Homework 5

DATA SCIENCE AND KNOWLEDGE DISCOVERY

Rushabh Barbhaya | CWID 10427219 | 4/22/2018

Data Understanding

Our goal is to perform text analysis on Hoboken_tweets.csv. The dataset contains the following information

- Username
- Timestamp
- Tweet

We need to perform network analysis on tweets and find patterns on common issues faced by citizens of Hoboken and present them in a network diagram.

Software Used

We were asked to clean the tweets and perform analysis on 5 bins of data. I used the following software:

1. Excel

   Used excel to clean data and generate bins.

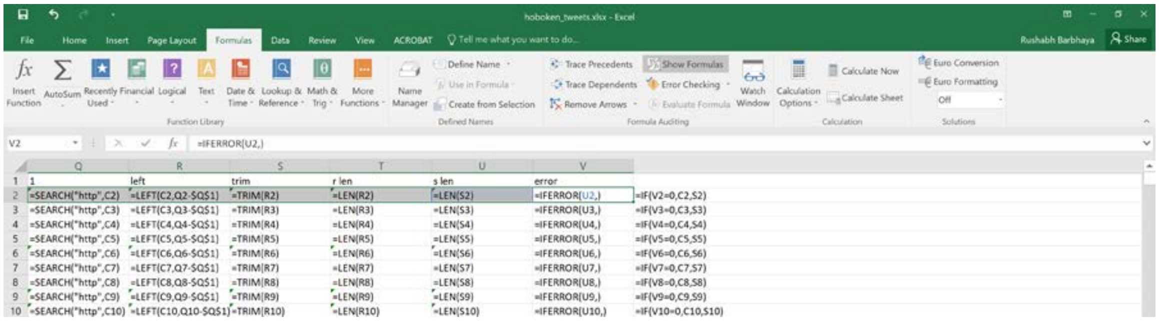
2. Wordij

   Used Wordij to perform network connections

3. Gephi

   Used to create modularity class. Applied data range filter and generated a network diagram.

Data Preparation



V2 =IFERROR(U2,)

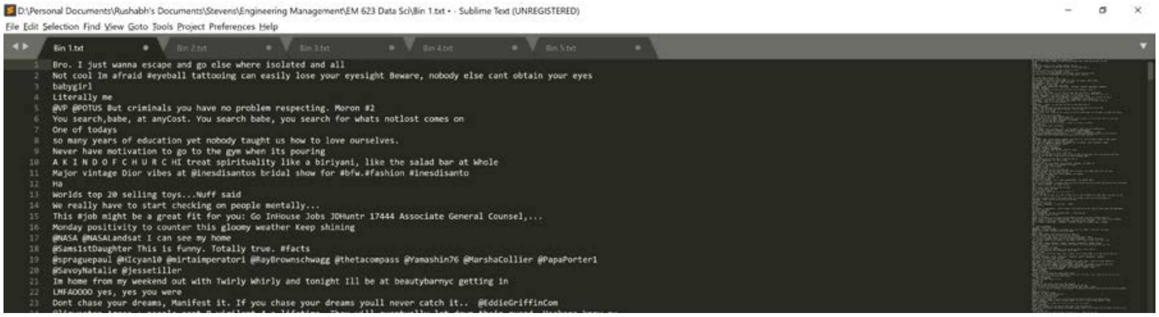
	Q	R	S	T	U	V	
1	1	left	trim	r len	s len	error	
2	=SEARCH("http",C2)	=LEFT(C2,Q2-Q1)	=TRIM(R2)	=LEN(R2)	=LEN(S2)	=IFERROR(U2,)	=IF(V2=0,C2,S2)
3	=SEARCH("http",C3)	=LEFT(C3,Q3-Q1)	=TRIM(R3)	=LEN(R3)	=LEN(S3)	=IFERROR(U3,)	=IF(V3=0,C3,S3)
4	=SEARCH("http",C4)	=LEFT(C4,Q4-Q1)	=TRIM(R4)	=LEN(R4)	=LEN(S4)	=IFERROR(U4,)	=IF(V4=0,C4,S4)
5	=SEARCH("http",C5)	=LEFT(C5,Q5-Q1)	=TRIM(R5)	=LEN(R5)	=LEN(S5)	=IFERROR(U5,)	=IF(V5=0,C5,S5)
6	=SEARCH("http",C6)	=LEFT(C6,Q6-Q1)	=TRIM(R6)	=LEN(R6)	=LEN(S6)	=IFERROR(U6,)	=IF(V6=0,C6,S6)
7	=SEARCH("http",C7)	=LEFT(C7,Q7-Q1)	=TRIM(R7)	=LEN(R7)	=LEN(S7)	=IFERROR(U7,)	=IF(V7=0,C7,S7)
8	=SEARCH("http",C8)	=LEFT(C8,Q8-Q1)	=TRIM(R8)	=LEN(R8)	=LEN(S8)	=IFERROR(U8,)	=IF(V8=0,C8,S8)
9	=SEARCH("http",C9)	=LEFT(C9,Q9-Q1)	=TRIM(R9)	=LEN(R9)	=LEN(S9)	=IFERROR(U9,)	=IF(V9=0,C9,S9)
10	=SEARCH("http",C10)	=LEFT(C10,Q10-Q1)	=TRIM(R10)	=LEN(R10)	=LEN(S10)	=IFERROR(U10,)	=IF(V10=0,C10,S10)

Figure 1 Data preparation

Used these formulas to remove the link form tweets and remove empty tweets from the dataset. SEARCH function was used to find http from the dataset and it returns the character location Used LEFT function was used to return #VALUE! For if there is no value before http. Used TRIM function to remove blank spaces. Used IFERROR to check the presence of http. Finally, used IF function to print clean text.

Bins of data

Crated 5 bins of data in text editor of 12000 tweets/bin.


D:\Personal Documents\Rushabh's Documents\Stevens\Engineering Management\EM 623 Data Sci\Bin 1.txt • - Sublime Text (UNREGISTERED)
File Edit Selection Find View Goto Tools Project Preferences Help
Bin 1.txt
Bin 2.txt
Bin 3.txt
Bin 4.txt
Bin 5.txt

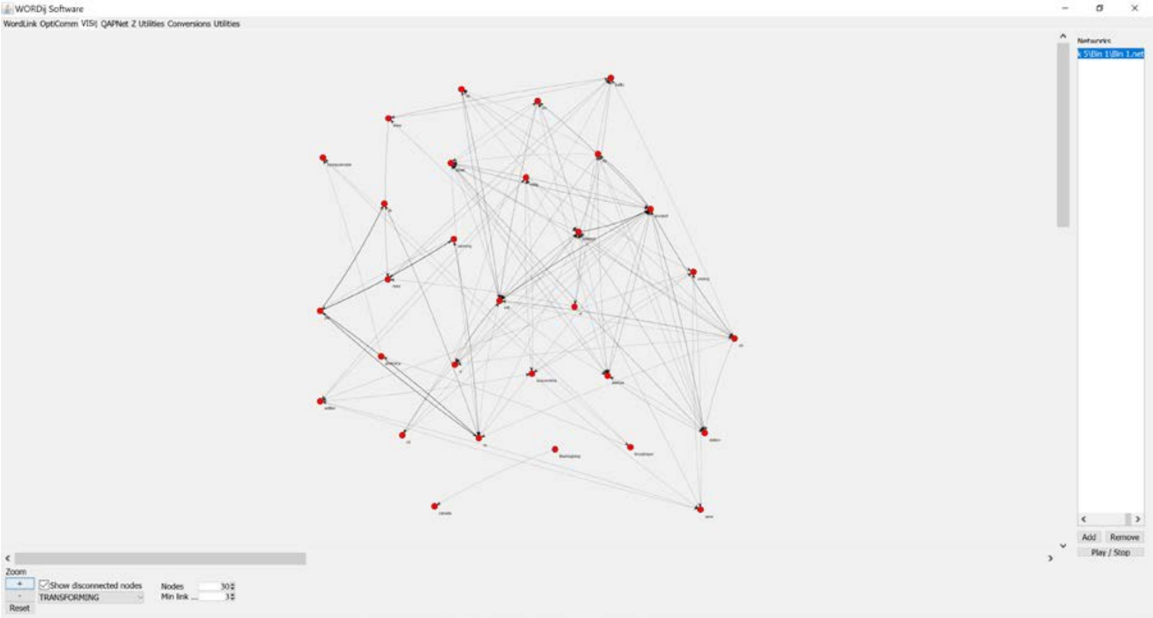
```
Bro. I just wanna escape and go else where isolated and all
Not cool Im afraid #eyeball tattooing can easily lose your eyesight Beware, nobody else cant obtain your eyes
babygirl
Literally me
@VP @POTUS But criminals you have no problem respecting. Moron #2
You search,babe, at anyCost. You search babe, you search for whats notlost comes on
One of todays
so many years of education yet nobody taught us how to love ourselves.
Never have motivation to go to the gym when its pouring
A K I N D O F C H U R C HI treat spirituality like a biriyani, like the salad bar at Whole
Major vintage Dior vibes at @inesdisantos bridal show for #bfw.#fashion #inesdisanto
Ha
Worlds top 20 selling toys...Nuff said
We really have to start checking on people mentally...
This #job might be a great fit for you: Go InHouse Jobs JDHuntr 17444 Associate General Counsel,...
Monday positivity to counter this gloomy weather Keep shining
@NASA @NASALandsat I can see my home
@Sams1stDaughter This is funny. Totally true. #facts
@spraguepaul @HIcyan10 @mirtaimperatori @RayBrownschwagg @thetacompass @Yamashin76 @MarshaCollier @PapaPorter1
@SavoyNatalie @jessetiller
Im home from my weekend out with Twirly Whirly and tonight Ill be at beautybarnyc getting in
LMFAOOOO yes, yes you were
Dont chase your dreams, Manifest it. If you chase your dreams youll never catch it..  @EddieGriffinCom
```

Figure 2 Sample of 5 bins

Saved these data as text files for worij.

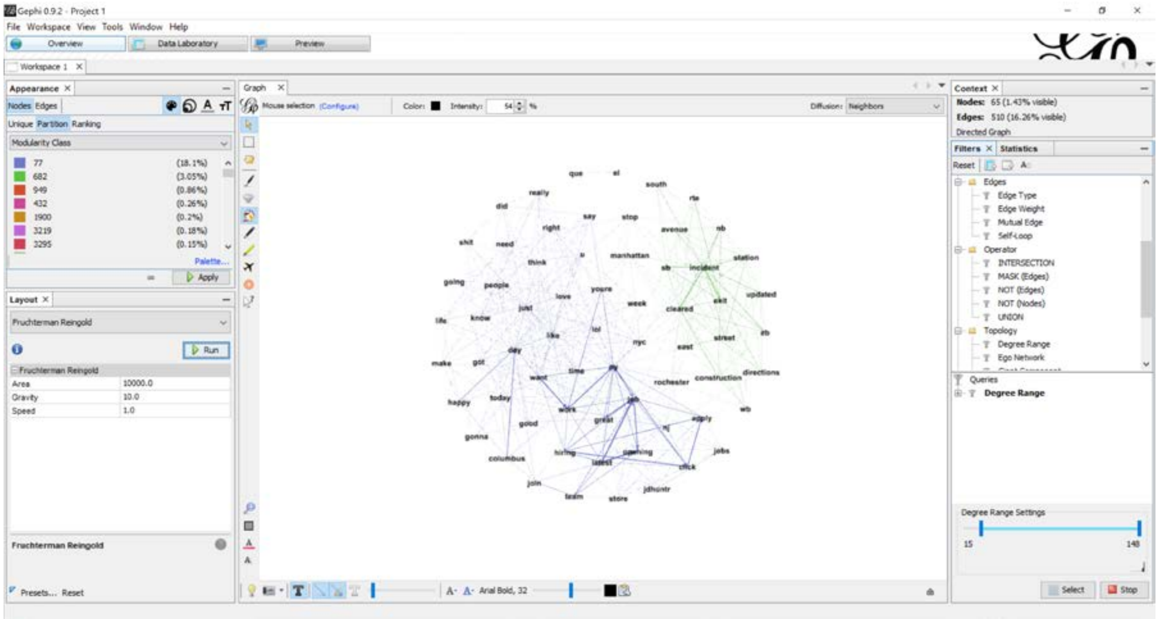
Bin analysis

BIN 1



☑ Show disconnected nodes
TRANSFORMING
Nodes 30
Min link 3

Figure 3 Wordij Network Diagram



Modularity Class

🟦	77	(18.1%)
🟩	682	(3.05%)
🟥	949	(0.86%)
🟪	432	(0.26%)
🟧	1900	(0.2%)
🟪	3219	(0.18%)
🟥	3295	(0.15%)

Fruchterman Reingold

Fruchterman Reingold	
Area	10000.0
Gravity	10.0
Speed	1.0

Color: ■ Intensity: 54 %

Diffusion: Neighbors

A· A· Arial Bold, 32

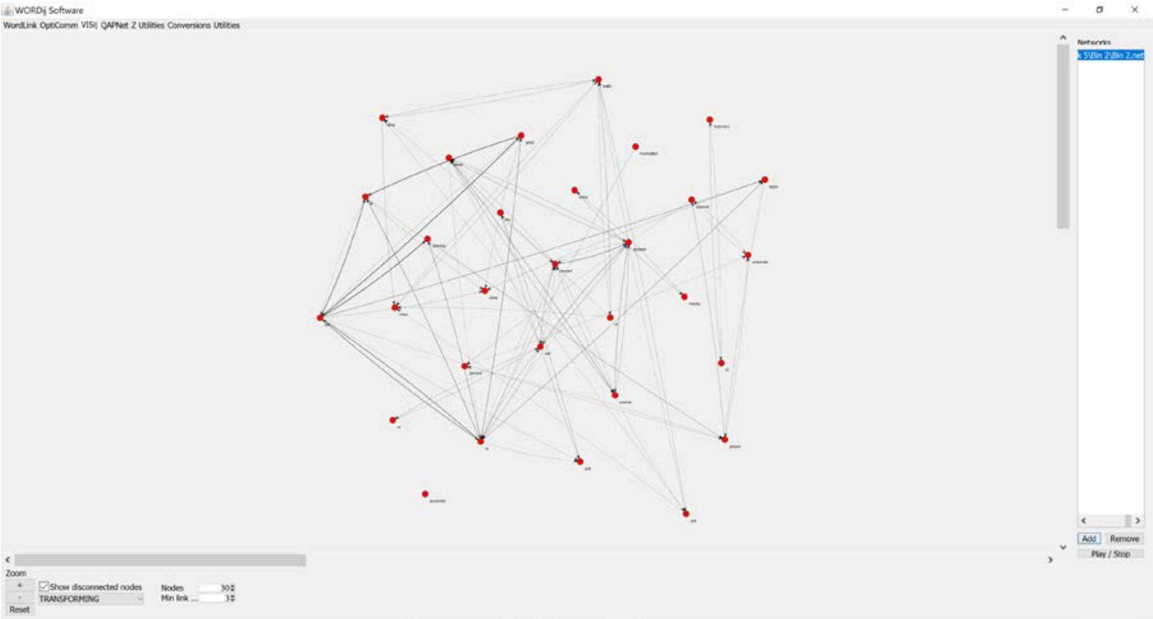
Nodes: 65 (1.43% visible)
Edges: 510 (16.26% visible)
Directed Graph

Degree Range Settings
15 148

Figure 4 Gephi network diagram

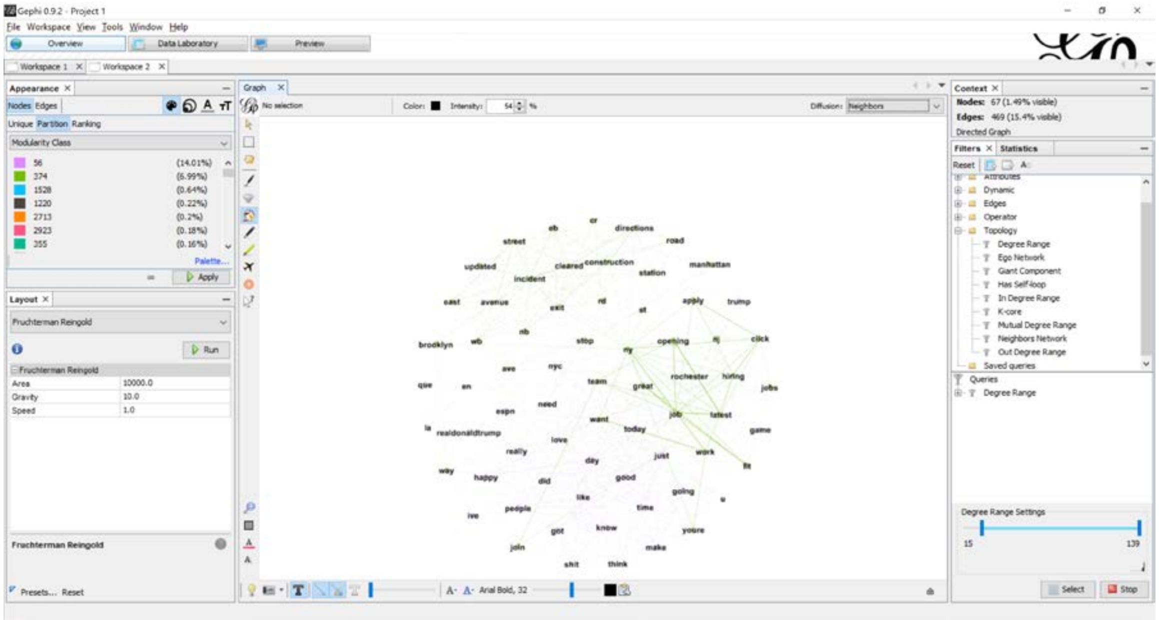
Generated modularity diagram. I got multiple modularity sets. To reduce the connections, I reduced the range using degree range filter. Applied Fruchterman layout to get a cleaner representation of data.

BIN 2



☑ Show disconnected nodes
TRANSFORMING
Nodes 30
Min link 3

Figure 5 Wordij network diagram



Modularity Class

🟪	56	(14.01%)
🟩	374	(6.99%)
🟦	1528	(0.64%)
⬛	1220	(0.22%)
🟧	2713	(0.2%)
🟥	2923	(0.18%)
🟩	355	(0.16%)

Fruchterman Reingold

Fruchterman Reingold	
Area	10000.0
Gravity	10.0
Speed	1.0

Color: ■ Intensity: 54 %

Diffusion: Neighbors

Arial Bold, 32

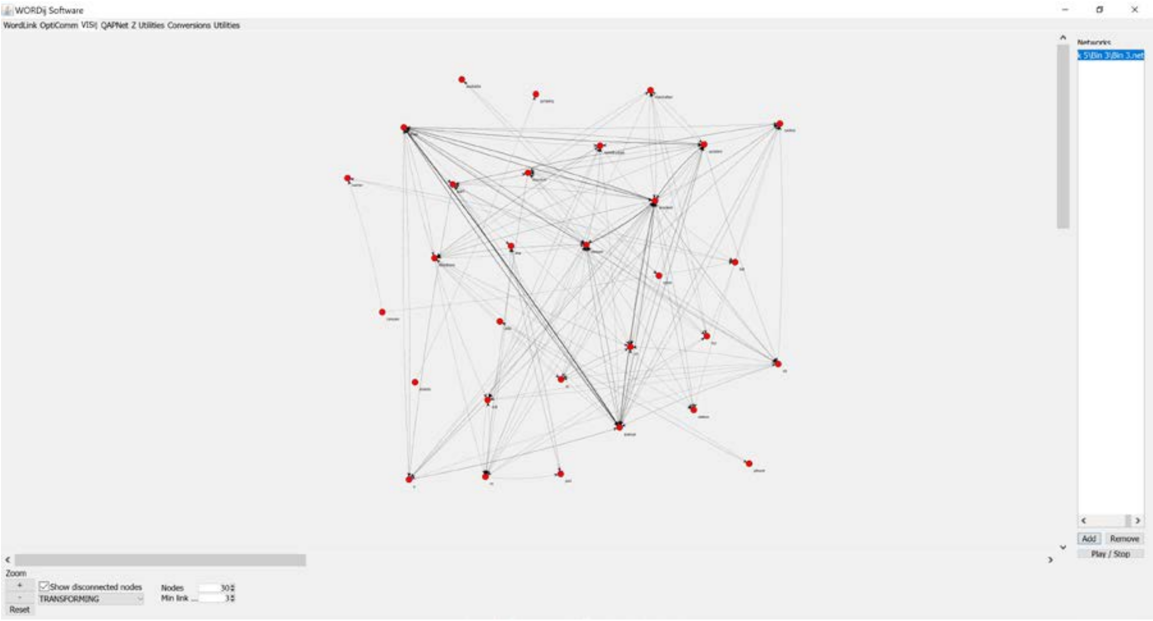
Nodes: 67 (1.49% visible)
Edges: 469 (15.4% visible)
Directed Graph

Degree Range Settings
15 139

Figure 6 Gephi network diagram

Generated modularity diagram. I got multiple modularity sets. To reduce the connections, I reduced the range using degree range filter. Applied Fruchterman layout to get a cleaner representation of data.

BIN 3



WORDij Software

WordLink OptiComm VISij QAPNet Z Utilities Conversions Utilities
Networks
Bin 3\Bin 3.net
Add
Remove
Play / Stop
Zoom
+
-
Reset

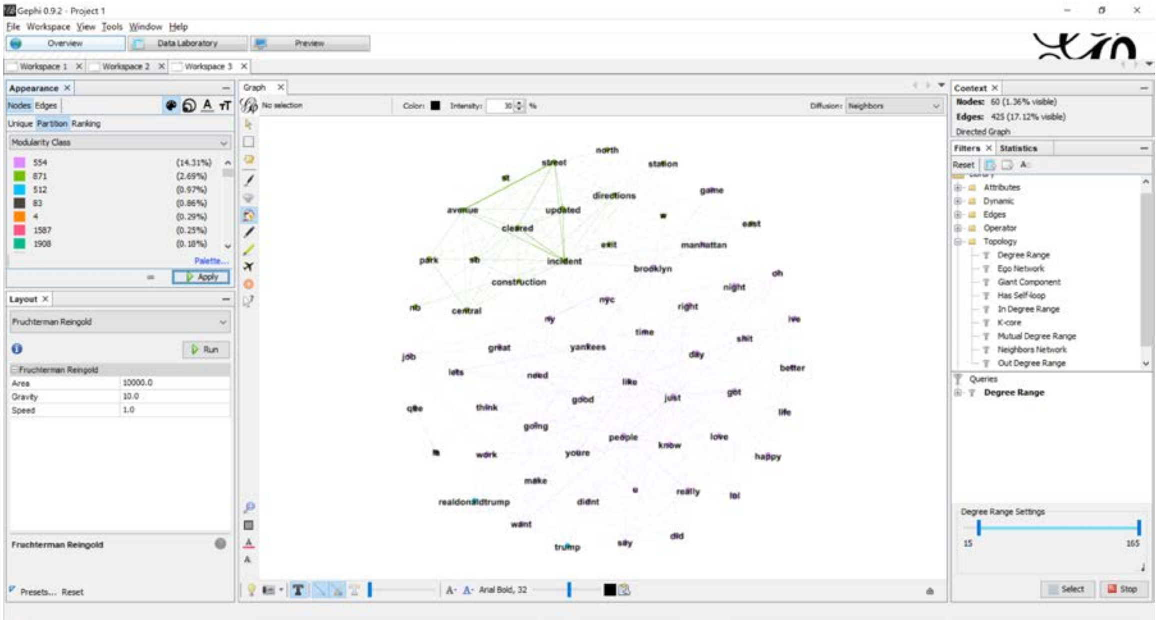
☑ Show disconnected nodes

TRANSFORMING

Nodes 30

Min link ... 3

Figure 7 Wordij Network diagram



Gephi 0.9.2 - Project 1

Modularity Class

🟪	554	(14.31%)
🟩	871	(2.69%)
🟦	512	(0.97%)
⬛	83	(0.86%)
🟧	4	(0.29%)
🟥	1587	(0.25%)
🟩	1908	(0.18%)

Fruchterman Reingold

Fruchterman Reingold	
Area	10000.0
Gravity	10.0
Speed	1.0

Context

Nodes: 60 (1.36% visible)

Edges: 425 (17.12% visible)

Directed Graph

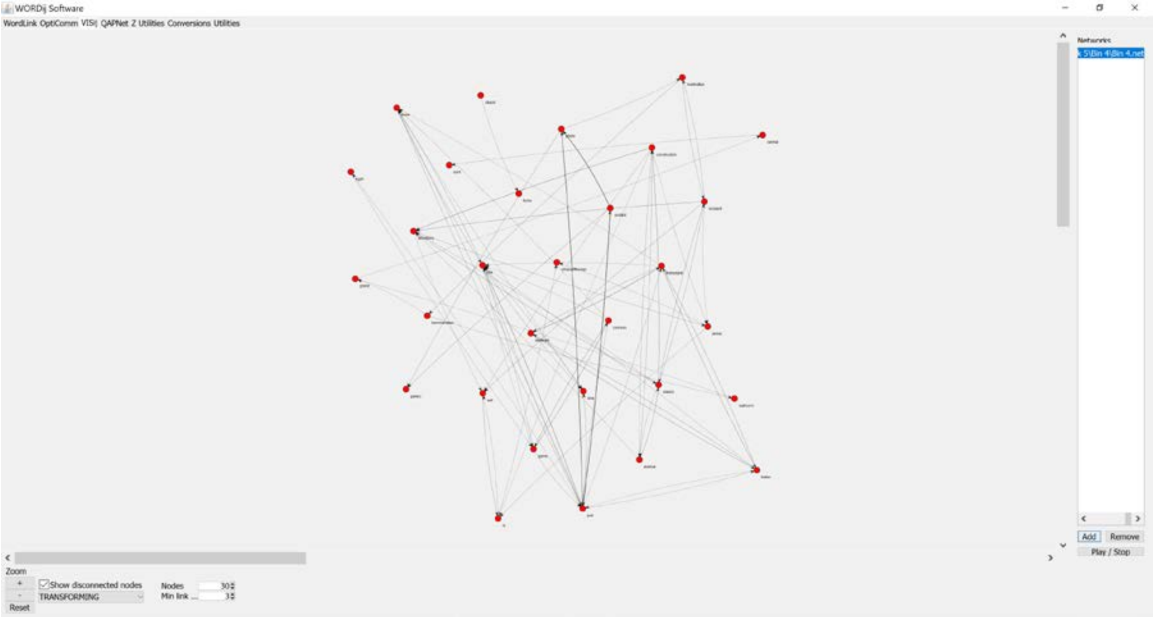
Degree Range

Degree Range Settings: 15 – 165

Figure 8 Gephi Network diagram

Generated modularity diagram. I got multiple modularity sets. To reduce the connections, I reduced the range using degree range filter. Applied Fruchterman layout to get a cleaner representation of data.

BIN 4

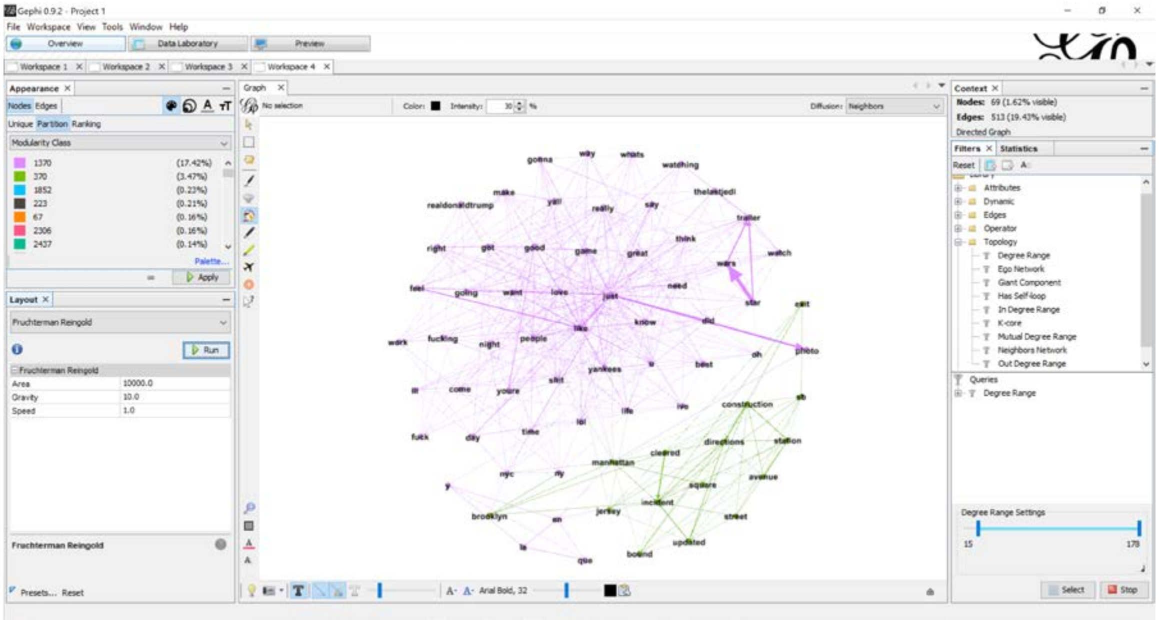


WORDij Software

WordLink OptiComm VISij QAPNet Z Utilities Conversions Utilities
Networks
Bin 4\Bin 4.net
Add
Remove
Play / Stop
Zoom
+
-
Reset

☑ Show disconnected nodes
TRANSFORMING
Nodes 30
Min link ... 3

Figure 9 Wordij Network Diagram



Gephi 0.9.2 - Project 1

Modularity Class

🟪	1370	(17.42%)
🟩	370	(3.47%)
🟦	1852	(0.23%)
⬛	223	(0.21%)
🟧	67	(0.16%)
🟥	2306	(0.16%)
🟩	2437	(0.14%)

Fruchterman Reingold

Fruchterman Reingold	
Area	10000.0
Gravity	10.0
Speed	1.0

No selection Color: ■ Intensity: 30 %

Diffusion: Neighbors

Arial Bold, 32

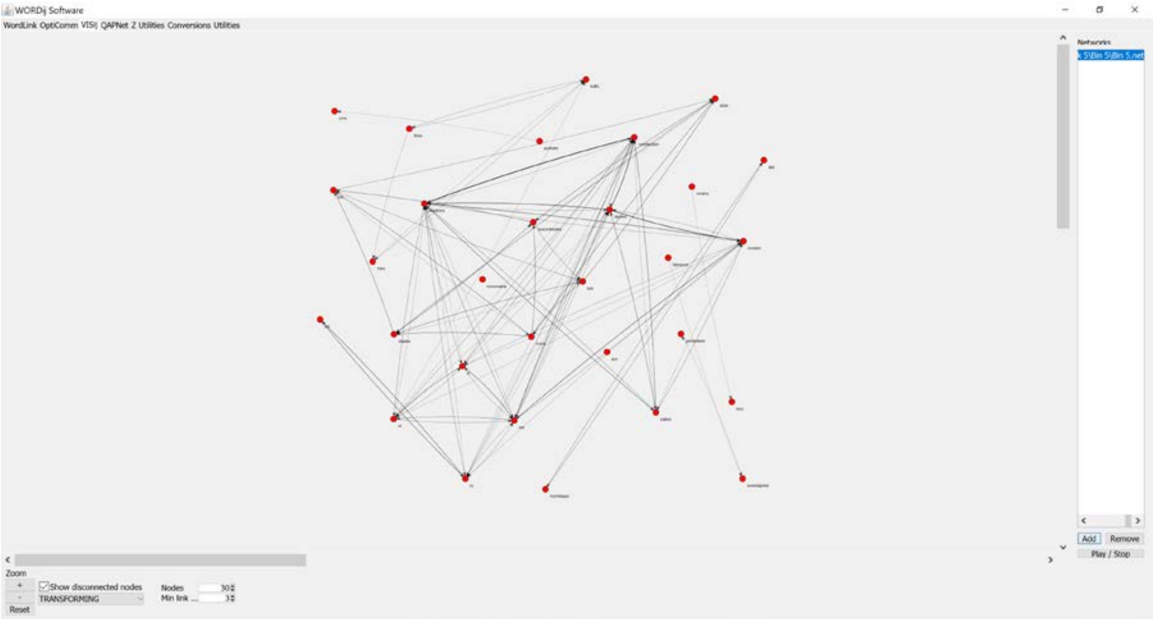
Context
Nodes: 69 (1.62% visible)
Edges: 513 (19.43% visible)
Directed Graph

Degree Range Settings
15 178

Figure 10 Gephi Network diagram

Generated modularity diagram. I got multiple modularity sets. To reduce the connections, I reduced the range using degree range filter. Applied Fruchterman layout to get a cleaner representation of data.

BIN 5



WORDij Software

Networks: 5\Bin 5\Bin 5.net

☑ Show disconnected nodes

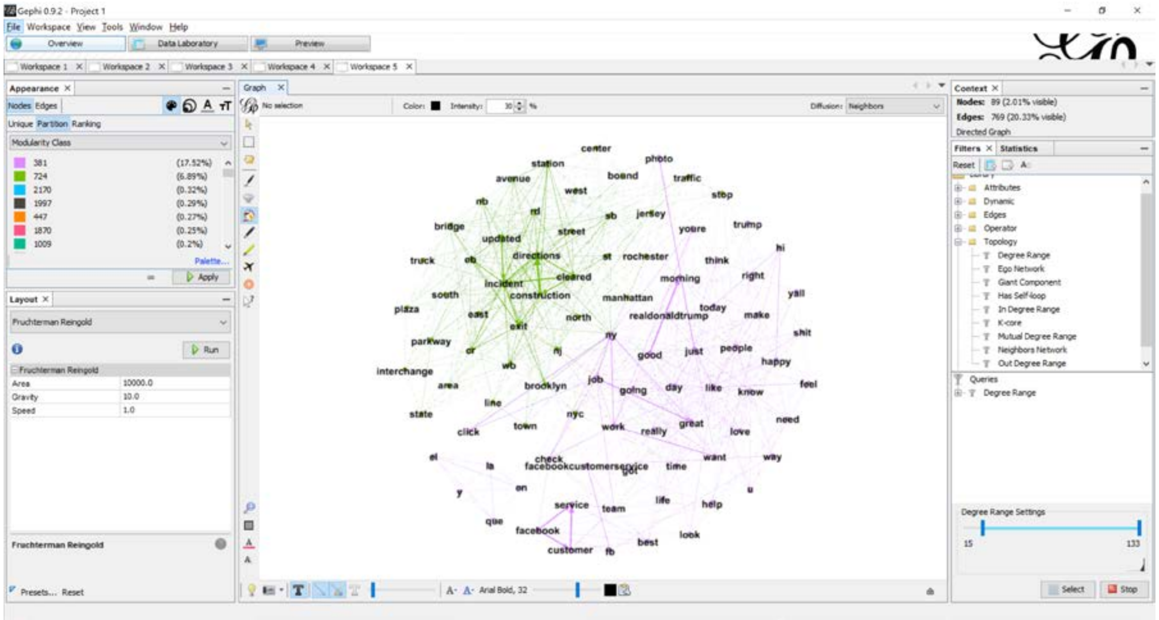
TRANSFORMING

Nodes: 30

Min link: 3

WordLink OptiComm VISij QAPNet Z Utilities Conversions Utilities
Add
Remove
Play / Stop
Zoom
+
-
Reset

Figure 11 Wordij Network Diagram



Gephi 0.9.2 - Project 1

Appearance: Nodes — Partition — Modularity Class

🟪	381	(17.52%)
🟩	724	(6.89%)
🟦	2170	(0.32%)
⬛	1997	(0.29%)
🟧	447	(0.27%)
🟥	1870	(0.25%)
🟩	1009	(0.2%)

Layout: Fruchterman Reingold

Area	10000.0
Gravity	10.0
Speed	1.0

Context

Nodes: 89 (2.01% visible)

Edges: 769 (20.33% visible)

Directed Graph

Queries: Degree Range

Degree Range Settings: 15 – 133

Figure 12 Gephi Network Diagram

Generated modularity diagram. I got multiple modularity sets. To reduce the connections, I reduced the range using degree range filter. Applied Fruchterman layout to get a cleaner representation of data.

Conclusion

Although there were some tweets regarding construction. There was more focus on a yankee game and converations regarding President Donald Trump.