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.

1. Wordij

Used Wordij to perform network connections

1. Gephi

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

# Data Preparation

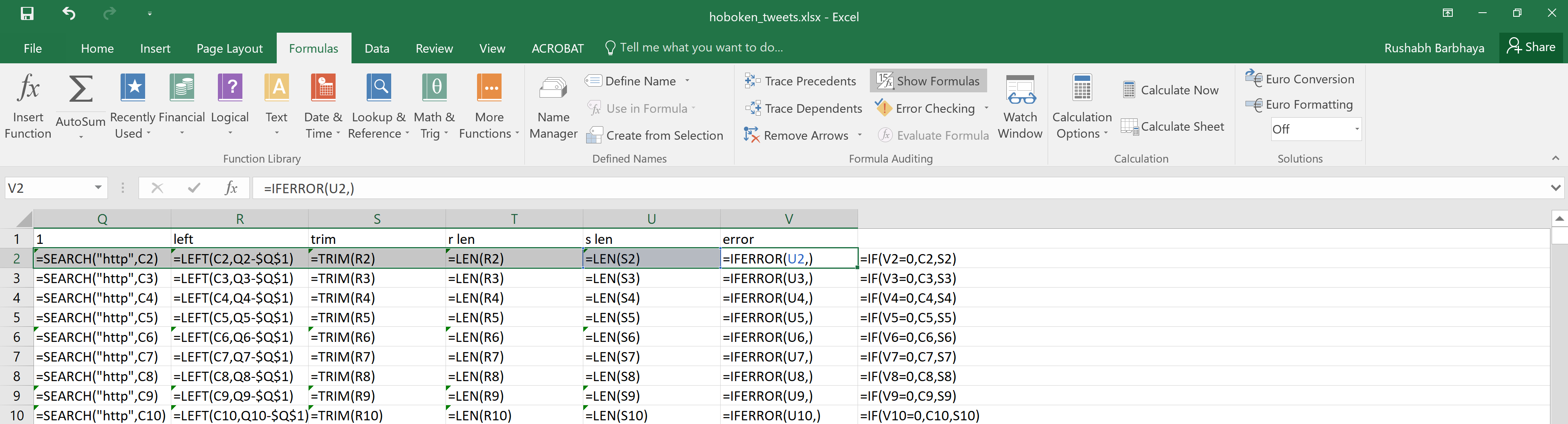


Figure Data preparation

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.

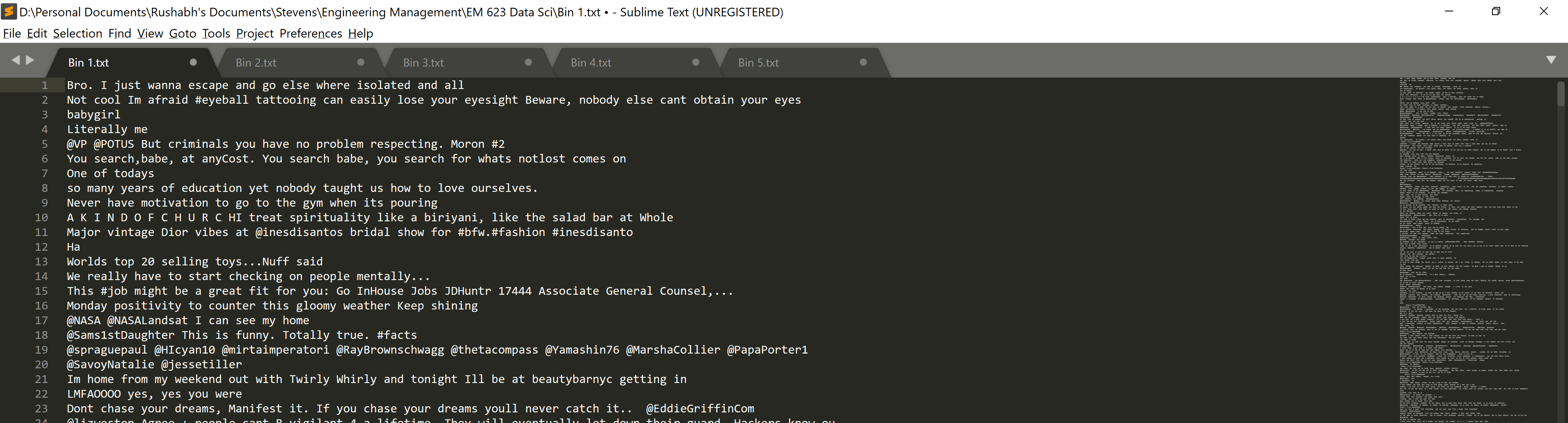


Figure Sample of 5 bins

Saved these data as text files for worij.

# Bin analysis

## Bin 1

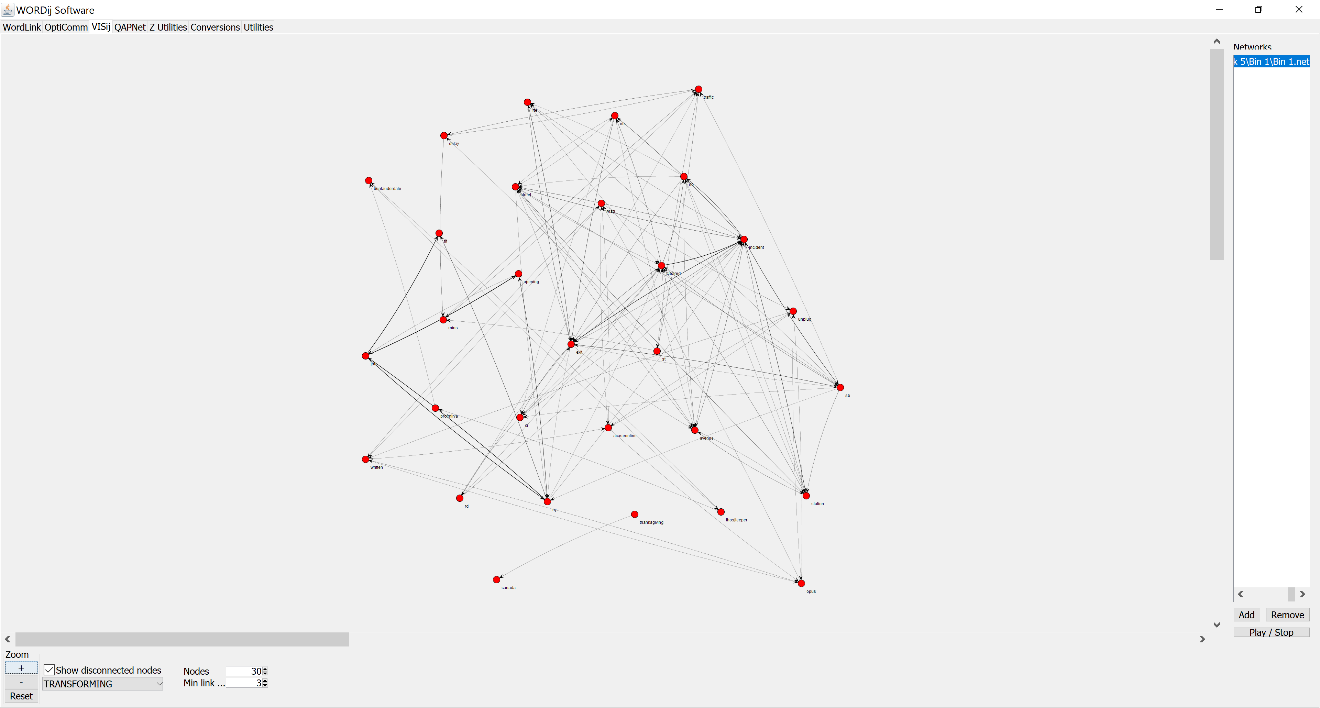


Figure Wordij Network Diagram

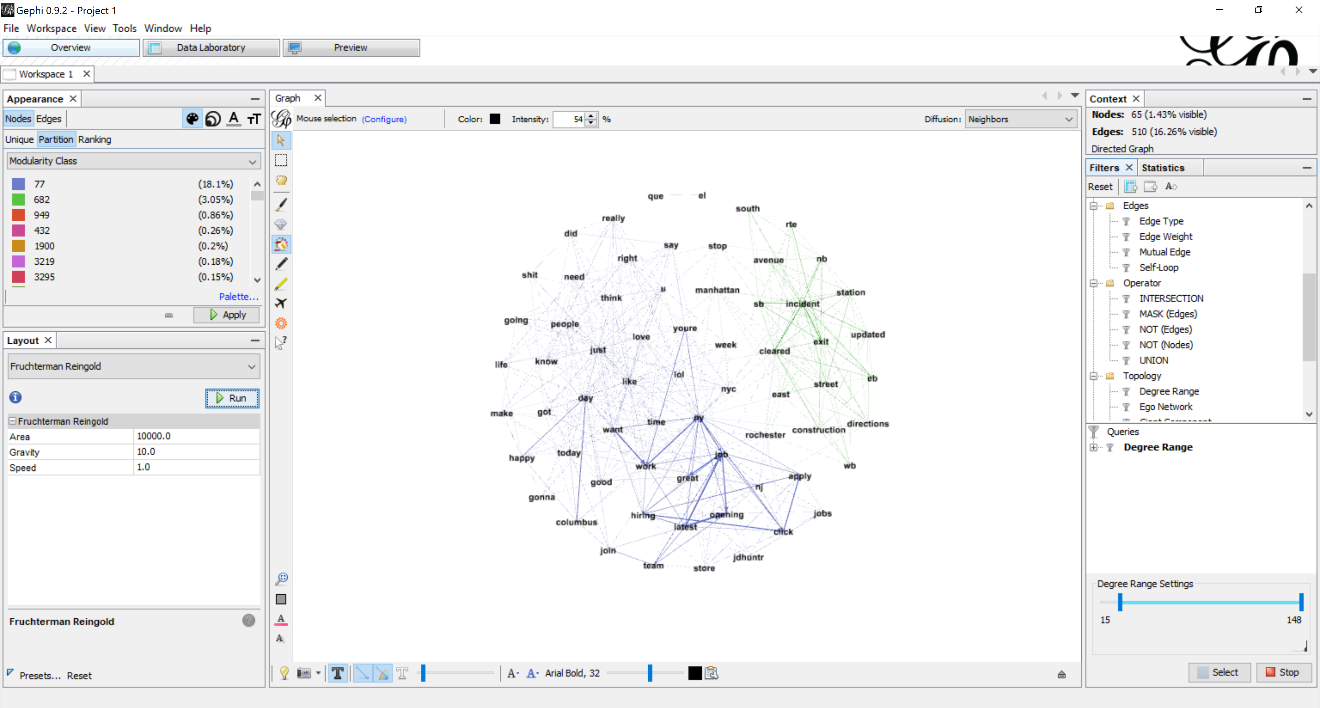


Figure Gephi network diagram

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

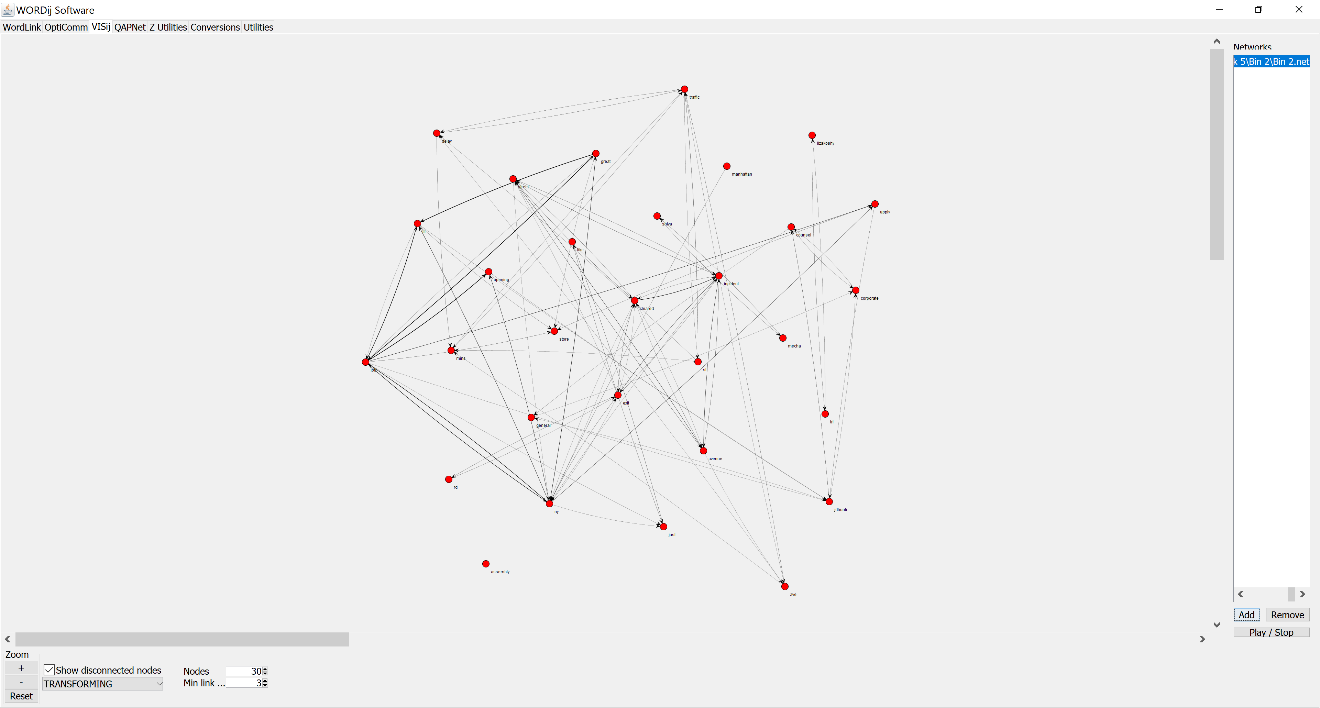


Figure Wordij network diagram

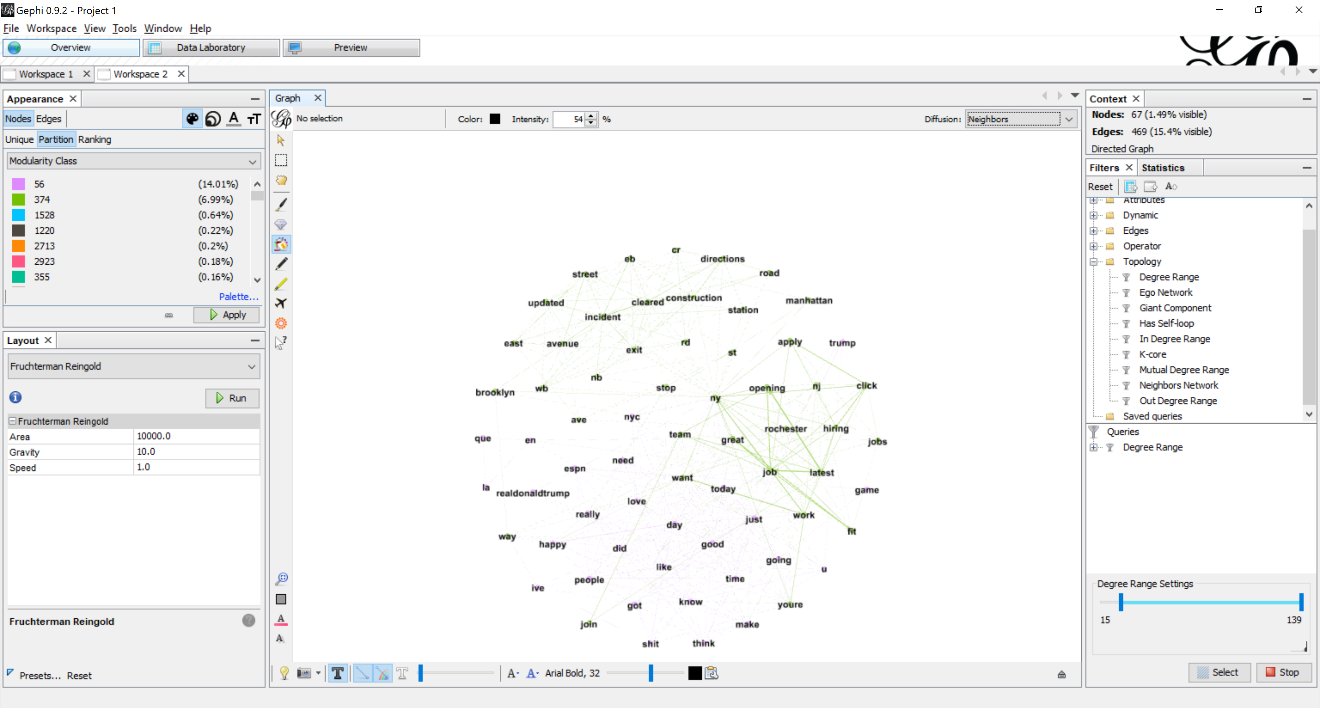


Figure Gephi network diagram

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

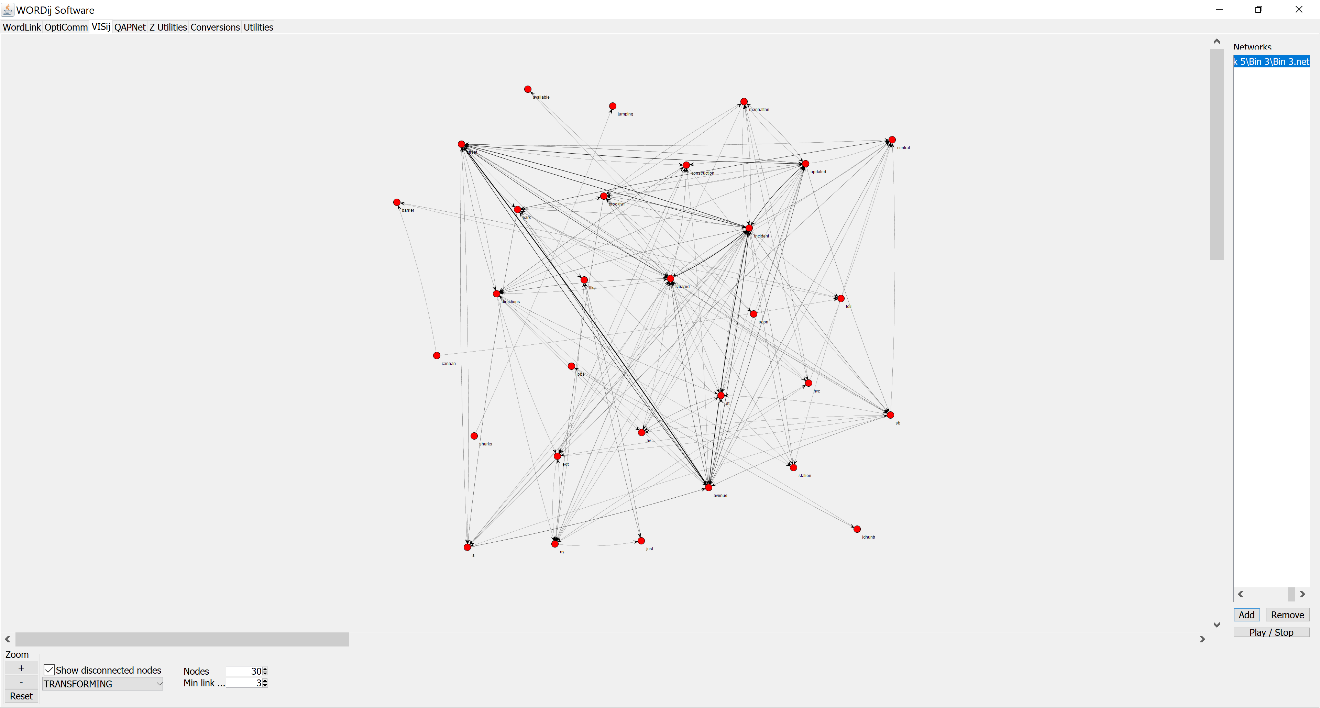


Figure Wordij Network diagram

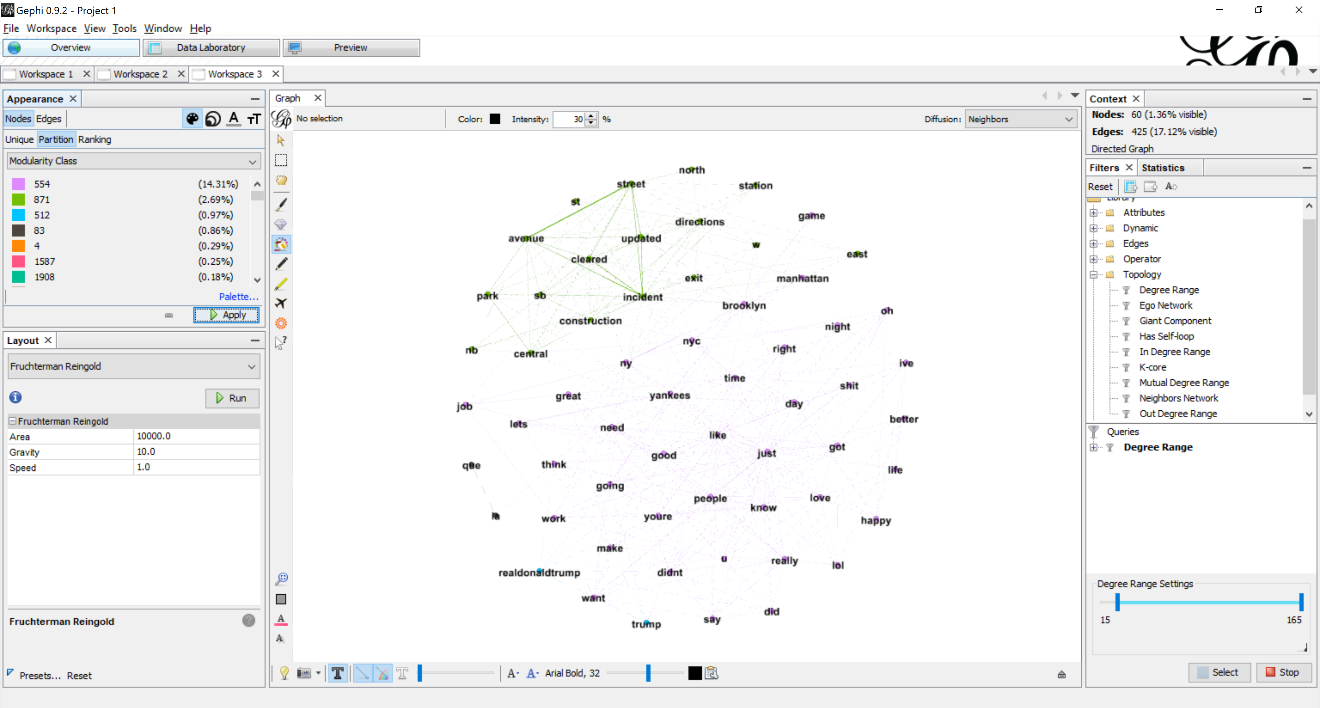


Figure Gephi Network diagram

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

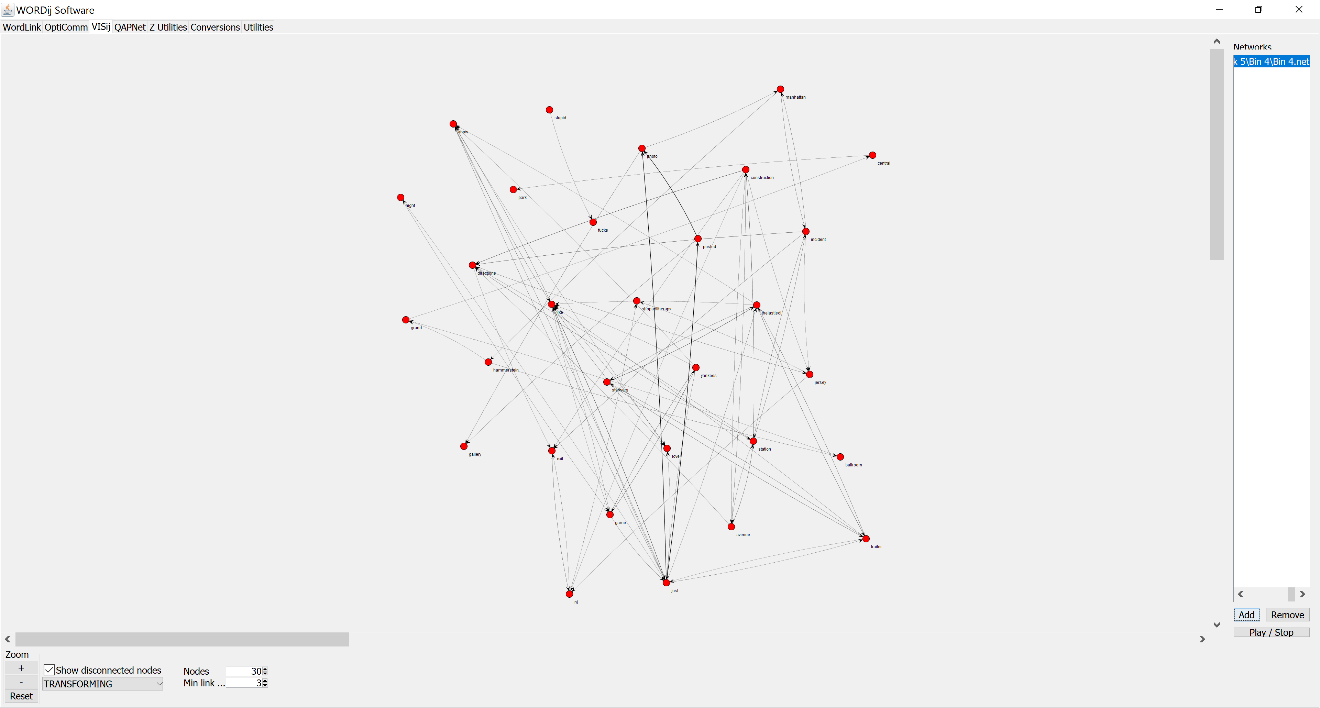


Figure Wordij Network Diagram

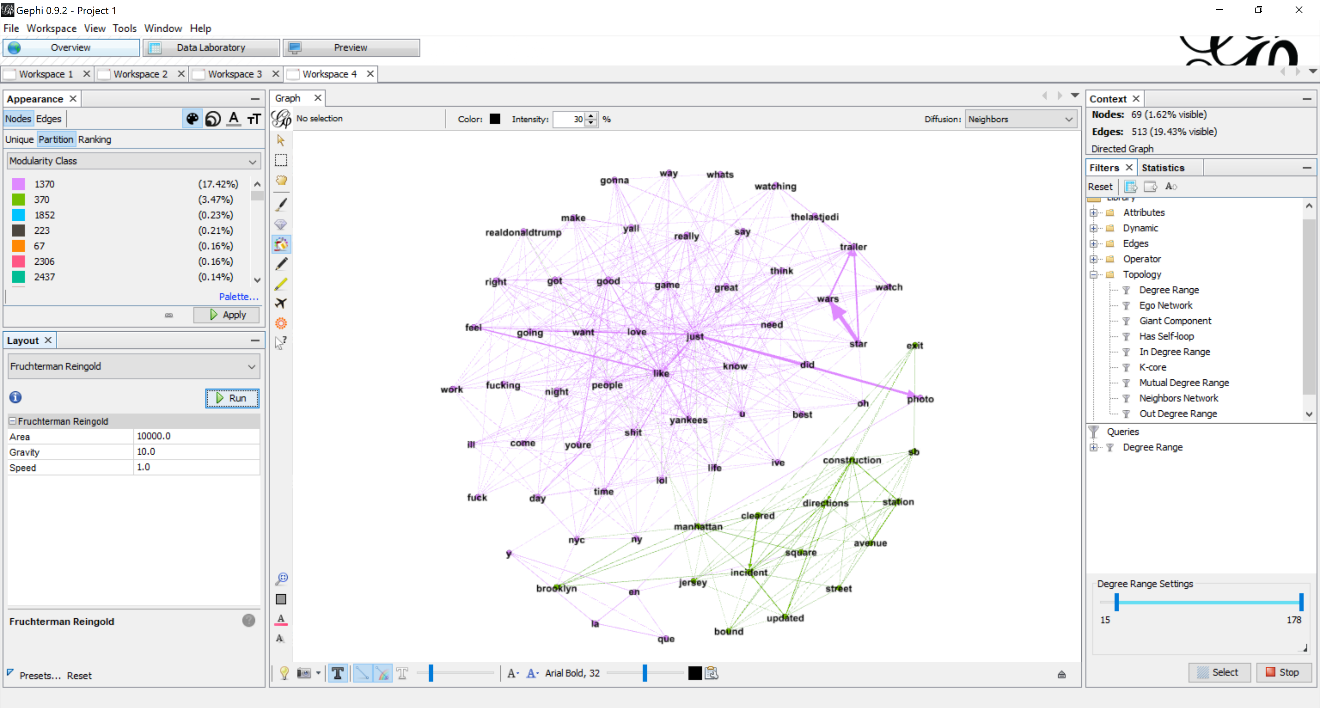


Figure Gephi Network diagram

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

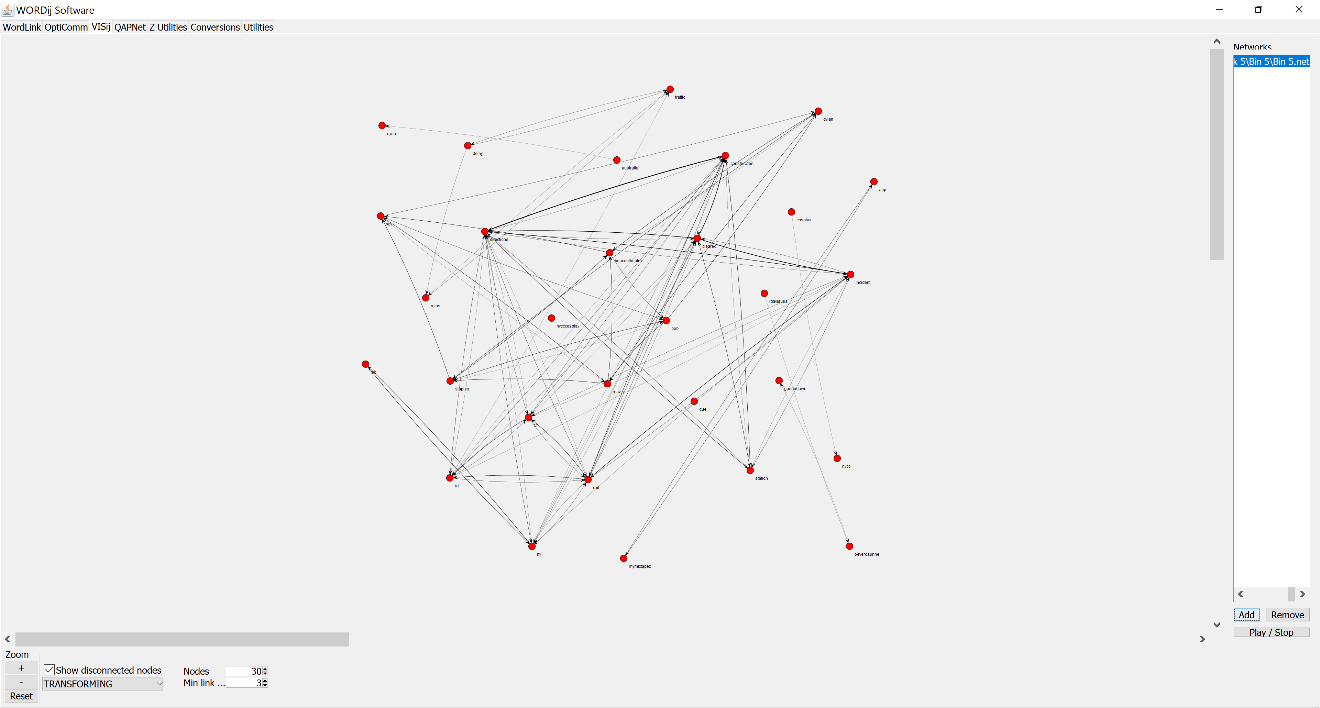


Figure Wordij Network Diagram

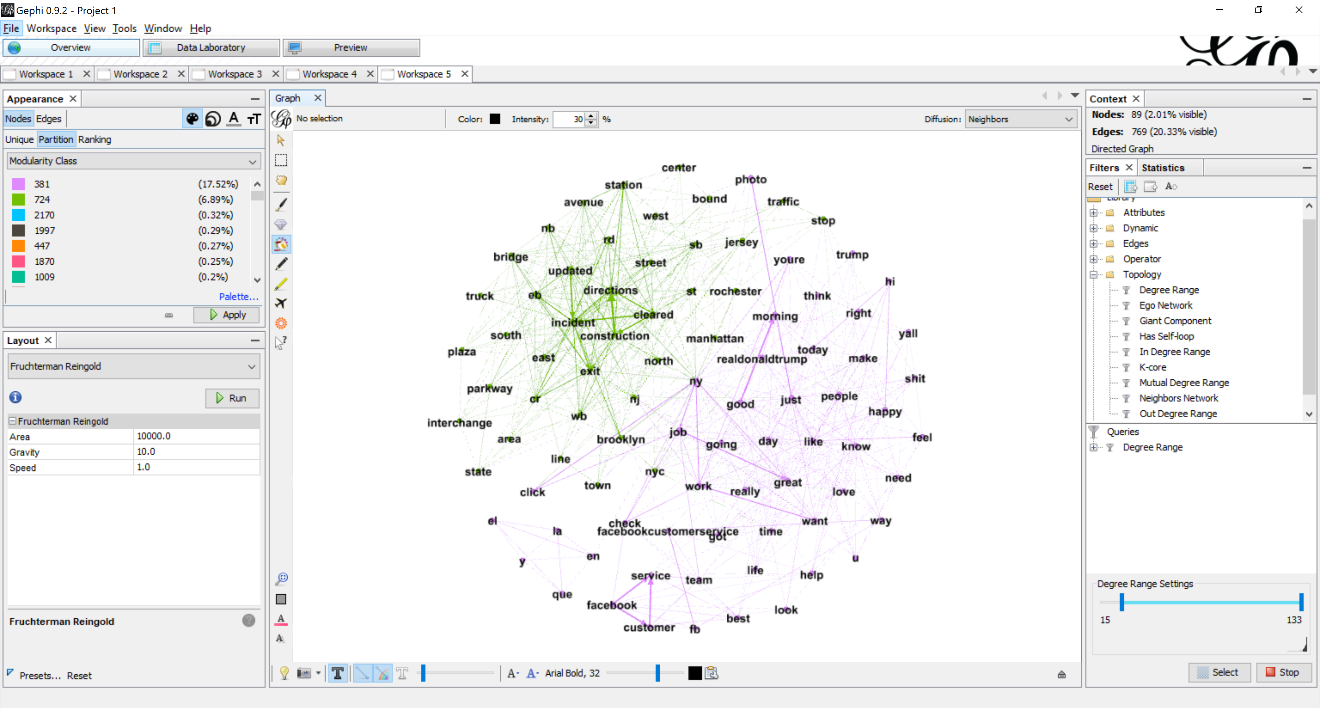


Figure Gephi Network Diagram

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.