7th Dec'2018

# Disaster Management

## INFORMATION, REPRESENTATION, PROCESSING & VISUALIZATION

Akshay Bitra
GEORGE MASON UNIVERSITY
G01087950

***Goal***: Hands-on experience to process data, to extract information, and discover patterns or knowledge using data mining method.

## MILESTONE 1: Data Acquisition

| MESSAGE | CREATED_AT |
|---|---|
| @Zuora wants to help @Network4Good with Hurricane Relief. Text SANDY to 80888 & donate $10 to @redcross @AmeriCares & @SalvationArmyUS #help | 2012-10-30 22:15:41 |

i. MESSAGE: message content from social media

- String

ii. DATETIME: date and time of message arrival

- Date-Time

iii. LATITUDE

- Numeric (or 'Null' if no such information is unavailable in tweet)

iv. LONGITUDE

- Numeric (or 'Null' if no such information is unavailable in a tweet)

-The data was imported into R and figured out that it is not clean dataset i.e. it consists of NA's/Missing values.

| | TWEET_TEXT | CREATION_TIME |
|---|---|---|
| 2 | | |
| 3 | Sheila Jackson Lee Confuses Hurricane Harvey f | Wed Aug 30 13:43:48 +0000 2017 |
| 4 | | |
| 5 | in other words bitch we bout to die _URL_ | Wed Aug 30 16:07:28 +0000 2017 |
| 6 | | |
| 7 | US Navy responding to Texas Coast _URL_ | Wed Aug 30 22:40:40 +0000 2017 |
| 8 | | |
| 9 | Fire destroyed a family s home during Harvey b | Thu Aug 31 22:52:04 +0000 2017 |
| 10 | | |
| 11 | IMPORTANT THREAD A list of great organizatio | Thu Aug 31 15:13:34 +0000 2017 |

**MILESTONE 2: Data Preprocessing**

I have used the Excel to separate the CREATION_TIME to Date, Time and Year columns respectively and to also delete the alternate empty rows.

| | TWEET_TEXT | DATE | TIME | YEAR |
|---|---|---|---|---|
| 2 | Sheila Jackson Lee Confuses Hurricane Harvey fc | WedAug30 | 13:43:48 | 2017 |
| 3 | in other words bitch we bout to die _URL_ | WedAug30 | 16:07:28 | 2017 |
| 4 | US Navy responding to Texas Coast _URL_ | WedAug30 | 22:40:40 | 2017 |
| 5 | Fire destroyed a family s home during Harvey bu | ThuAug31 | 22:52:04 | 2017 |
| 6 | IMPORTANT THREAD A list of great organization | ThuAug31 | 15:13:34 | 2017 |

I cleaned the TWEET_TEXT column, removed the unnecessary characters. For example, the _URL_, two letter words etc.

```
#Removing unnecessary characters from Message Column
#------------------------------
data$TWEET_TEXT <- gsub("@", " " , data$TWEET_TEXT)
data$TWEET_TEXT <- gsub("#", " " , data$TWEET_TEXT)
data$TWEET_TEXT <- gsub("(s?)(f|ht)tp(s?)://\\S+\\b", "", data$TWEET_TEXT)
data$TWEET_TEXT <- gsub("[[:punct:]]", " " , data$TWEET_TEXT)
data$TWEET_TEXT <- gsub("[0-9]", " " , data$TWEET_TEXT)
data$TWEET_TEXT <- gsub("\\ can ", " " , data$TWEET_TEXT)
data$TWEET_TEXT <- gsub("\\ to ", " " , data$TWEET_TEXT)
data$TWEET_TEXT <- gsub("\\ is ", " " , data$TWEET_TEXT)
data$TWEET_TEXT <- gsub("\\ the ", " " , data$TWEET_TEXT)
data$TWEET_TEXT <- gsub("\\ URL ", " " , data$TWEET_TEXT)
data$TWEET_TEXT <- gsub("\\ did ", " " , data$TWEET_TEXT)
data$TWEET_TEXT <- gsub("\\You ", " " , data$TWEET_TEXT)
data$TWEET_TEXT <- gsub("\\Your ", " " , data$TWEET_TEXT)
data$TWEET_TEXT <- gsub("\\The ", " " , data$TWEET_TEXT)
data$TWEET_TEXT <- gsub("\\ _ ", " " , data$TWEET_TEXT)
data$TWEET_TEXT <- tolower(data$TWEET_TEXT)

library(NLP)
library(stopwords)
stopwords_regex = paste(stopwords('en'), collapse = '\\b|\\b')
stopwords_regex = paste0('\\b', stopwords_regex, '\\b')
data$TWEET_TEXT = stringr::str_replace_all(data$TWEET_TEXT, stopwords_regex, '')
data$TWEET_TEXT<- gsub(" *\\b[[:alpha:]]{1,2}\\b *", " ", data$TWEET_TEXT)
```

The data looks like this after the previous step:

| | | TWEET_TEXT | DATE | TIME | YEAR |
|---|---|---|---|---|---|
| 2 | 1 | sheila jackson lee confuses hurricane harve | WedAug30 | 13:43:48 | 2017 |
| 3 | 2 | words bitch bout die | WedAug30 | 16:07:28 | 2017 |
| 4 | 3 | navy responding texas coast | WedAug30 | 22:40:40 | 2017 |
| 5 | 4 | fire destroyed family home harvey virgin | ThuAug31 | 22:52:04 | 2017 |
| 6 | 5 | important thread list great organizations | ThuAug31 | 15:13:34 | 2017 |

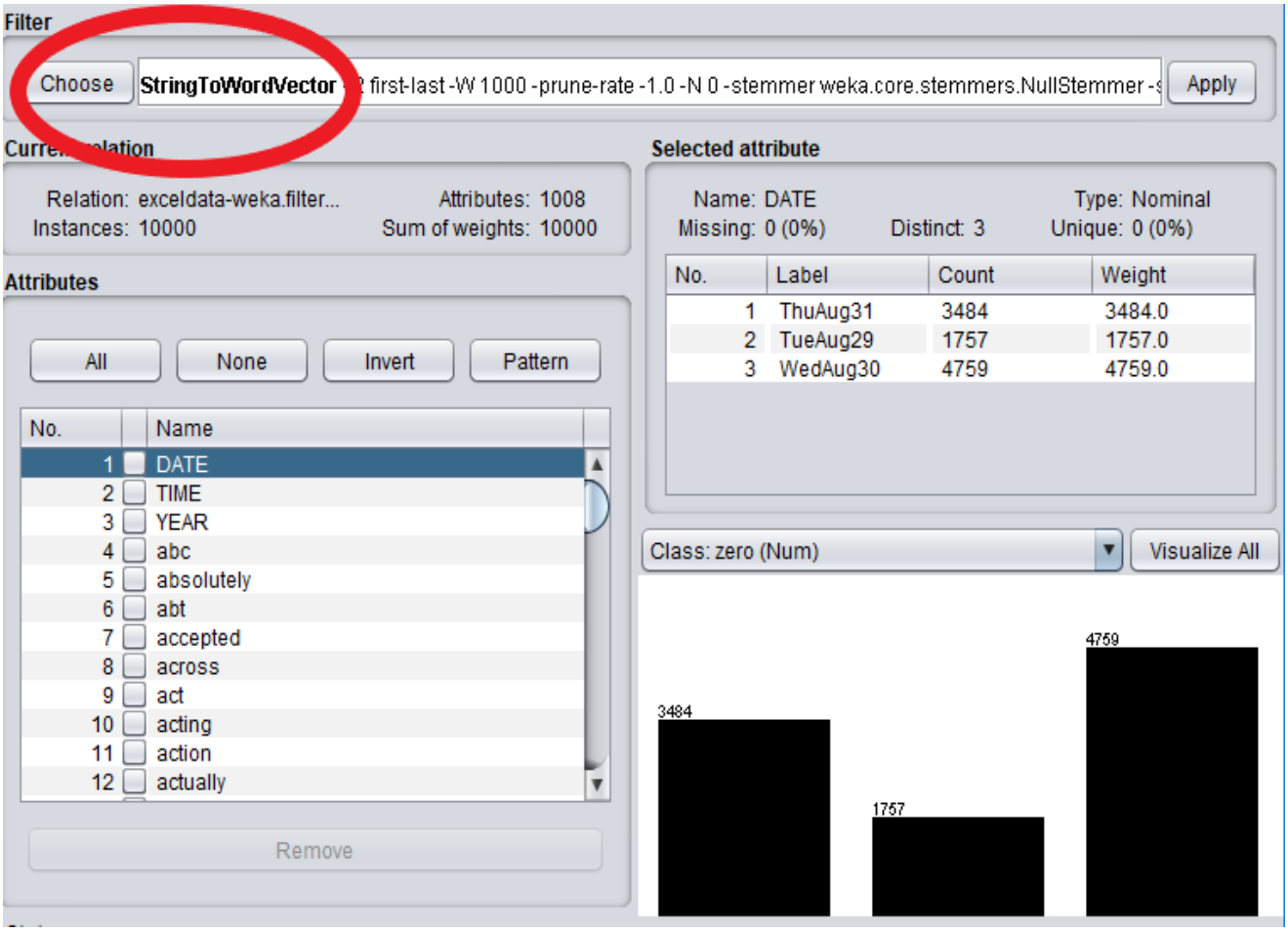
## MILESTONE 3: Mining Tool Preparation

This Milestone needs the file to be loaded to Weka, I have converted the CSV to ARFF with the help of R.

```
#Writing to CSV to ARFF
#-------------------------------
write.csv(data,"E:/PROJECT/hurriane.csv")
exceldata = read.csv('E:/PROJECT/hurriane.csv')
write.arff(exceldata, file='E:/PROJECT/hurriane.arff')
```

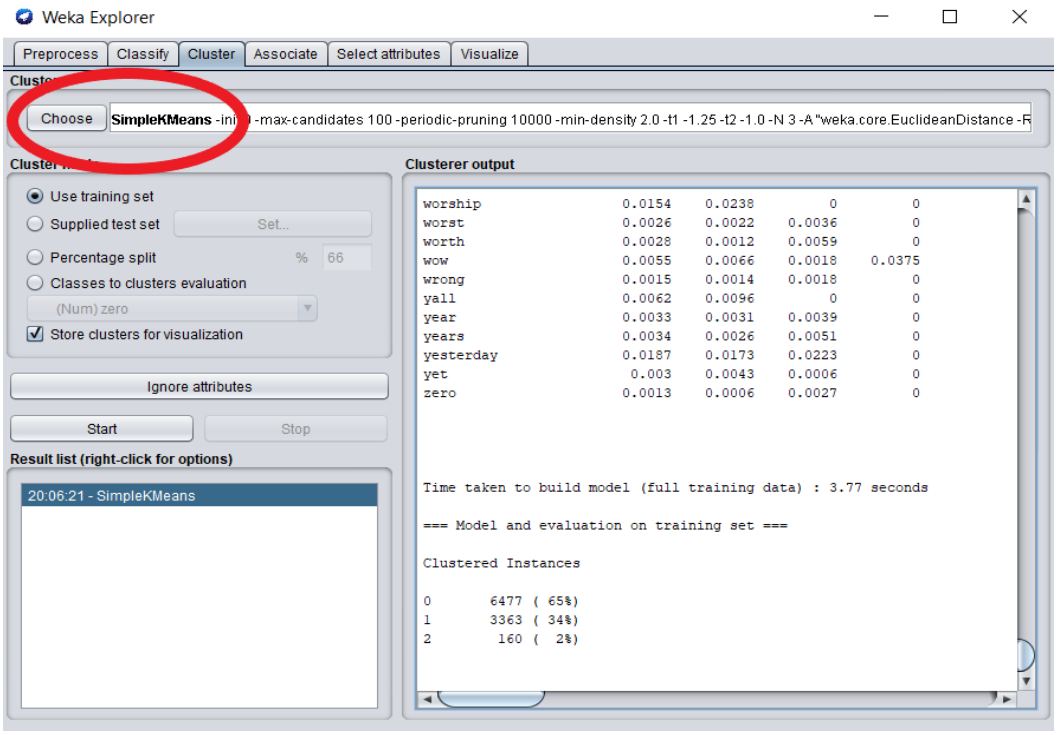
It is converted to ARFF as the TWEET is nominal.

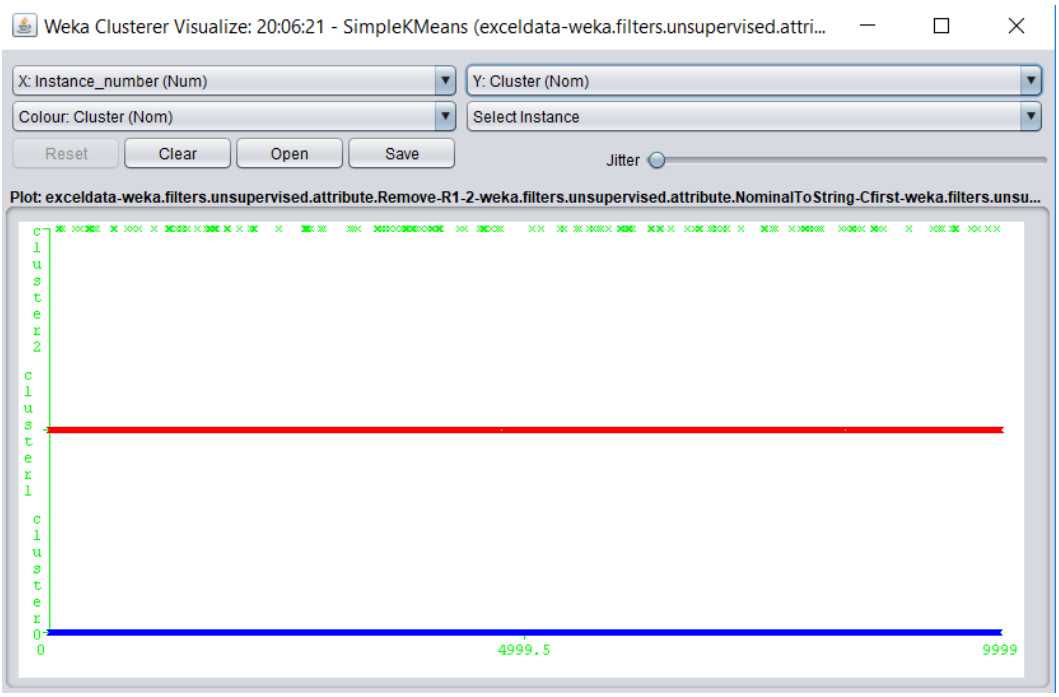
Steps: Unsupervised Learning -> Attributes->Nominal to String->String to Word Vector Filters.



## MILESTONE 4: Clustering Analysis

Simple K-means algorithm is used in this Milestone and applied on the TWEET column, here K is set to 3.



What we observe in above snapshot having 3 clusters is that, the Inter cluster distance is maximized and intra cluster distance is minimized.

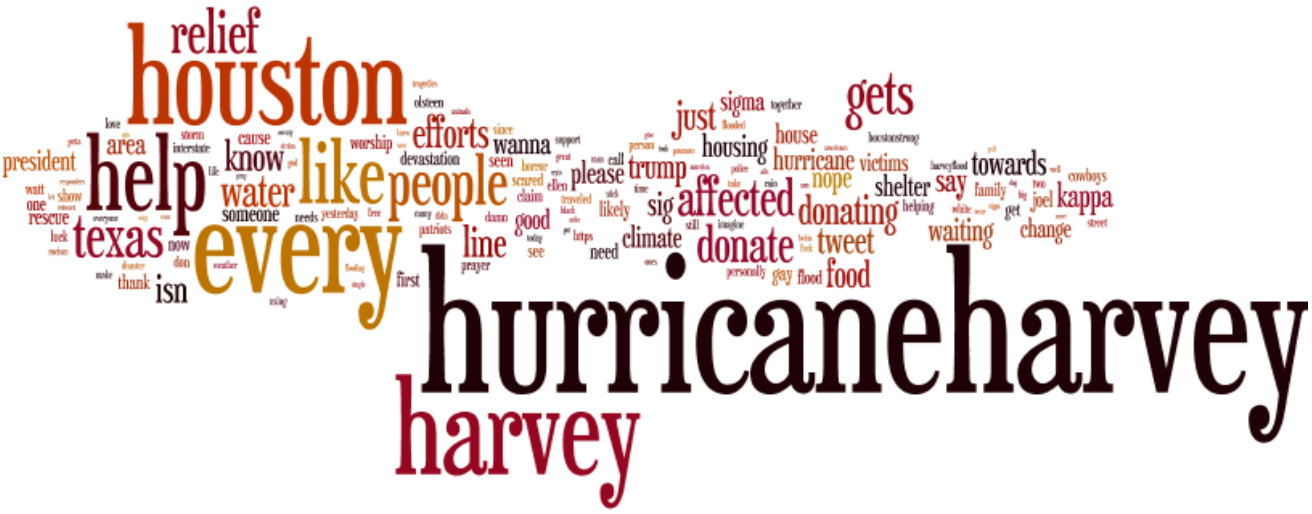
## MILESTONE 5: Visualization

The output of ARFF is converted to CSV for visualizing the data.

| | | |
|---|---|---|
| 3251 | agree think needs make significant contribution organization helping | cluster0 |
| 3252 | sandra bullock donated million harvey relief efforts politics eight feet water | cluster2 |
| 3253 | fuckin cryin ain scared flood nigga scared yall | cluster0 |
| 3254 | houston wanna know line food water housing nope people waiting lin | cluster0 |
| 3255 | beaumont loses water supply wake historic flooding hurricane harvey abc news | cluster1 |
| 3256 | white house reporters must demand others administration explain | cluster0 |
| 3257 | magic houston native jonathon simmons details escape floods hurricane harvey espn | cluster1 |

**Word clouds are made for each of the Cluster from the CSV file. It is further given an explanation using the DIKW.**

**Cluster-1**



**Data:**

Hurricane Harvey, Houston, relief, gets, help, Texas, like, every, affected, donate, water, housing, shelter, people.

**Information:**

Harvey affected people get donations. Houses are affected by Hurricane in Texas. People help hurricane victims.

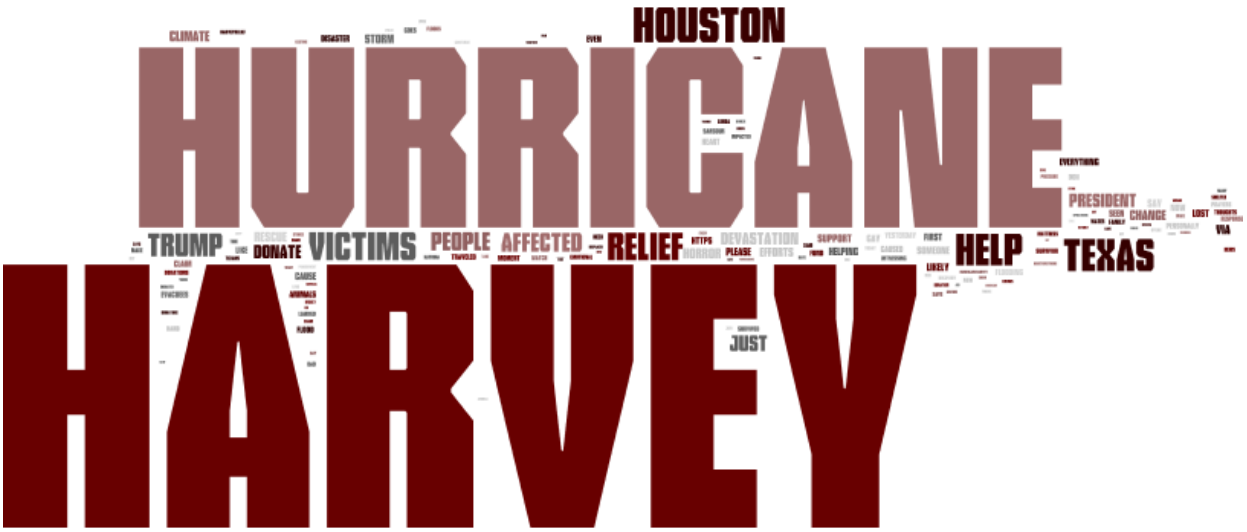
**Knowledge:**

Hurricane Harvey victims in Texas gets help from people by various donations and shelter.

**Wisdom:**

The victims of the Hurricane Harvey in Texas get help from people in the form of shelter, food and water.

**Cluster-2:**



**Data:**

Hurricane Harvey, trump, help, Texas, people, affected, relief, victims, Houston, President, devastation.

**Information:**

Trump helps Harvey affected people. Victims from devastation are in relief after help.

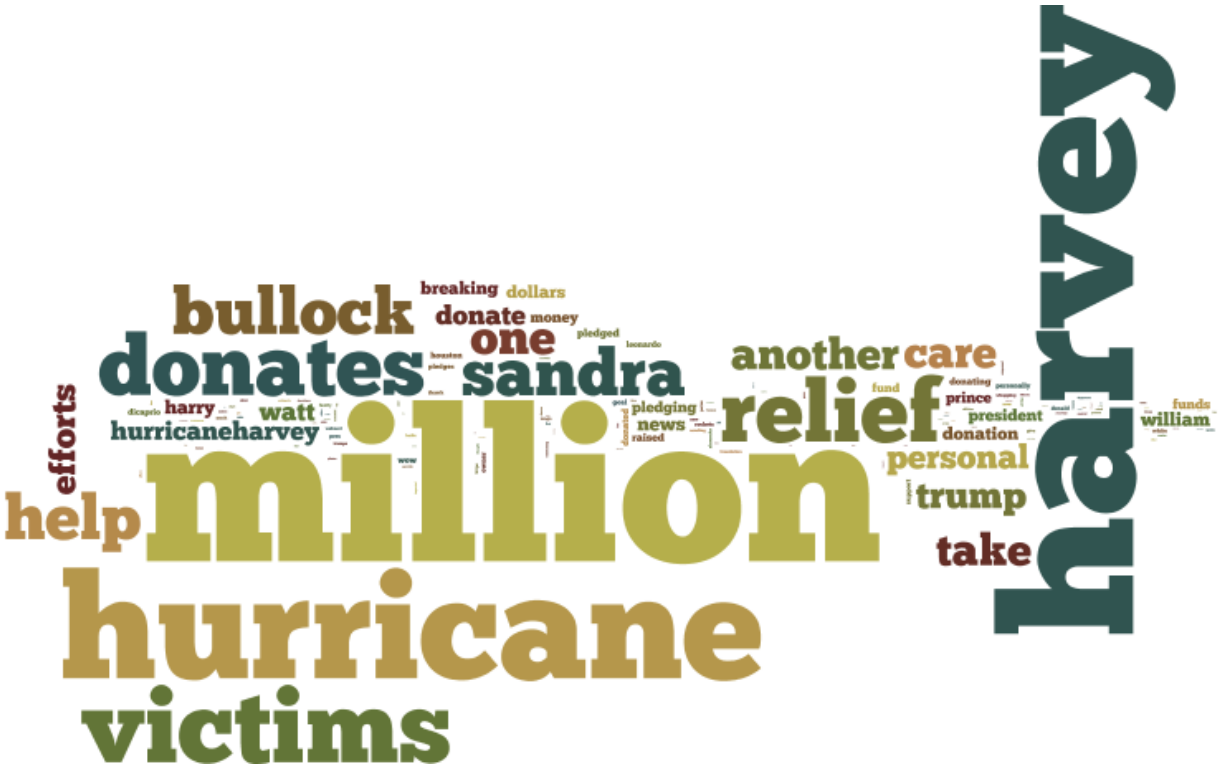
**Knowledge:**

President Trump helps the people in Texas affected by Hurricane Harvey.

**Wisdom:**

The victims of the Hurricane Harvey who were devasted gets help from President Trump and other people in the form donations to spread relief.

**Cluster-3:**




**Data:**

Help, million, hurricane, victims, Sandra Bullock, donate, one, Harvey, relief, efforts, personal.

**Information:**

Sandra Bullock helps by donating money. One million is donated towards the relief.

**Knowledge:**

Sandra Bullock personally donates one million dollars towards the relief for affected people.

**Wisdom:**

Sandra Bullock donates one million dollars personally towards the people who were the victims of Hurricane Harvey in Texas.

**Top 10 Words for each Cluster.**

**Cluster-1**

```
> #for Cluster-1
>   data2 <- Corpus(VectorSource(vdata$CLUSTER0))
> data1 <- TermDocumentMatrix(data2)
> mat <- as.matrix(data1)
> type <- sort(rowSums(mat),decreasing=TRUE)
> freq <- data.frame(word = names(type),freq=type)
> head(freq, 10)
                          word freq
hurricaneharvey hurricaneharvey 1546
every                   every 1133
houston                 houston 1059
harvey                   harvey 1045
help                       help  810
like                       like  686
people                   people  540
texas                     texas  502
gets                       gets  496
relief                   relief  481
```

**Cluster-2**

```
> #for Cluster-2
> data2 <- Corpus(VectorSource(vdata$CLUSTER1))
> data1 <- TermDocumentMatrix(data2)
> mat <- as.matrix(data1)
> type <- sort(rowSums(mat),decreasing=TRUE)
> freq <- data.frame(word = names(type),freq=type)
> head(freq, 10)
              word freq
harvey       harvey 3299
hurricane hurricane 2259
houston     houston  438
texas         texas  403
help           help  395
victims     victims  352
relief       relief  345
trump         trump  287
affected   affected  239
people       people  230
```

**Cluster-3**

```
> #for Cluster-3
> data2 <- Corpus(VectorSource(vdata$CLUSTER2))
> data1 <- TermDocumentMatrix(data2)
> mat <- as.matrix(data1)
> type <- sort(rowSums(mat),decreasing=TRUE)
> freq <- data.frame(word = names(type),freq=type)
> head(freq, 10)
              word freq
million     million  160
harvey       harvey  138
hurricane hurricane  110
victims     victims   77
relief       relief   65
donates     donates   63
bullock     bullock   49
sandra       sandra   49
help           help   47
one             one   36
```

```
#for Cluster-1
  data2 <- Corpus(VectorSource(vdata$CLUSTER0))
data1 <- TermDocumentMatrix(data2)
mat <- as.matrix(data1)
type <- sort(rowSums(mat),decreasing=TRUE)
freq <- data.frame(word = names(type),freq=type)
head(freq, 10)
```

**Insights retrieved from the top 10 words for each Cluster.**

**Cluster-1:** Hurricane Harvey at Houston gets Relief from people.

**Cluster-2:** Trump helped Hurricane Harvey affected people in Texas.

**Cluster-3:** Sandra Bullock donates one Million for Harvey Relief.

*Code for Top 10 Words.*

```
#for cluster-1
  data2 <- Corpus(VectorSource(vdata$CLUSTER0))
data1 <- TermDocumentMatrix(data2)
mat <- as.matrix(data1)
type <- sort(rowSums(mat),decreasing=TRUE)
freq <- data.frame(word = names(type),freq=type)
head(freq, 10)
```