An approach to find change of topics using twitter data

By

Pranay Gouru

Committee Members

Dr. K.M. George(Advisor)

Dr. Johnson Thomas

Dr. Christopher Crick

Department of Computer Science, Oklahoma State University.

Stillwater, OK, USA

ABSTRACT

Social Media is a central domain for dissemination of real time information. To see the information flow change over social networks, we consider text of tweets and derive results, which is computationally intensive process.

Objective of this Projects is to find an efficient method to compute information flow change over a period of time in twitter. For this, we collect and analyze hashtags, user names and retweets.

Twitter API is used to collect the data using the keyword “TRUMP” and the corresponding text is parsed to collect hashtags and retweets.

TABLE OF CONTENTS

I. INTRODUCTION

In twitter lot of discussions happen around the globe on different topics and out of those topics some of them will be discussed over a period of time. Hashtags and text from the tweets represented as information has been used as tools to see flow of information. For example, people prefer the most recent information, and they discover information from various sources like social networks, media sites, blogs or employing their social networks. Existing models may be too constrained to capture the underlying phenomenon.

The flow of information for hashtags is inseparable from retweeting by users. Its network is the growth of total popularity which can be represented by number of users who discuss about the hashtag. Network of tweets is performed two ways to observe the information I) A network of hashtags and users II) A network of retweets and users.

In the First method implementation of hashtags are considered as nodes and users are used as connections in a network. This network is a weighted undirected graph where the connections are users which adds upon through the process of building a network. Hashtags are extracted from the “text” column of the tweets. Tweets are collected over a period of time and chosen the files with high volume to define information. In the second way of modelling the network is between retweets and users in this users are nodes and connections are retweets.

Related Work:

Earlier work has been done using usernames and hashtags relation where usernames are used as the nodes and hashtags for connections. Linear influence model is one such technique used where nodes be infected when they adopt the information. Influence is calculated using a selected group of nodes and results are taken accordingly. But, using this approach gives the relation between nodes without grouping.

In this approach connected components explain the connected information over the network which explains how related topics are within the cluster. From this we can explain how the information flow change can be changed over the time.

Pooling of hashtags into clusters is also mentioned in different research papers using LDA to form clusters and we are using a spectral clustering method to pool the hashtags.

II.DATA COLLECTION

Data is collected from twitter through twitter Application programming interface(API). The data is stored in HDFS and the data is collected through a domain named trump.

Apache flume is an open-source software where a webserver generates log data and this data is collected by an agent in Flume. The channel buffers this data into a sink, which finally pushes it to centralized stores. The process of collecting data is mentioned in the following steps,

- Create a twitter account by signing up.
- Go to https://developer.twitter.com/ and apply for developer account.
- After it is approved then collect consumer secret, consumer token and access token.
- There are three components for a twitter agent, namely source, sink and channel.
- Flume connects to twitter API and receives data in JSON format and stored in the HDFS.
- Add the flume source to the flume class-path.
- Now, create a configuration file for the flume agent the specifying the consumer key, consumer secret, access token and access token secret and keywords, hdfs sink path.

Data is collected in JSON format which contains key value pairs for various fields for which we are using for out analyses and data modelling. Following is the sample JSON file and hashtags and text fields are highlighted.

Fig 1. A sample JSON file:

Keywords used for the trump dataset are,

Trump: Donald Trump

III.IMPLEMENTATION

As mentioned earlier, collected data is in JSON format and the process of implementation is

shown below. A sample JSON format is shown below where text field, hashtag and screen name are highlighted.

{

"is_quote_status": false,

"id_str": "1046654277060624385",

"in_reply_to_user_id": null,

"favorite_count": 0,

"id": 1046654277060624385,

"text": "Check out this open letter to her neighbors on Donald Trump #ActStrong #PinkOutTheVote ",

"place": null,

"lang": "en",

"quote_count": 0,

" "screen_name": "GeekyPoliWonk",

"possibly_sensitive": false,

"coordinates": null,

"truncated": false,

"timestamp_ms": "1538376800204",

"reply_count": 0}

The following is the model we will use to see the information diffusion which is represented as fig 2.

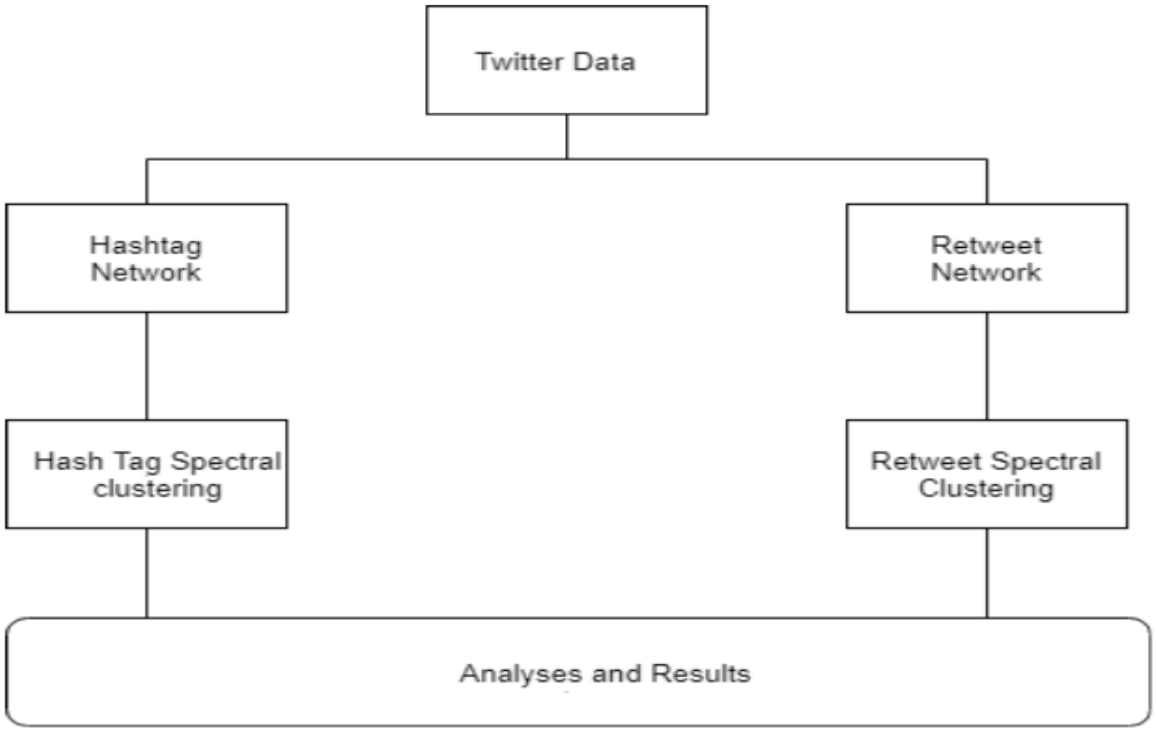

Fig 2. Model of Information diffusion process

Spectral Clustering:

Input: A undirected graph with nodes and edges number k of clusters to construct.
- Let W be its weighted adjacency matrix.
- Compute the unnormalized Laplacian L.
- Compute the first k eigenvectors v1, . . ., vk of L.
- Let V ∈ R n×k be the matrix containing the vectors v1, . . ., vk as columns.
- For i = 1, . . ., n, let yi ∈ R k be the vector corresponding to the i-th row of V.
- Cluster the points (yi)i=1,n in R k with the k-means algorithm into clusters C1, . . . , Ck.

Output: Clusters A1, . . . , Ak with Ai = {j| yj ∈ Ci}.

Laplacian Matrix:

Laplacian matrix of a graph G = (V, E) is an undirected graph without graph loops or multiple edges from one node to another, V is the vertex set, n = |V| and E is the edge set, is a n by n

symmetric matrix with one row and column for each node defined by

$L = D - A$

Where D = diag(d1,…….,dn) is the degree matrix, which is the diagonal matrix from the vertex degrees and A is the adjacency matrix.

Eigen values and Eigen vectors:

$$\mathbf{A}\cdot\mathbf{v}=\lambda\cdot\mathbf{v}$$

In the above equation A is an n-by-n matrix, v is a non-zero n-by-1 vector and λ is a scalar (which may be either real or complex). Any value of λ which this equation has a solution is known as an eigenvalue of the matrix A.

The vector, v, which corresponds to this value is called an eigenvector. The eigenvalue problem can be rewritten as

$$\begin{aligned}\mathbf{A}\cdot\mathbf{v}-\lambda\cdot\mathbf{v}&=0\\ \mathbf{A}\cdot\mathbf{v}-\lambda\cdot\mathbf{I}\cdot\mathbf{v}&=0\\ (\mathbf{A}-\lambda\cdot\mathbf{I})\cdot\mathbf{v}&=0\end{aligned}$$

If **v** is non-zero, this equation will only have a solution if

$$|\mathbf{A}-\lambda\cdot\mathbf{I}|=0$$

This equation is called the characteristic equation of A, and nth order polynomial in λ with n roots. These roots are called the eigenvalues of A. We will only deal with the case of n distinct roots, though they may be repeated. For each eigenvalue there will be an eigenvector for which the eigenvalue equation is true.

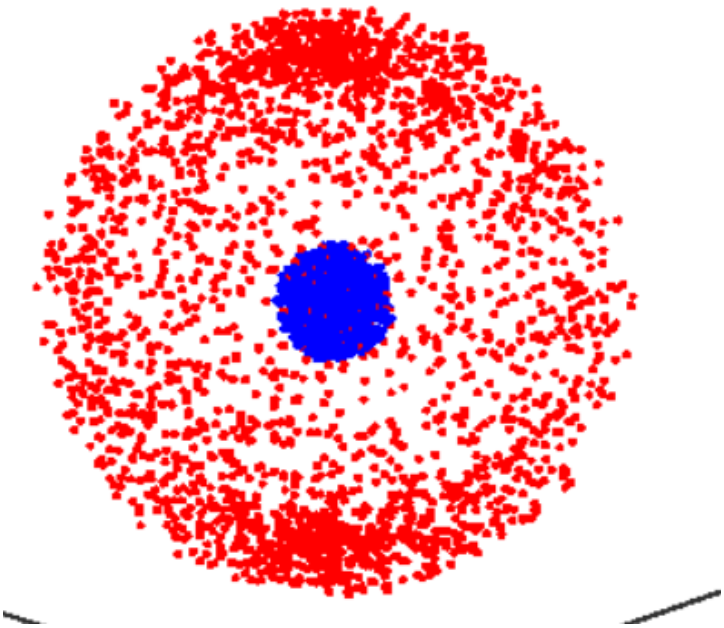

Fig 3. An example for spectral clustering.

Implementation Using Hashtags and Usernames:

- Collected the data using Apache flume.
- Loading the JSON files into a dictionary which contains key value pairs of which we collected several keys like “text” for text,”full text” in “extended tweet” and language selected is “en” whereas in the user tags I have chosen “screen_name” for the usernames.
- Now the “text” key value is cleaned by removing urls, whitespaces etc.
- Text is now used to collect hashtags using split by “#”.
- Built an adjacency matrix for the graph G where nodes are hashtags and connections are users.
- Follow the steps of the spectral clustering and clusters are formed which is further used for analyses and results.

Implementation Using Retweets and Usernames:

- Collected the data using Apache flume.
- Loading the JSON files into a dictionary which contains key value pairs of which we collected several keys like" retweeted_status" for status, "text" for text," full_text" in "extended tweet" and language selected is "en" whereas in the user tags I have chosen "screen_name" for the usernames.
- Now the "text" key value is cleaned by removing urls, whitespaces etc.
- We check for the "retweeted_status" tag to check the retweets and parse the text for data cleaning and preprocessing.
- Built an adjacency matrix for the graph G where nodes are usernames and connections are retweets and the network follows parent child relationship.
- Follow the steps of the spectral clustering and clusters are formed which is further used for analyses and results.

IV. RESULT

#WEEK	#CLUSTERS FOR RETWEETS	#CLUSTERS FOR HASHTAGS	#No of hashtags	#No of tweets
week1	186	204	25561	1514173
week2	158	176	18073	1399554
week3	196	186	19354	1622543
week4	209	221	29687	2825487
week5	189	165	14256	1895635
week6	196	192	21896	2688456
week7	183	201	22555	2898645
week8	201	236	29675	2996354
week9	189	209	26987	2568696
week10	189	203	25524	2364578
week11	206	211	26723	1958765
week12	178	198	23589	2257869
week13	173	173	18564	1955863
week14	169	165	16647	1623458
week15	148	156	21289	1569873
week16	159	153	18536	1789546
week17	168	143	17994	1456832
week18	139	127	15662	1257856
week19	149	145	14256	1254578
week20	185	165	17225	1867596
week21	187	179	19957	1785469
week22	185	185	19758	1569856
week23	207	209	27896	2369875
week24	245	254	25785	2855372
week25	233	245	21456	2512833
week26	208	221	26789	2245127
week27	206	254	29368	2386559
week28	236	236	27895	2457812
week29	176	176	24758	2054478
week30	154	154	26789	1687429
week31	188	188	27289	1668759
week32	163	163	19473	1475586
week33	185	193	24984	1922458
week34	202	202	24421	1336475
week35	176	186	19073	1887546
week36	146	156	17378	1336547

Fig 4. Statistics for the hashtags for twitter data.

The above diagram shows the data collected from June 2018 – Feb 2019 which contains 72 million tweets and contains 807 thousand hashtags and the clusters are represented for each week of the data.

The correlation coefficient for the #hashtags and tweets is 0.79853.

The correlation coefficient for the clusters of retweets and tweets is 0.69953

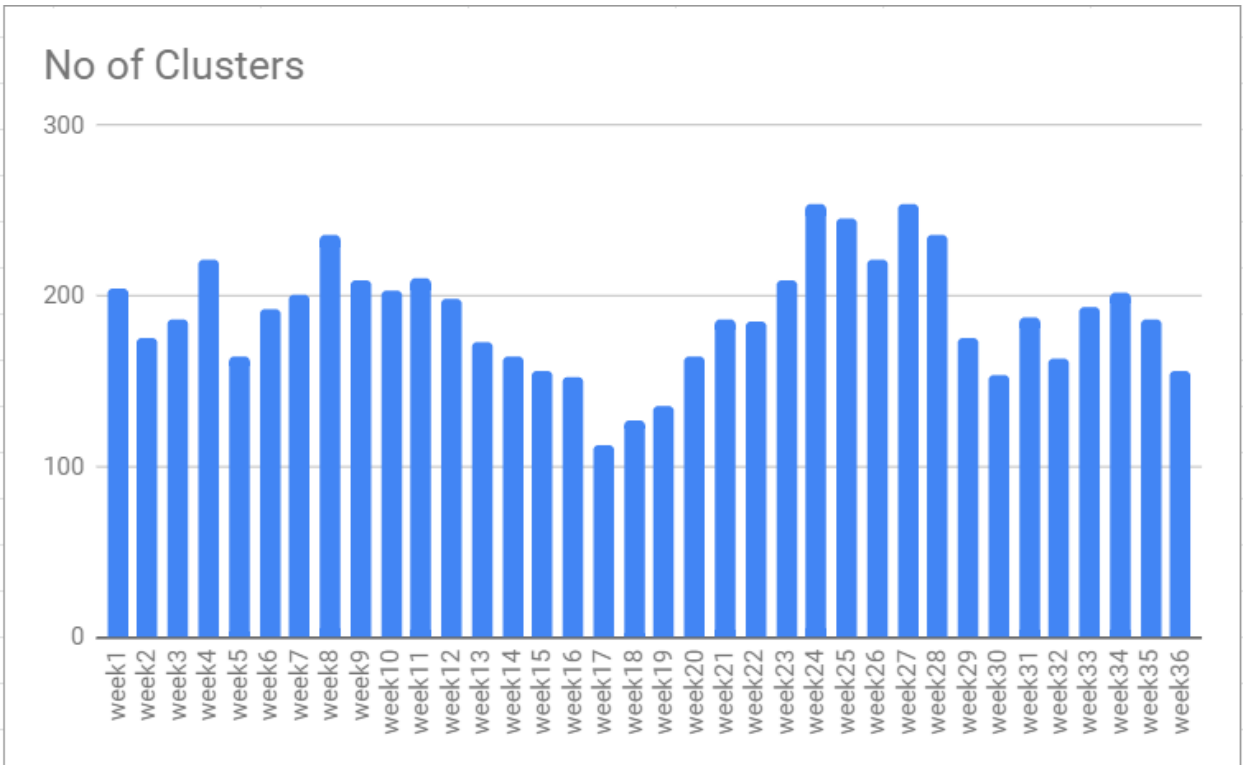

Fig 5. Clusters for hashtags network

The above picture represents the number of clusters formed each week for the Hashtags network.

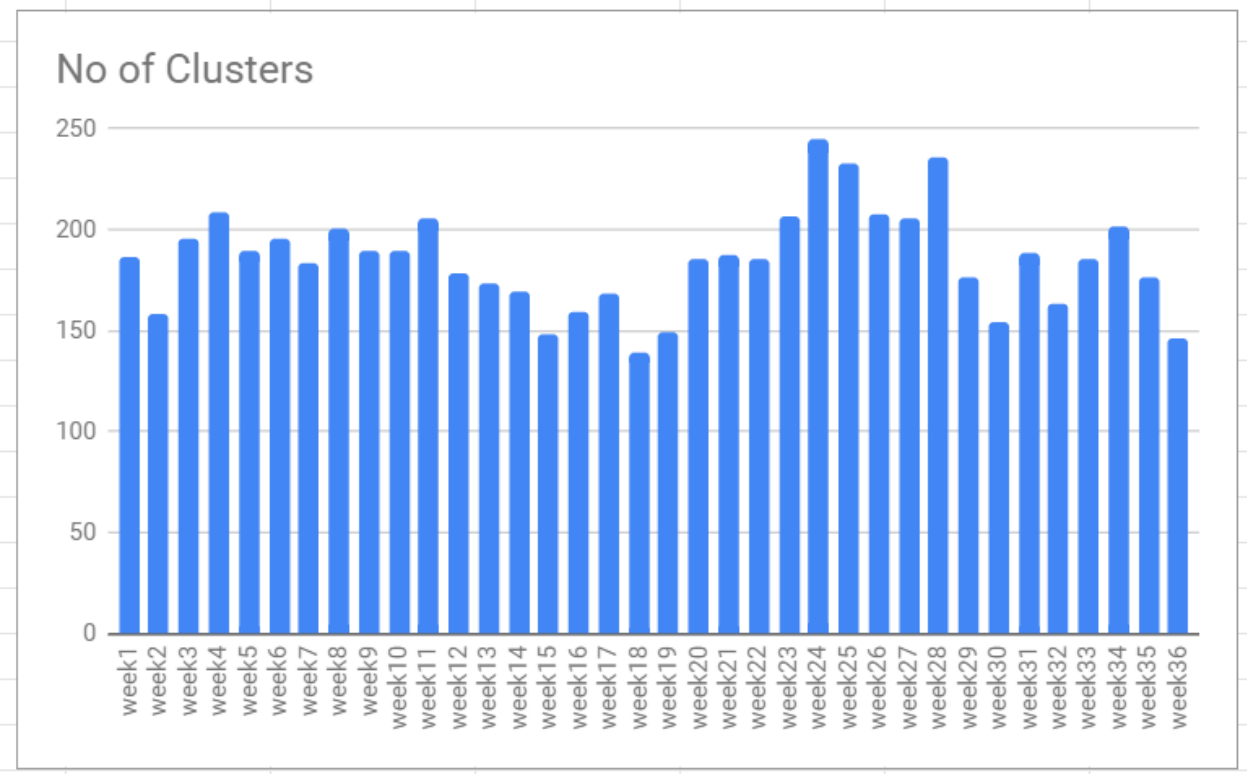

Fig 6. Clusters for Retweet network

The above picture represents the number of clusters formed each week for the retweets network.

The results observed during the week 17 – week 19 where the number of clusters formed are less when compared to the number of clusters during week 24 to week 27.

During week 17 – week 19 we have searched several articles on CNN news and Twitter news feed to see the content and found out topics discussed during this time are on “mid-term elections” and trump going to Tennessee, North Carolina for campaigning in the month of October. Most used hashtag during this period is **#midterm.**

From the week 24 – week 28 which is in December 2018 to the start of Jan 2018.On searching the articles during this period we have observed that most of the articles are on government shutdown and building the wall. But the articles from January the topics are more on **#MAGA**(MAKE AMERICA GREAT AGAIN). **#ShutDown**,**#Trumpwall #buildthewall** are some tags mostly used.

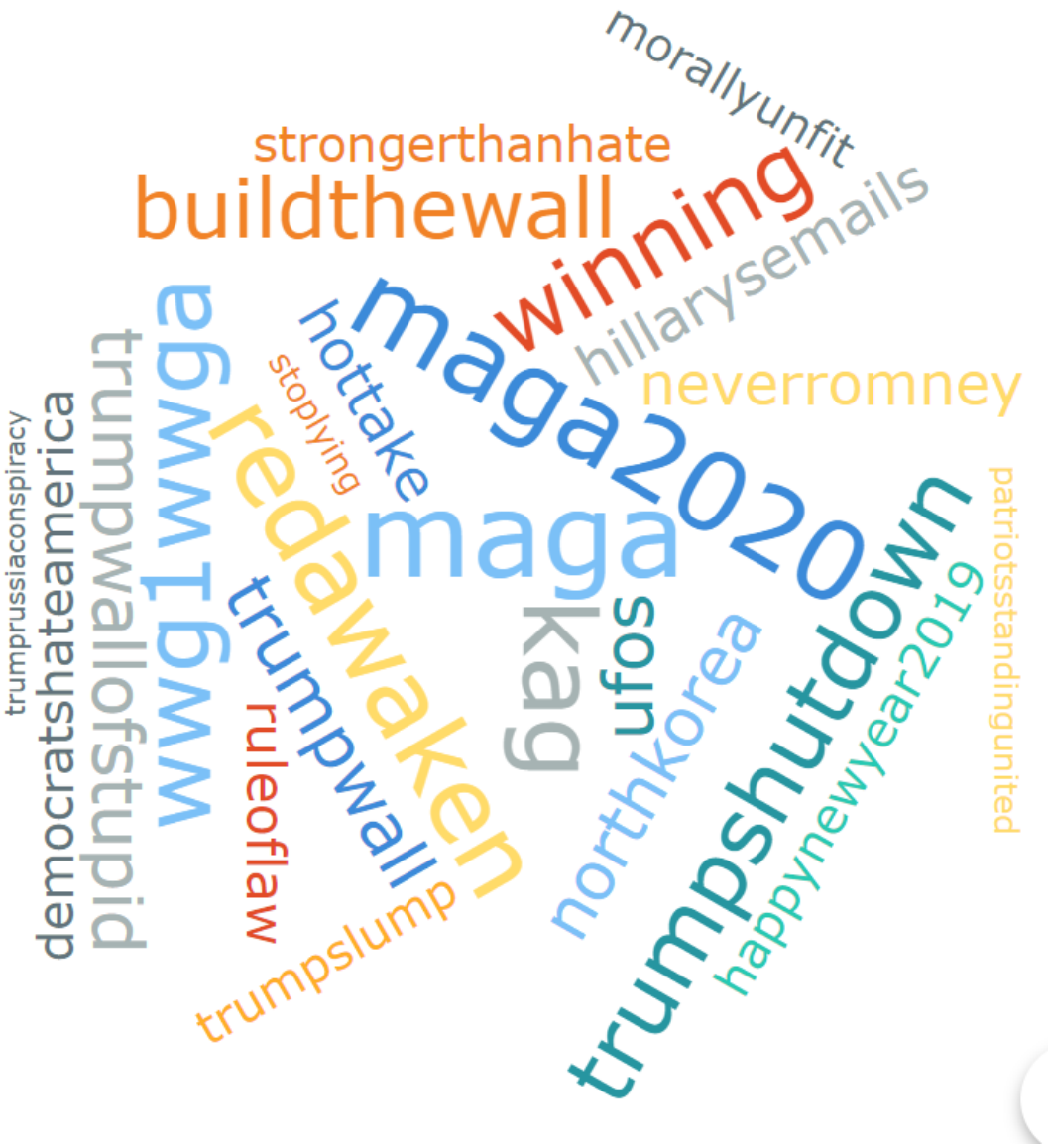

Fig 7. Words cluster.

The above figure is a word cluster of hashtags for frequently used hashtags during the period of December 2018 – Jan 2019.They are classified based on their frequencies.

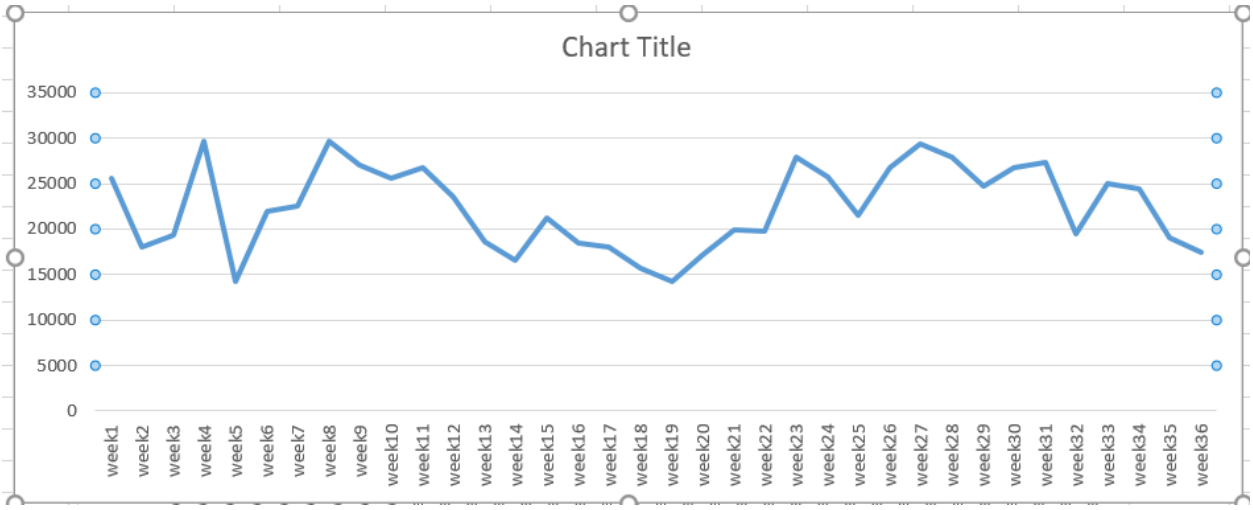

Fig 8. Hashtags vs weeks

The above diagram is the number of hashtags collected over the period from June 2018 to Feb 2019. You also can see a slight variation in the number of hash tags collected during week 17 to week 18.

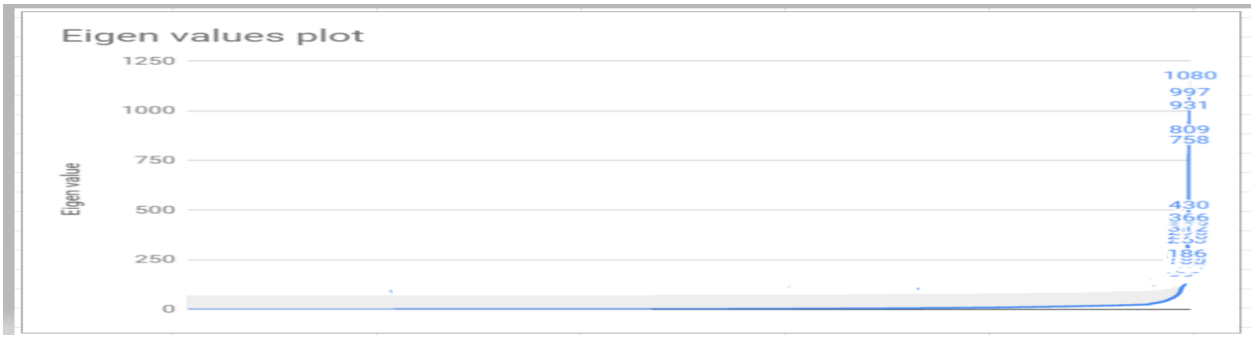

Fig 9. Eigen values plot

The above plot represents the eigen values and is used to determine the number of clusters.

CONCLUSION

From the above observations made from week 17 and week 24 are some takeaways of this project and the similar method can be used to see the information flow change over the period of time for several data sets.

I have found that the information flow change can be seen through various approaches but my way choosing the clustering algorithm will help find the related topics present in the clusters.

We can also find the relation between hashtags and the number of tweets. The choice of networks can be many ways like choosing hashtags and users, users and hashtags, retweets and users, followers count can be used to see the flow of information.

This approach can have handled using the large sets of data but having followers in twitter data would have helped us to build another network to observe the relations of another choice.

REFERENCES

1Jaewon Yang,Jure Leskovec, Modelling information diffusion in implicit networks

2Shushen Fu,Chungjin Hu,Ying Hu,Bo Sun,Wenrui Ying,Peng Shi ,Information Diffusion mechanisms in online social networks

3,De Wang,Aibek Musaev,Calton Pu, Information Diffusion analysis over Rumor dynamics over a social-interaction based model

4,Nitin Sukhija, Mahidhar Tatineni, Nicole Brown, Mark Van Moer, Paul Rodriguez, and Spencer Callicott Topic modelling and visualization for Big data in social sciences.

5. https://www.bbc.com/news/blogs-trending-45040614

6. https://www.cnn.com/2018/12/20/politics/donald-trump-shutdown-border-wall-funding/index.html

7. https://www.cnn.com/politics/live-news/trump-rally-north-carolina-october-2018/index.html

8. https://www.cnn.com/2018/10/05/politics/week-in-review-headlines/index.html