Information, Representation, Processing, and Visualization

By:- Nikhitha Lingutla

Objective:

To process the data, extract information, and discover patterns or knowledge from tweets focusing on Hurricane Harvey to use in disaster management using data mining. Weka tool is used to perform clustering and identify patterns.

Milestone 1- Data Acquisition:

CSV file is downloaded programmatically from the dropbox using wget command.

"wget -q https://www.dropbox.com/s/pytnxphfuhqv9pn/hurriane.csv?dl=0"

After downloading the csv file programmatically the dataset has two columns. First one is a string with column name "TWEET_TEXT" holding tweet messages and the second column is "CREATION_TIME" holding day, date, year, time and time zone information.

TWEET_TEXT	CREATION_TIME
Sheila Jackson Lee Confuses Hurricane Harvey for Sandy Hook on LIVE TV _URL_	Wed Aug 30 13:43:48 +0000 2017
in other words bitch we bout to die _URL_	Wed Aug 30 16:07:28 +0000 2017
US Navy responding to Texas Coast _URL_	Wed Aug 30 22:40:40 +0000 2017

Milestone 2- Data Preprocessing:

1. Since dataset has blank rows in it which needs to be deleted. Used "Go To" tool to select blank rows. When the blanks rows get selected these need to be deleted.
2. Time, time-zone and year are removed from the column "CREATION_TIME".
3. Once the time, time-zone and year are eliminated, duplicates in the data set were removed which narrowed it down from 10000 to 5758 rows.
4. Using R coding I removed all the stopwords and then converted all the words to lowercase.
5. Downloaded the csv file after stopwords removal and converting them to lower case.

week_day	Tweet_Text
30 Aug Wed	sheila jackson lee confuses hurricane harvey sandy hook live
30 Aug Wed	words bitch bout die
30 Aug Wed	navy responding texas coast
31 Aug Thu	fire destroyed family home harvey virgin mary statue survived
31 Aug Thu	important thread list great organizations donations can make real impact wecanhelp
30 Aug Wed	dog rescue this lumberton texas street moms house that brother black shirt
31 Aug Thu	redneck army saves national guard thisisamerica hurricane harvey houston strong
31 Aug Thu	knew good person every since took big mike
30 Aug Wed	hurricane harvey texas first lady makes quiet difference
31 Aug Thu	thank responders private citizens helping people ground devastation left hurricane
31 Aug Thu	join help hurricane harvey relief text harvey donate
29 Aug Tue	wow awesome idea how week tell students excited see
31 Aug Thu	one silver lining hurricane harvey providing plenty inspiration new statues replace tho
31 Aug Thu	can kenosha fill semi hurricane harvey relief support donate
30 Aug Wed	kappa sig donating every every like tweet gets towards hurricane harvey relief efforts

csv file after preprocessing

Milestone 3- Mining tool preparation.

Loading the preprocessed csv file into weka tool. Weka is a mining tool which is responsible for tokenization and cluster assignments. Once the csv file is loaded it shows up that both the attributes "Tweet_Text" and "week_day" are in 'Nominal' format, where the TWEET_TEXT needs to be converted to string format.

No.	1: week_day Nominal	2: Tweet_Text Nominal
1	30 Aug W...	sheila jacks...
2	30 Aug W...	words bitch...
3	30 Aug W...	navy respo...
4	31 Aug Thu	fire destroy

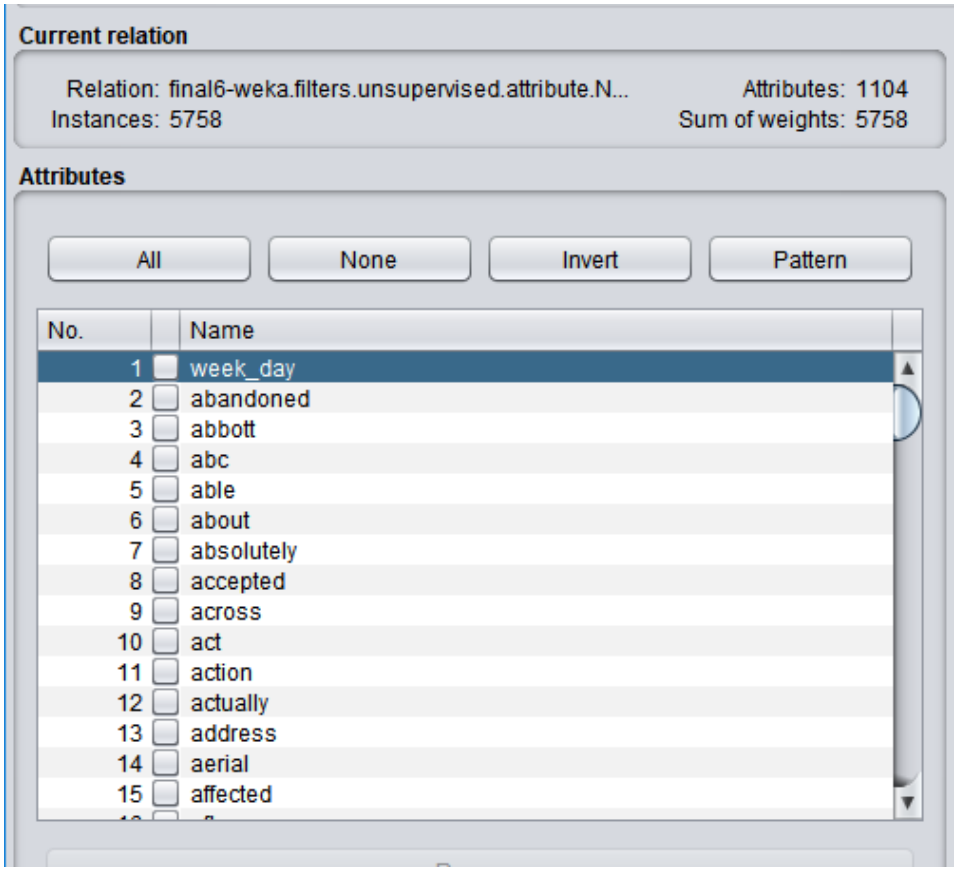
Both columns in Nominal format

To do that we apply a filter tool "NominalToString". Now when the TWEET_TEXT column is converted to string, we apply another filter tool "StringToWordVector" for tokenization purpose.

No.	1: week_day Nominal	2: Tweet_Text String
1	30 Aug W...	sheila jacks...
2	30 Aug W...	words bitch...
3	30 Aug W...	navy respo...

Tweet_Text converted to string

When the words are tokenized from string Tweet_Text by applying a filter tool "String ToWordVector", it looks this way



Strings converted to word with TF and IDF transforms left TRUE

So the dataset after converting to string to word and making them attributes looks this way with TF-IDF values in it.

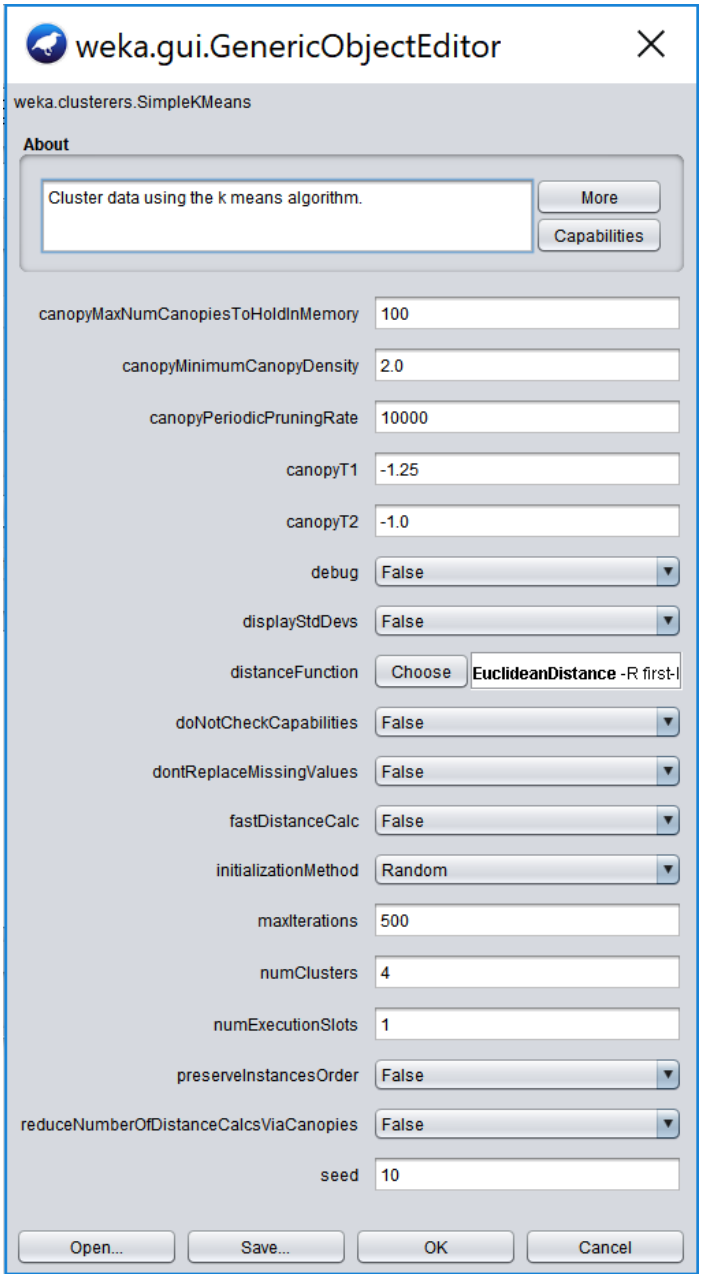
No.	1: week_day	2: abandoned	3: accepted	4: across	5: act	6: action	7: actually	8: address	9: aerial	10: affected
	Nominal	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric	Numeric
1	30 Aug W...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	30 A						0.0	0.0	0.0	0.0
3	30 A						0.0	0.0	0.0	0.0
4	31 Aug Thu	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
5	31 Aug Thu	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
6	30 Aug W...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
7	31 Aug Thu	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
8	31 Aug Thu	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
9	30 Aug W...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
10	31 Aug Thu	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
11	31 Aug Thu	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
12	29 Aug Tue	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
13	31 Aug Thu	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14	31 Aug Thu	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
15	30 Aug W...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
16	30 Aug W...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	1.99246...
17	30 Aug W...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

Sort view: left click = ascending / Shift + left click = descending
Menu: right click (or left+alt)

Word “affected” has TF-IDF value 1.992

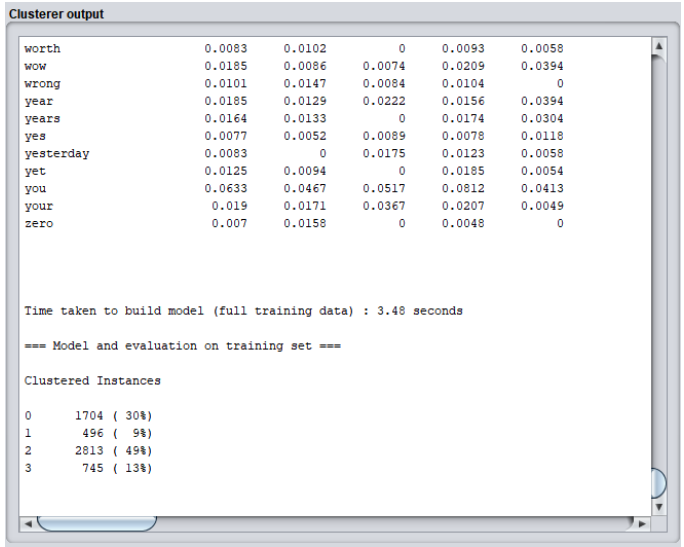
Milestone 4- Clustering Analysis.

Using the file obtained after "StringToWordVector" filter tool to cluster assignment process by moving to 'Cluster' tab. We choose a particular clusterer. I chose "SimpleKMeans" clusterer with 4 clusters and 500 iterations.



SimpleKMeans Properties window

Hit the start button to begin the clustering process. It shows that cluster assignments for the words and give a report of number of words in each cluster.

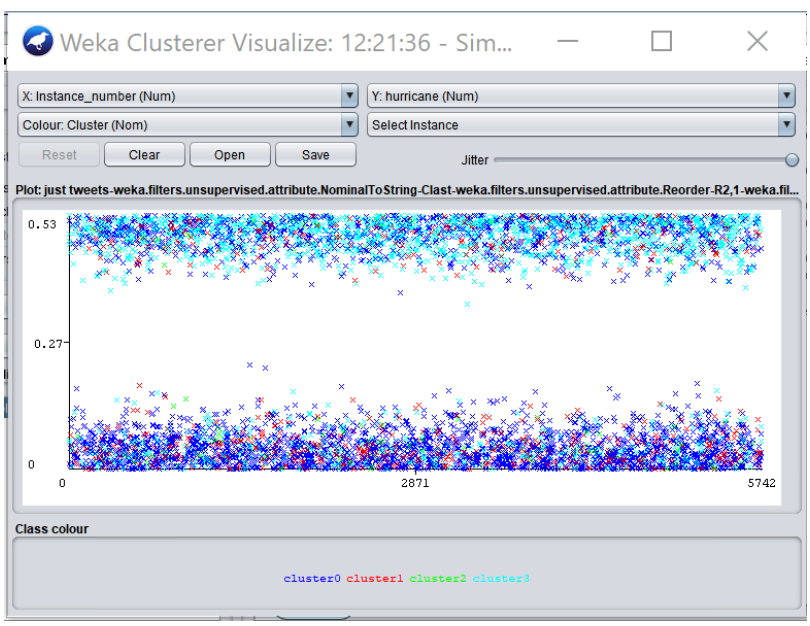


Clusterer output

```
worth                0.0083    0.0102         0    0.0093    0.0058
wow                  0.0185    0.0086    0.0074    0.0209    0.0394
wrong                0.0101    0.0147    0.0084    0.0104         0
year                 0.0185    0.0129    0.0222    0.0156    0.0394
years                0.0164    0.0133         0    0.0174    0.0304
yes                  0.0077    0.0052    0.0089    0.0078    0.0118
yesterday            0.0083         0    0.0175    0.0123    0.0058
yet                  0.0125    0.0094         0    0.0185    0.0054
you                  0.0633    0.0467    0.0517    0.0812    0.0413
your                  0.019    0.0171    0.0367    0.0207    0.0049
zero                  0.007    0.0158         0    0.0048         0



Time taken to build model (full training data) : 3.48 seconds

=== Model and evaluation on training set ===

Clustered Instances

0       1704 ( 30%)
1        496 (  9%)
2       2813 ( 49%)
3        745 ( 13%)
```

Cluster Output

To save the results of the cluster assignments into a csv file we go back to tab "Preprocess" and choose the filter "AddCluster" to get the cluster assignments for each document/message. To save the results once the cluster assignments are done we click on the save button on top right corner and save it in csv format. Visualizations of the clusters from weka looks something like this. Where it looks like word hurricane is frequent in 1st cluster and 4th cluster.

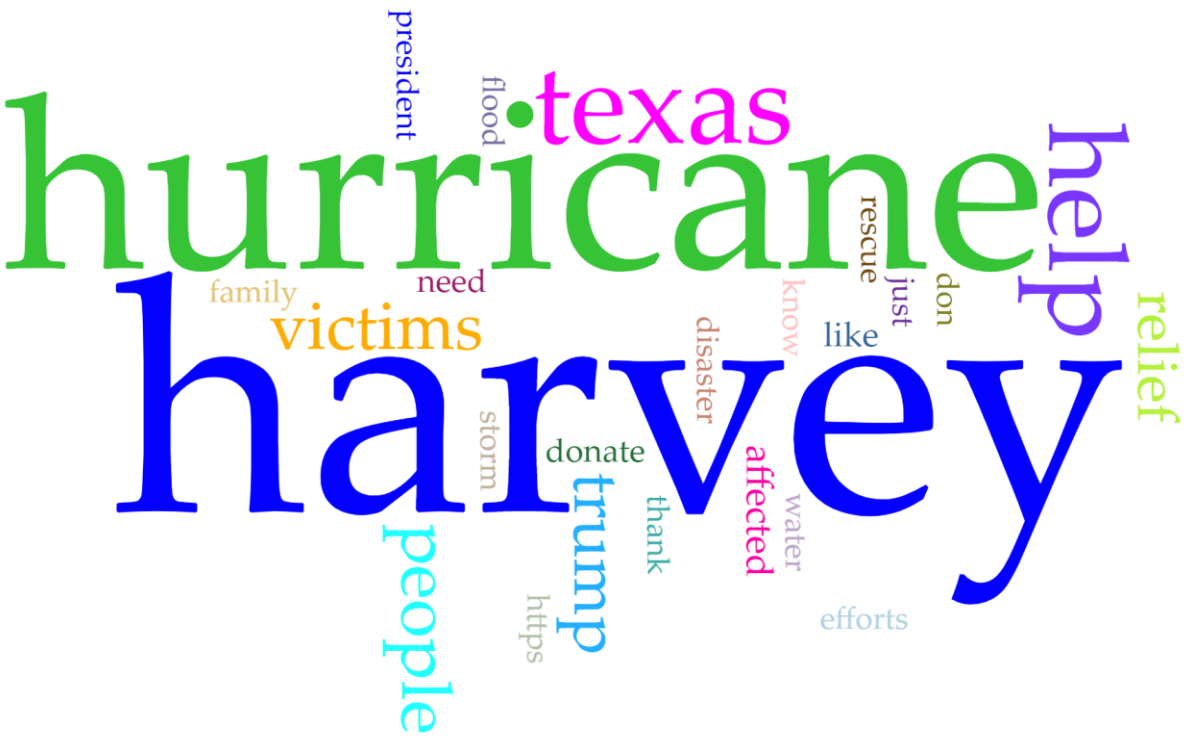


Spread of the word hurricane on different clusters

Milestone 5: Visualization.

Using the outputs of clustering process I made a word cloud for each clusters using voyant tools.

Cluster 1:

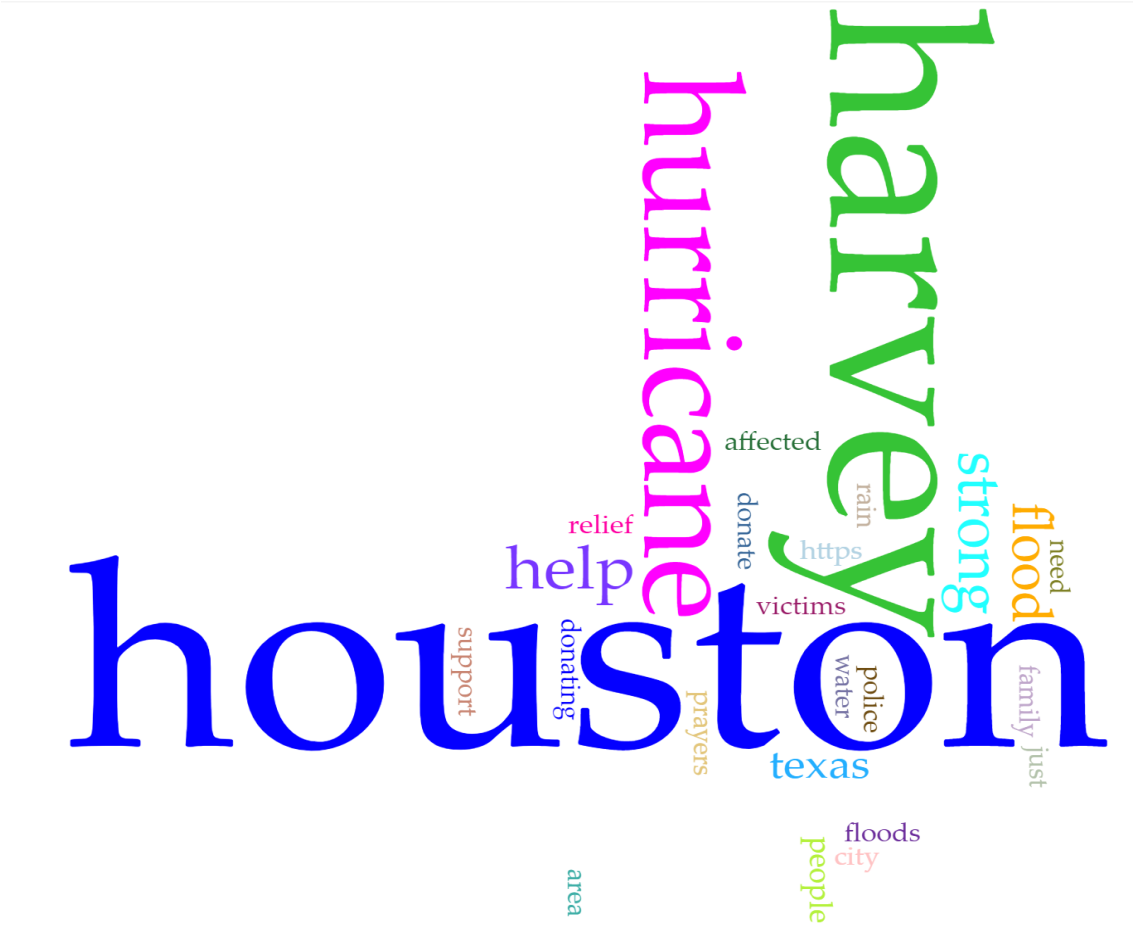


Where in cluster 1 words “harvey” and “hurricane” seems to be most frequently used.

Word Count:

1. hurricane – 1683
2. harvey – 1042
3. texas – 339
4. help – 326
5. people – 199
6. victims –192
7. relief – 186
8. trump –185
9. affected – 128
10. like – 123

Cluster 2:

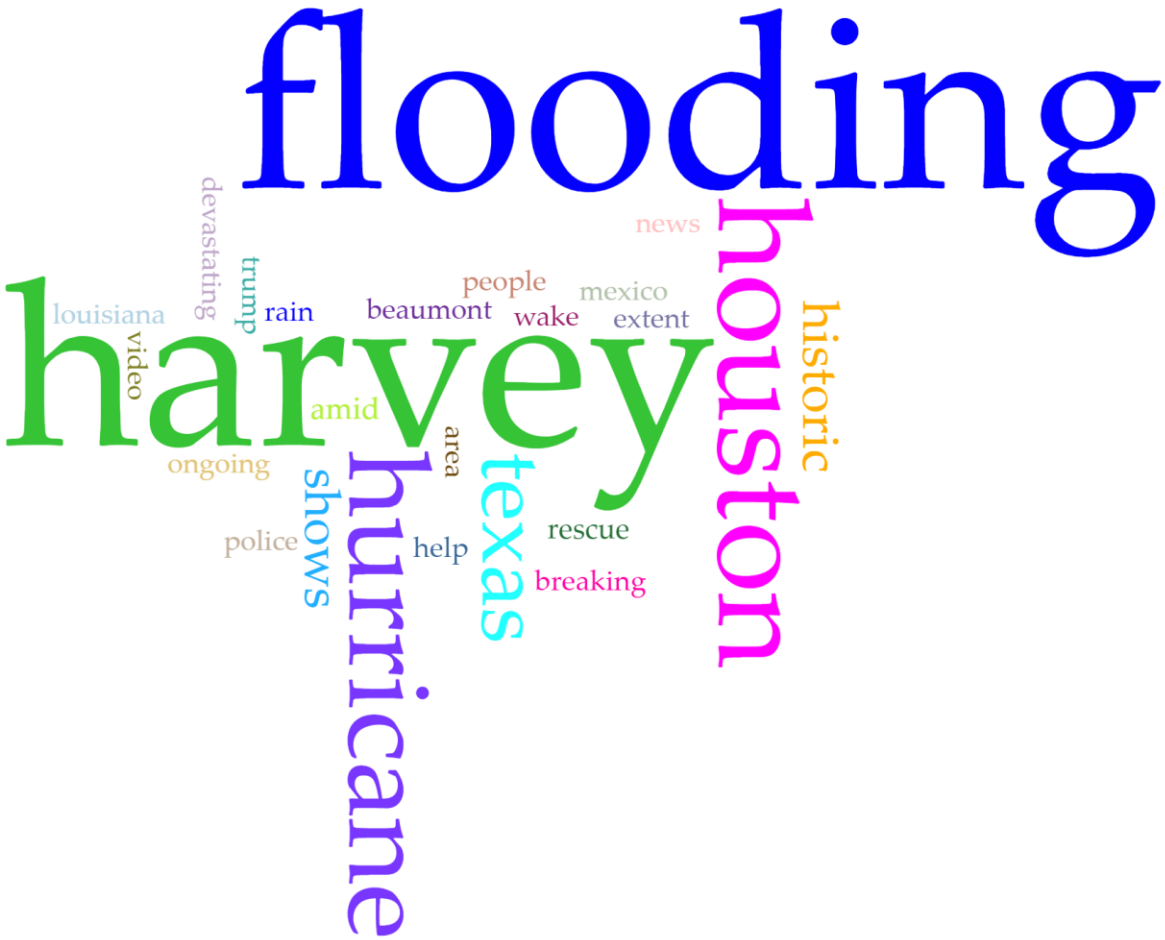


Where in cluster 2 words “houston”, “harvey” and “hurricane” seems to be most frequently used.

Word Count:

1. houston – 750
2. harvey – 489
3. hurricane – 238
4. help – 88
5. strong – 85
6. flood – 78
7. texas – 59
8. relief – 45
9. people – 44
10. donate – 35

Cluster 3:

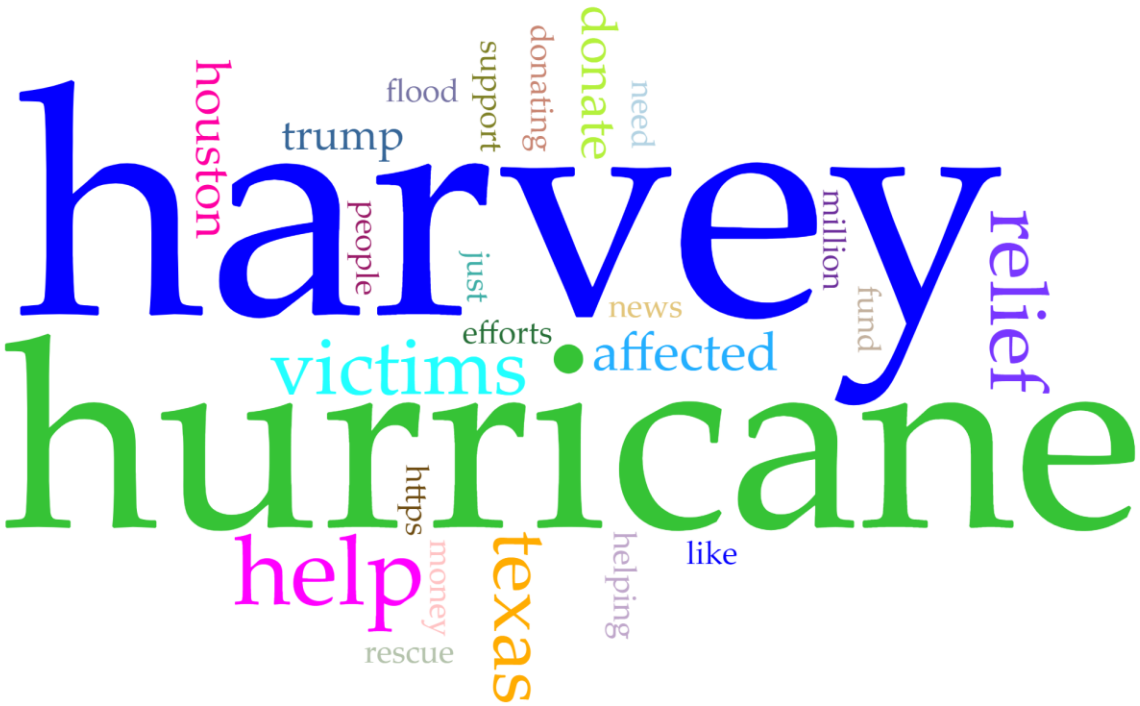


Where in cluster 3 words “flooding”, “harvey” and “houston” seems to be most frequently used.

Word Count:

1. flooding – 106
2. harvey – 88
3. houston – 35
4. hurricane – 31
5. texas – 21
6. historic – 14
7. shows – 12
8. amid – 9
9. breaking – 9
10. help – 9

Cluster 4:



Where in cluster 4 words “harvey” and “hurricane” seems to be most frequently used.

Word Count:

1. harvey – 1881
2. hurricane – 1414
3. help – 302
4. relief – 274
5. victims – 241
6. texas – 229
7. affected – 164
8. donate – 150
9. houston – 146
10. trump – 132

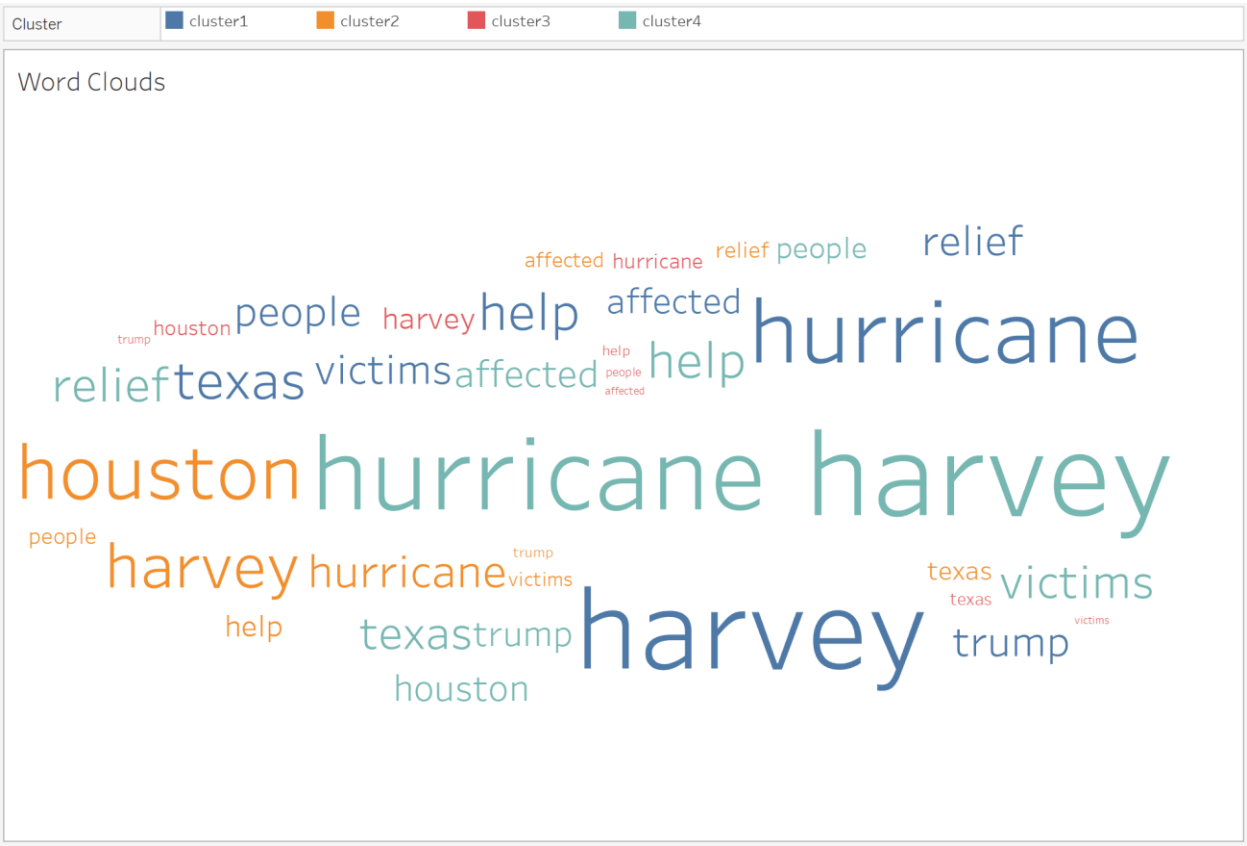
Data: Collection of strings in column one and time and date in column two.

Information: Column one contains the information of tweet texts and its corresponding column has information about when that particular tweet was published.

Knowledge: Most of the tweets were about the Hurricane Harvey which occurred at Houston, Texas.

Wisdom: Many were concerned about the Hurricane Harvey occurred and raising relief funds, collecting donations, asking for help performing rescue operations.

All together along with cluster assignments tableau gives the word cloud with which shows that the words “hurricane” and “harvey” are frequently repeated in clusters 1 and 3.



Conclusion: It very well may be seen that most tweets produced amid the times of 29, 30 and 31 of August, 2017 focused on the victims affected by Hurricane Harvey around Houston, Texas and the endeavors made to give donations for their help.

References:

1. Hurriane.csv. (n.d.). Retrieved from https://www.dropbox.com/s/pytnxphfuhqv9pn/hurriane.csv?dl=0

2. Shams, R. (2013, November 18th). Weka Tutorial 31: Document Classification 1 (Application).

Retrieved from: https://www.youtube.com/watch?v=jSZ9jQy1sfE

3. Prashant Bhowmik (2016, April 8th). Weka Tutorial Unsupervised Learning (Simple K-Means Clustering)
Retrieved from: https://www.youtube.com/watch?v=TtBgfXmIDHQ

4. Voyant Tools to obtain word frequency and word clouds – https://voyant-tools.org/

5. Tableau to obtain word cloud from different clusters in one single image