STEVENS
INSTITUTE of TECHNOLOGY
THE INNOVATION UNIVERSITY®

Demonetization In India

A Text Mining Approach

EM-623 Data Science and Knowledge Discovery

Vinay Rajesh Gor

CRISP-DM of Demonetization in India using Text Mining

Introduction

On November 8th 2016, *The Honorable Prime Minister of India, Mr. Narendra Modi* announced demonetization of all Rs.5,00 and Rs.1,000 notes accounting 86% of the country's currency in order to combat Black money and corruption going on from years.

Project Goals

The purpose of this project is to analyze the trend of twitter and visualize the time series evolution of virtual conversation post demonetization. This will help them to analyze common issues faced by people, thinking and reactions of opposition leaders, effect on the economy, knowing public decisions etc.

Business Understanding

The objective of this study is to analyze *tweets containing #demonetizationinindia* as a *Data Scientist* for the Government *of India* to determine the general sentiment of tweets and to visualize the time series virtual conversation happened after Demonetization in India.

This study can used by *Government of India* to analyze the trends of twitter which can state the following points:-

a. What are the problems faced by people?
b. Are the people supporting this decision or are they against this decision?
c. What are conversations happening in the country post demonetization?
d. How are the opposition leaders reacting to it?
e. How is it going to affect the economy?
f. Is the decision going to affect the next elections?

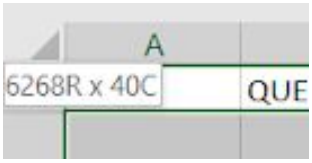
This trend can be helpful for the Government of India to analyze and work upon the trends. For example if the people are facing shortage of money due to current notes being nulled, the Government can implement online & UPI transactions. Another example would be to know whether the people are or not with this decision.

The study will conclude trending topics on Twitter and time series evolution of conversations from tweets.

The data set will be extracted from a *csv* file, will be imported to *txt* file and finally will be analyzed using the combination of *Gephi and Wordij. Also the project will be evaluated using Knime Analytics.*

Data Understanding

To understand the data and its characteristics the data was first explored in it's basic format .csv (comma separated values). Using the *Excel sheet* and visualizing it was explored that the dataset contains *6268 rows* and *40 columns*.



Observations and Variables

Trying to explore more about data we analyze that the data does not contain any missing values but it consists of *tweets* which might have non printable characters. Hence cleaning will be required. The database also contains *40 variables* such as I*D, tweetpic, source, datecreated, entities mentioned, username etc*. which will be eventually removed except tweets.

Data Preparation

From data understanding it can be seen that the tweets which are tweeted may have non printable characters and hence can be cleaned using ***=CLEAN()*** function in Excel. The data also has *40 variables* such as *ID, tweetpic, source, datecreated, entities mentioned, username etc*. hence will be required to remove all the data except tweets which are not much contributing to our dataset. This will be done step by step.

Cleaning the tweet column using Excel

The tweets text contains non printable and erroneous characters which need to be cleaned first which can be done using ***=CLEAN()*** function.

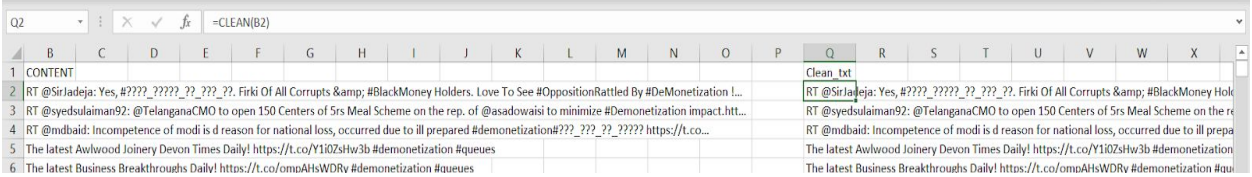


Fig. Using CLEAN() function

Selecting the tweets from .csv file and copy it into a text editor and save it as txt file by dividing into 6 files each one with tweets per day

All the files containing tweets are copied into the text editor file and are saved as txt files. They are named as nov20, nov21, nov22, nov23, nov24, nov25 each containing tweets tweeted per day from November 20,2016 to November 25,2016.

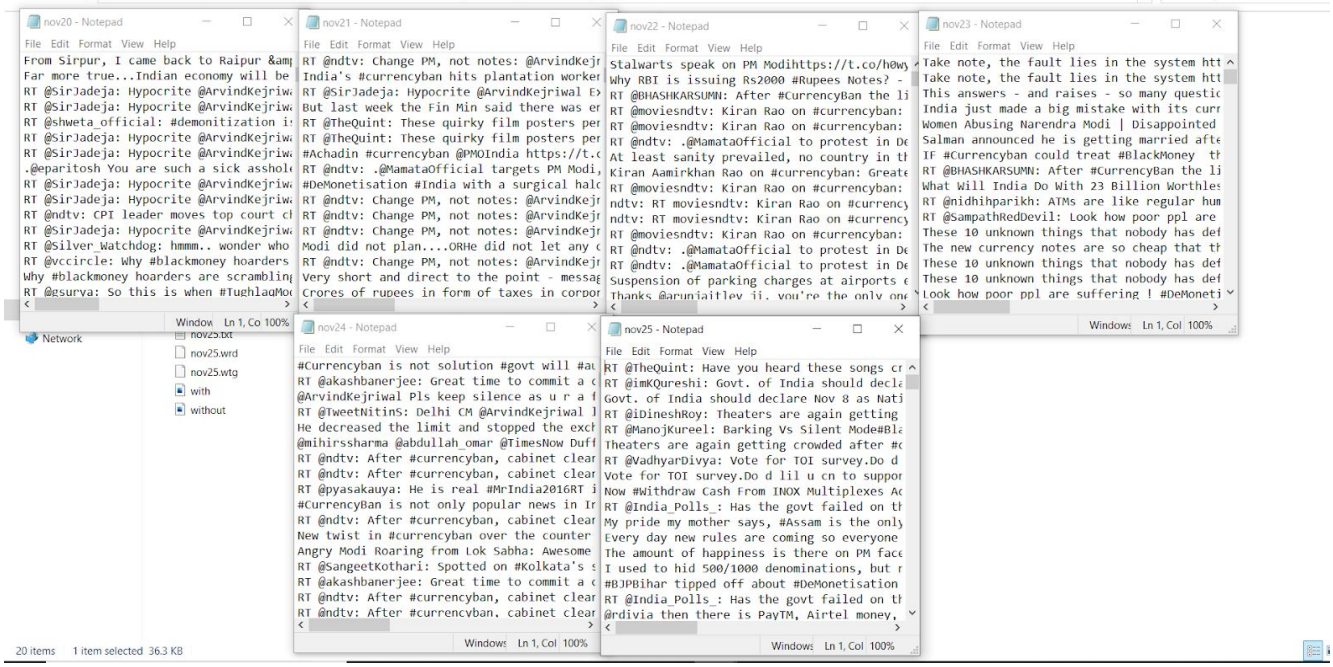
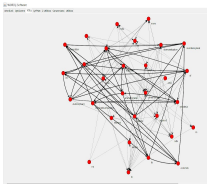
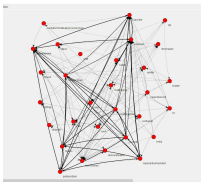
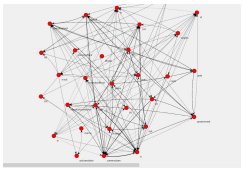
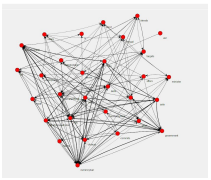
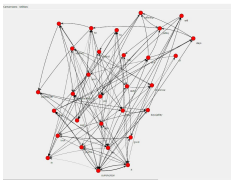
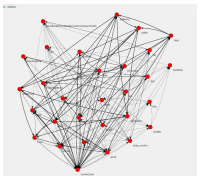
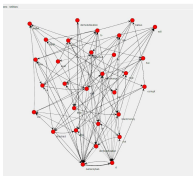
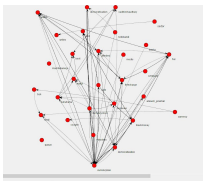
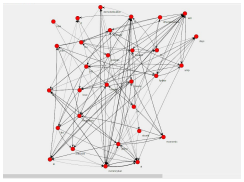
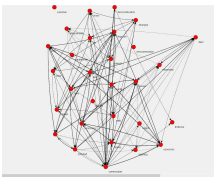
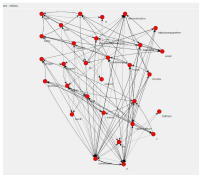
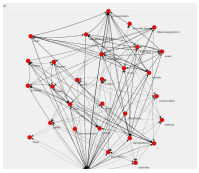


Fig. 6 files each containing tweets per day

Processing the file using Wordij

Each file is imported in Wordij and executed using the default parameters. and then using the drop list menu and using the file droplist.txt stop words are removed. This creates a *.net* file in the same folder.

Day	Before removing stopwords	After removing stopwords
nov20		
nov21		
nov22		
nov23		
nov24		
nov25		

Importing .net in to Gephi and Understanding the Key metrics

Analyzing the network from the structural standpoint

The *.net* which is obtained from *Wordij* is imported in *Gephi* and executed using default parameters to obtain the number of nodes and edges.

Days	Number of Nodes	Number of Edges
nov20	224	980
nov21	337	1137
nov22	310	1056
nov23	303	980
nov24	345	1088
nov25	256	826

Statistics

The Degree Range Filter is applied to all .net files and statistics are observed.
PageRank is a way of measuring the importance.

Average path length is the shortest path for all pairs of network nodes - The shorter the path the more desirable path. Nov21 (shortest) and nov25 (highest).

The average clustering coefficient indicates how nodes are embedded.It also indicates the overall clustering of the network - Node 3 (lowest) which means that it is connected to the least nodes and Node 1 (maximum) which means that it is connected to the maximum nodes.

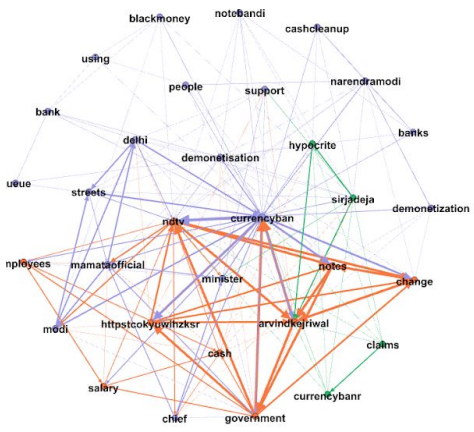
Statistics	nov20	nov21	nov22	nov23	nov24	nov25
Page Rank						
Average Path Length	2.32	2.00	2.18	2.14	2.25	2.41
Average Clustering Coefficient	0.32 	0.41 	0.42 	0.50 	0.42 	0.47 

Modeling

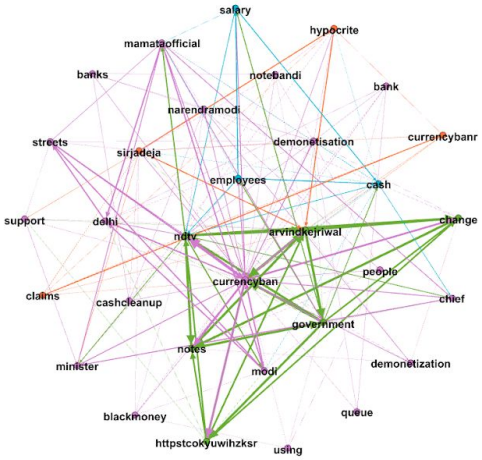
Modularity

Modularity measures the strength of community structure to see if the nodes of the network can be easily grouped into (potentially overlapping) sets of nodes such that each set of nodes is connected internally.

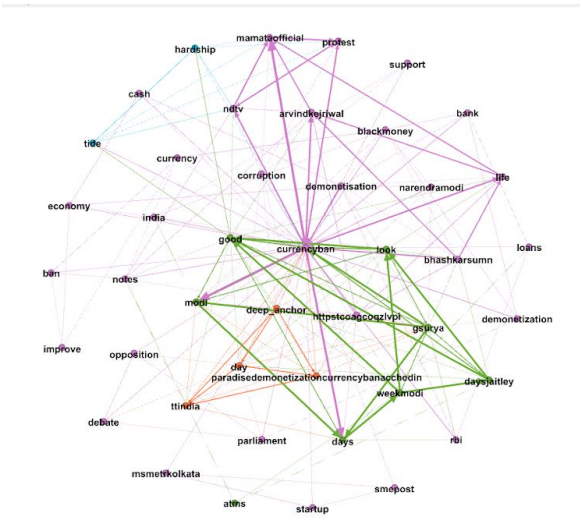
Text	*Modularity*	*No. of Communities*
nov20	*0.196*	*3*
nov21	*0.134*	*4*
nov22	*0.159*	*4*
nov23	*0.197*	*5*
nov24	*0.165*	*5*
nov25	*0.282*	*3*



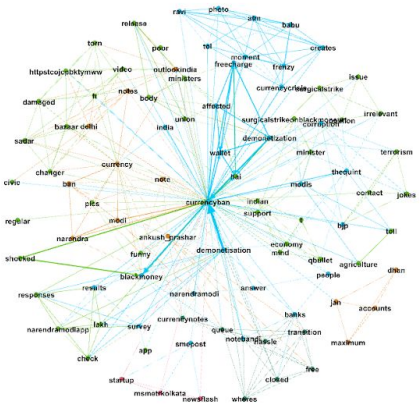
Modularity nov20



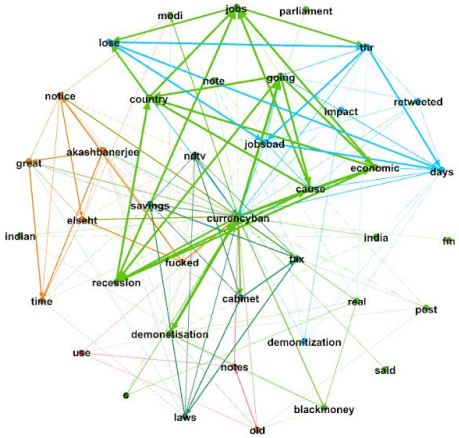
Modularity nov21



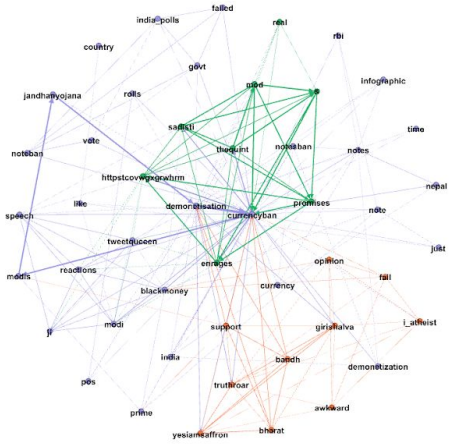
Modularity nov22



Modularity nov23



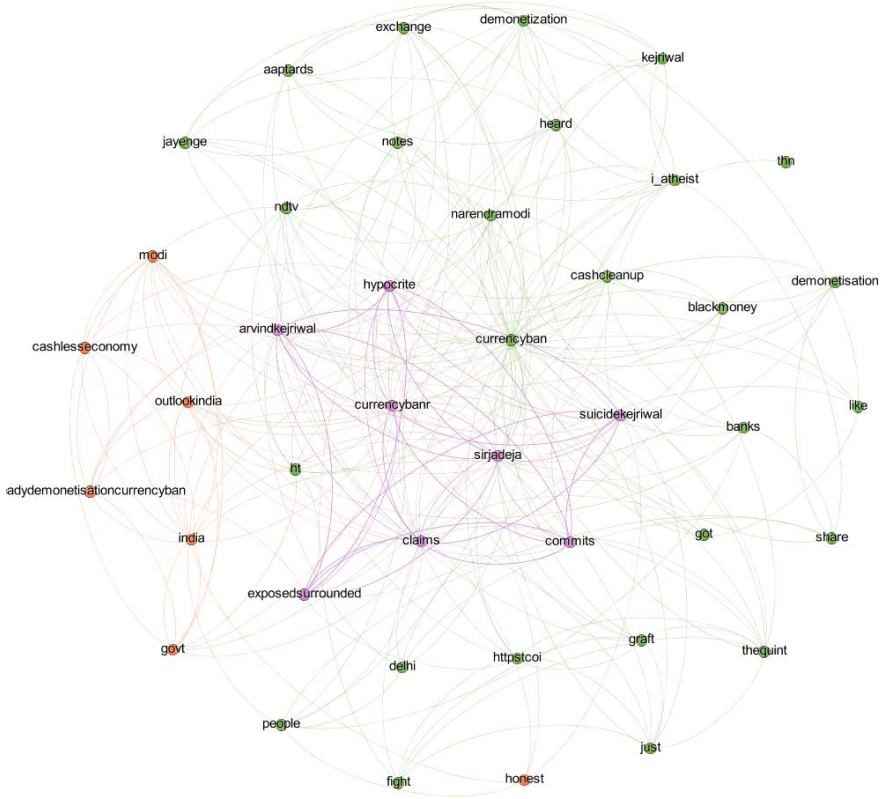
Modularity nov24



Modularity nov25

Using Fruchterman Reingold Layout and Modularity Class Partitioning to analyze the Virtual Conversation.

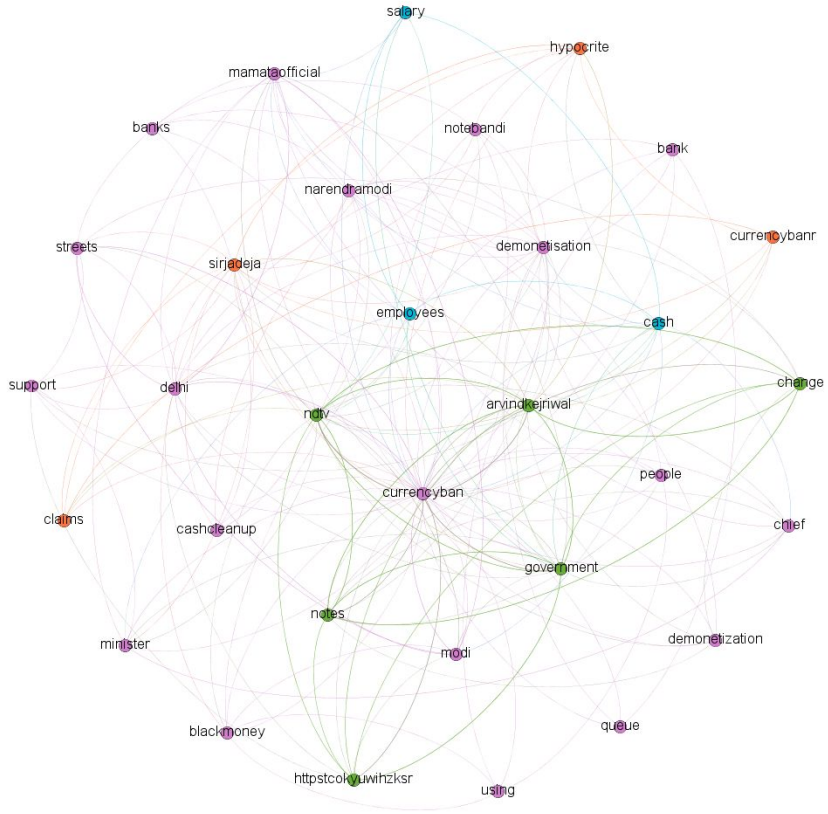
nov20



In this situation there are 3 Virtual Conversation (Topics) which are observed:-

1. *Due the currency ban Arvind Kejriwal which is one of the leaders of the opposition party is abused to be a hypocrite.*
2. *The quint, ndtv, ht all are media firms which are talking about prime minister of India Narendra Modi, banks and currency ban.*
3. *Due to Narendra modi and currency ban India is now turning out to be a cashless economy.*

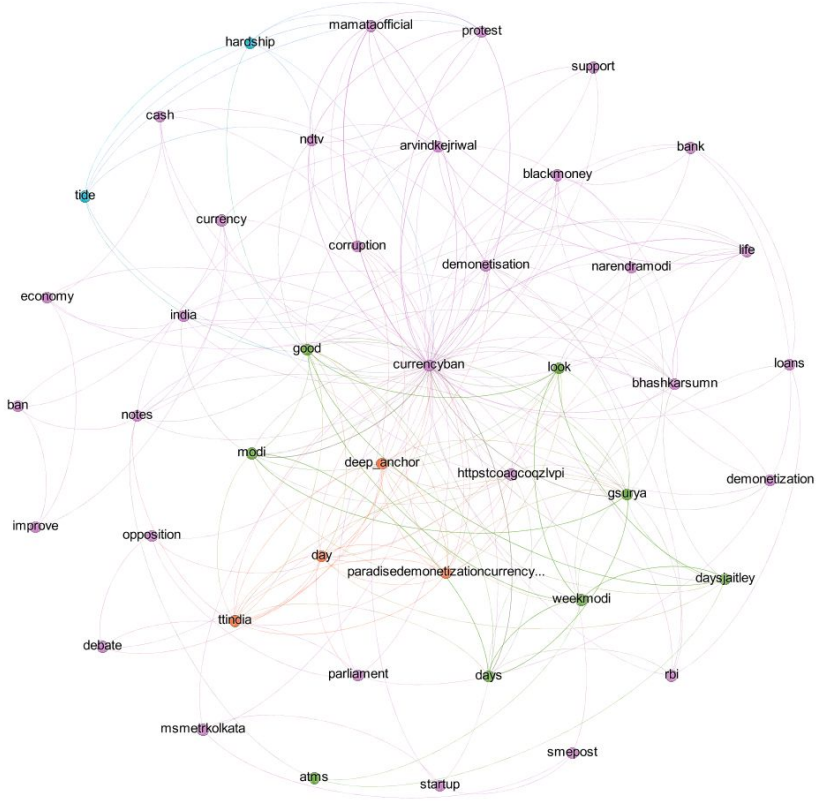
nov21



In this situation there are 3 Virtual Conversation (Topics) which are observed:-

1. *Due to the currency ban Mamta which is one of the leaders of the opposition party is thinking to come on streets to protest.*
2. *Due the currency ban Arvind Kejriwal which is one of the leaders of the opposition party is abused to be a hypocrite*
3. *The quint, ndtv, ht all are media firms which are talking about prime minister of India Narendra Modi, Arvind Kejriwal, banks and currency ban.*

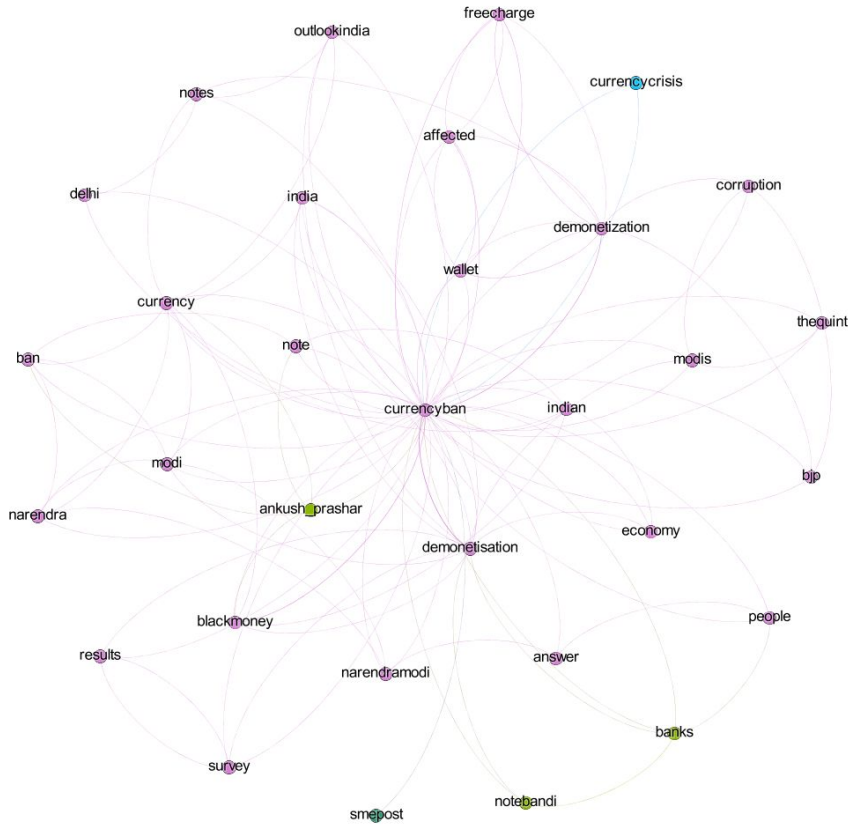
nov22



In this situation there are 3 Virtual Conversation (Topics) which are observed:-

1. *Due to the currency ban Mamta and Arvind Kejriwal which is one of the leaders of the opposition party is thinking to come on streets to protest.*
2. *Due the currency ban Modi and Finance Minister are supported with positive words starting good days “acchedin” will come.*
3. *Currency ban, good days are considered to be a paradise.*

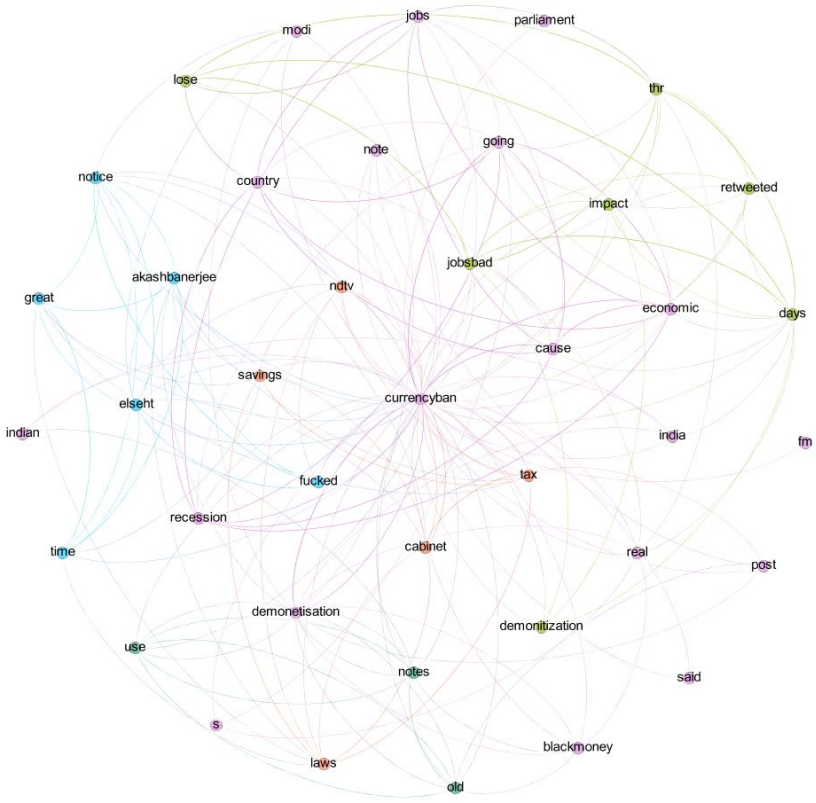
nov23



In this situation there are 2 Virtual Conversation (Topics) which are observed:-

1. *Due to the currency ban Freecharge which is an upi app is affected. People are asking Modi and their party bjp about the black money, corruption.*
2. *Due to the currency ban there is a currency crisis.*

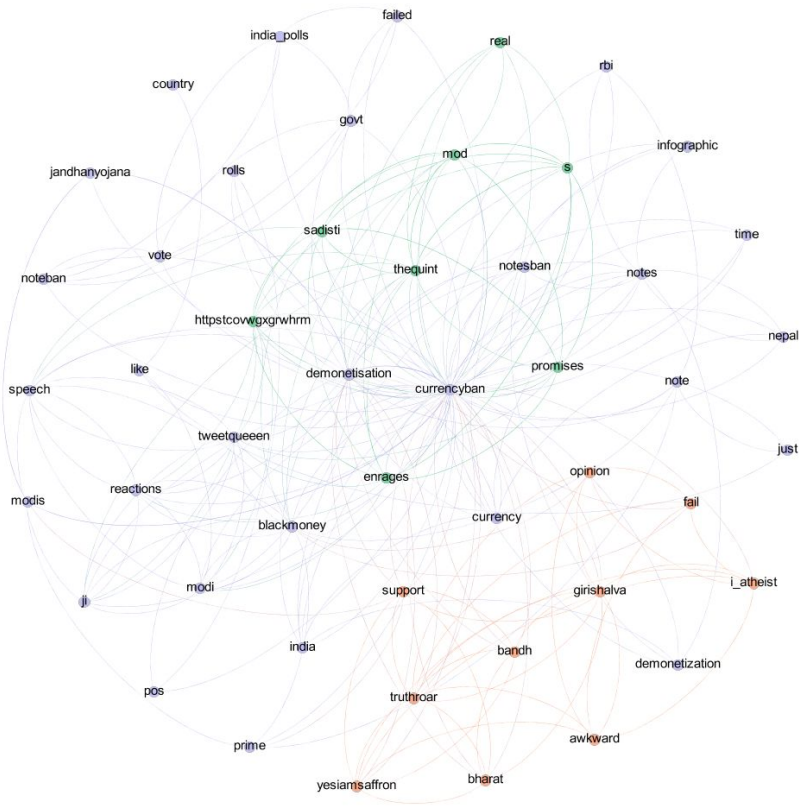
nov24



In this situation there are 4 Virtual Conversation (Topics) which are observed:-

1. *Due to the currency ban the country is going to have economic and job recesion.*
2. *Due to the currency ban the country is going to lose many jobs.*
3. *The topic is about a new tax law which can be soon implemented in the cabinet.*
4. *What is happening to the old notes and black money.*

nov25



In this situation there are 3 Virtual Conversation (Topics) which are observed:-

1. *Due to the currency ban the country the people are enraged and they are waiting for Modi to give a speech.*
2. *There is demand for Lockdown in India (Bharat band).*
3. *There are reactions to Modi and Jandhan Yojna.*

Evolution of Topics.

The conversation evolved from the currency ban which will help India to create a Cashless Economy and due to which there will be Good days. But this can also lead to currency crisis, economic recession and many people might lose their jobs.

The conversation evolved from Opposition leaders such as Arvind kejriwal and Mamta coming on streets to protest and finally can lead to lock down in India due to this protest.

The conversation evolved from currency ban and blackmoney to media firms such as ndtv, ht, quint who are dealing with the news related to Demonetization. The people are waiting to get answers from the Prime minister of India if this ban is helping or not and what are the results.

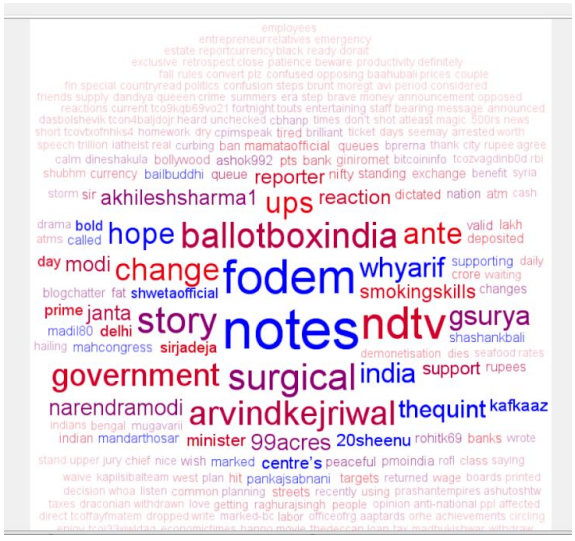
Evaluation

Evaluation Can be incorporated with *Knime Analytics* to see if the same conversations are resulted or not.

Nov20

In Knime Analytics the words which are most linked and observed are aaptards, hypocrite, ndtv which incorporates the same conversation and evolution as in analyzed by Gephi.

Nov21

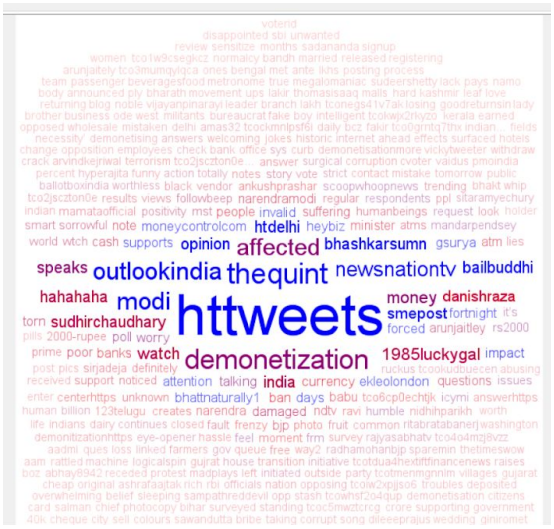


In Knime Analytics the words which are most linked and observed are arvind kejriwal, notes, streets which incorporates the same conversation and evolution as in analyzed by Gephi.

nov22

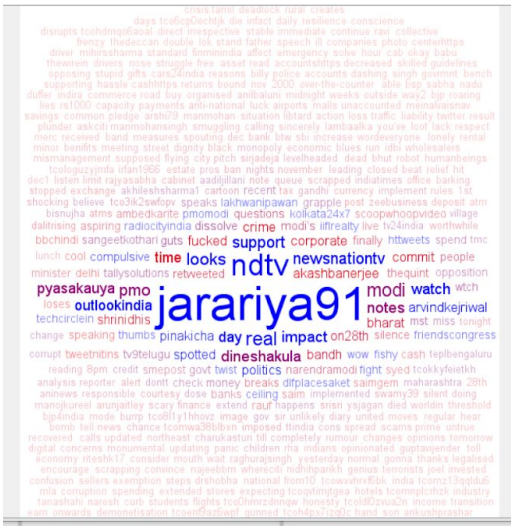
In Knime Analytics the words which are most linked and observed are mamta, modi, paradise, good days, protest which incorporates the same conversation and evolution as in analyzed by Gephi.

nov23



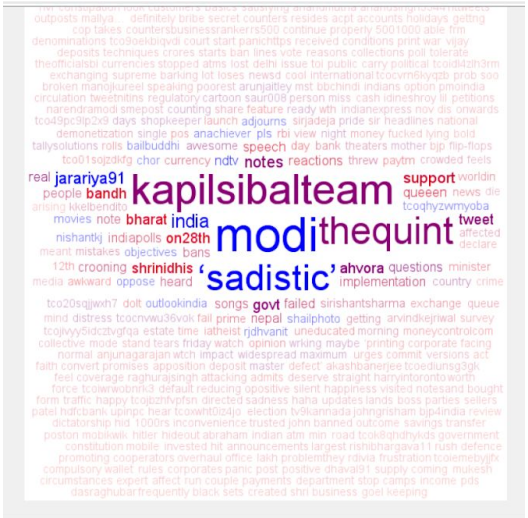
In Knime Analytics the words which are most linked and observed are affected, opinion which incorporates the same conversation and evolution as in analyzed by Gephi.

nov24



In Knime Analytics the words which are most linked and observed are India Lockdown, impact which incorporates the same conversation and evolution as in analyzed by Gephi.

nov25



In Knime Analytics the words which are most linked and observed are India Lockdown, support which incorporates the same conversation and evolution as in analyzed by Gephi.

Deployment

The results created can be used by the Government *of India* to improve the situation created by Demonetization and Currency Ban. This can be helpful for them to understand the common problems which common public are facing, what are the opposition party leaders talking and reacting about, how can this affect economy and what can lead this step in to next election

Conclusions

To conclude, the *tweet* file of *#demonetizationinindia* containing *6268 rows* and *40 columns* is reduced to *1 column containing clean tweet. This file is then imported in txt file and created network analysis in Wordij* removing stop words which is finally imported to Gephi file where using several statistical tools and analytical thinking the twitter trend and virtual conversation is observed. This can be used by Government of India to think and work upon the aspects related to Demonetization in India.