.174891
3.	hiring	1049	44	734.607837	1	0.237844
4.	careerarc	649	40	749.592916	1	0.220513
5.	incident	581	27	800.698678	5	0.239316

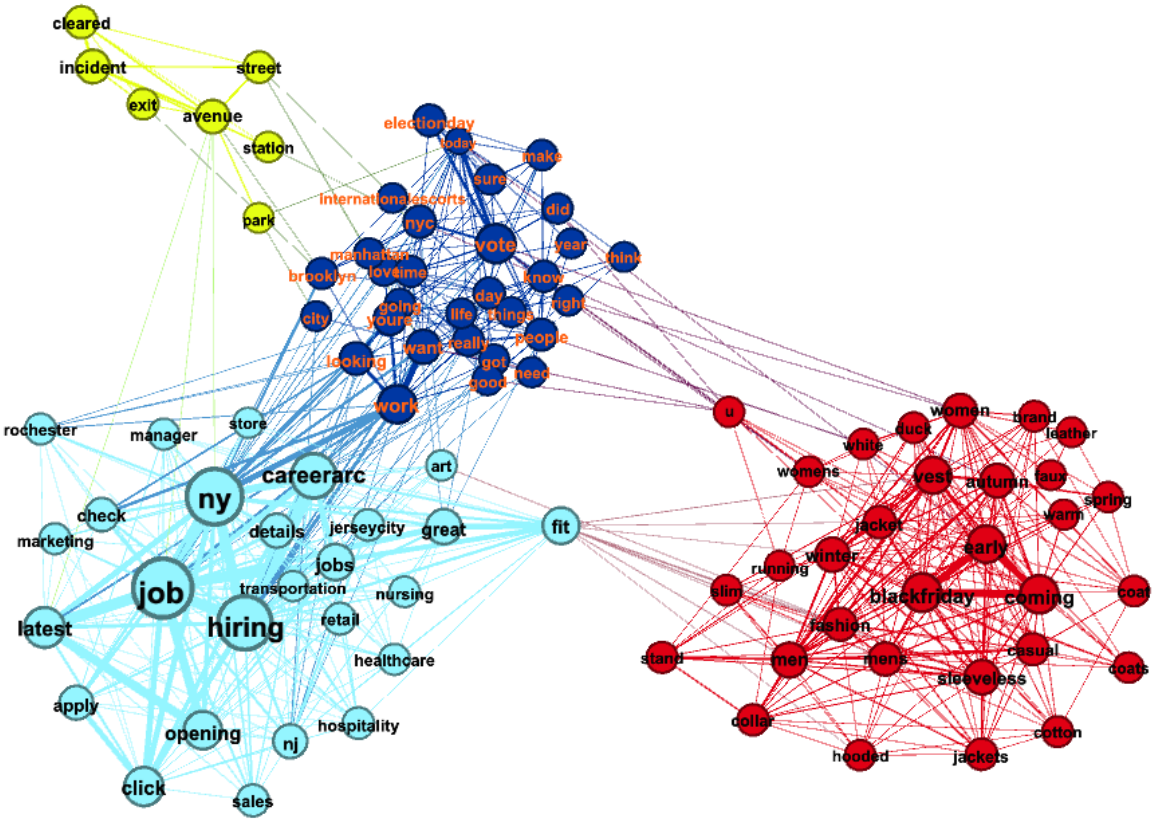
Note: The entire excel sheet of statistic metrics is attached in the submission section.

In the above table the word “job” has the highest weighted degree meaning the maximum tweets by people are related to the word “job” for this particular file.

File 5

Structural analysis:

Structure of network is shown below:



The above structure also depicts similar topics to the above files. The topics which can be easily identified are jobs and hiring in NY, Voting and election day and blackfriday coming early with all the sales and deals.

Statistical Metrics Analysis

The table shown below depicts the top 5 words based on the highest weighted degree:
Note: The entire excel sheet of statistic metrics is attached in the submission section.

Sr. no	Label	Degree	Weighted Degree	Betweenness Centrality	Modularity Class	Clustering Coefficient
1.	job	31	1123	259.597832	0	0.348387
2.	ny	35	980	476.031501	0	0.280672
3.	hiring	34	917	240.168172	0	0.322638
4.	careerarc	28	572	213.461228	0	0.333333
5.	latest	17	412	30.601704	0	0.5

Combined analysis of all the 5 files.

Comparison based on the following parameters:

Number of nodes and edges, clustering coefficient and topics.

The nodes and edges mentioned below are after adding a filter of "degree range" to filter out the nodes with less than approximately 10 degree.

File no.	Nodes	Edges	Clusterring Coefficient	Topics identified
1	99 (2.73% visible)	318 (18.25% visible)	0.451	3
2	125 (3.58% visible)	415 (27.67% visible)	0.466	3
3	84 (2.34% visible)	508 (14.85% visible)	0.571	3
4	125 (3.27% visible)	653 (25.41% visible)	0.528	3
5	111 (2.9% visible)	664 (25.61% visible)	0.552	3

The conversation and topics can be considered cohesive in terms of the topic of "election day" and "voting" as it appears in more than 2 files. Moreover, the conversation about the topic "Black Friday" has an evolution toward "deals of the day" and people talking about different deals that are being offered on men and women fashion.

Conclusion

After analyzing the network of all the 5 files, it can be seen that the virtual conversations between people are mostly based on topics like:

Black Friday: People are seen talking about Black Friday coming early and the different deals are the offered because of it.

Incident: People are seen talking about some incident which has taken place in New York, which has led to clearing of exits, avenues, stations and streets.

Election day: People are seen talking about election day where they talk about voting, barack Obama and election day related things.

Job opportunities: People are seen talking about the job openings in New York and Jersey City, different groups and websites associated with the same.