**AIT 664: Information Processing: Representation, Mining and Visualization**

**Individual Class Project Report Spring 2018**

**Meena Rapaka**

**G01101759**

**Project Overview:**

To gain hands-on experience to process data and extract information and utilize metadata to discover patterns in the data mining process.

**Scenario of the Project:**

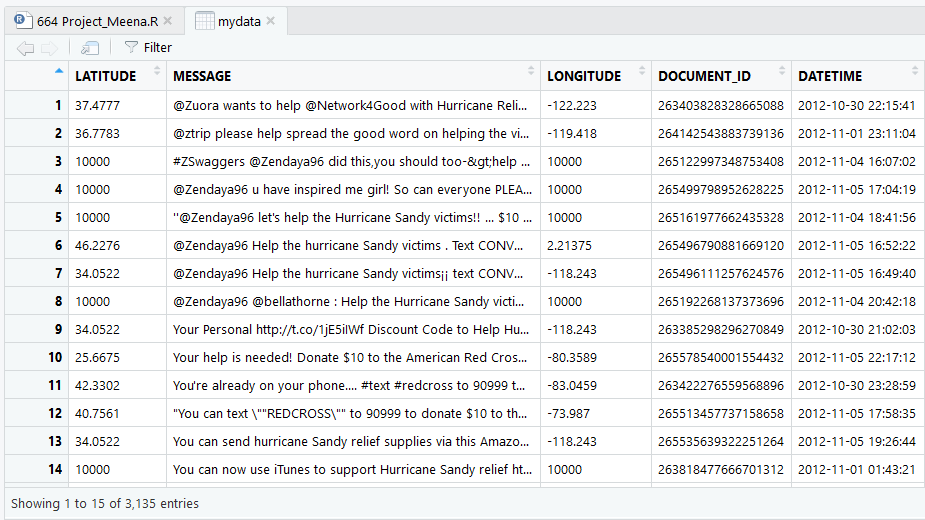
You are an information professional for an emergency response organization, you analyze and monitor social media data streams during a disaster event. You want to identify the patterns from information that are helpful to understand the situation for the unfolding event.

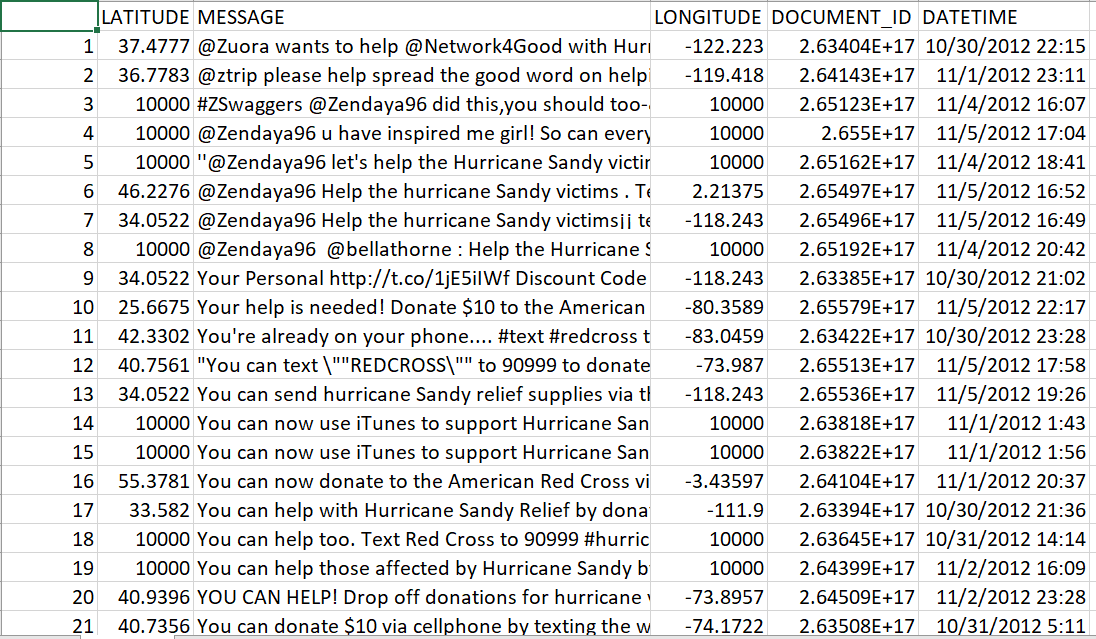
**Milestones:**

We have five milestones to fulfill the key task completion.

1. **Data Acquisition and Conversion**

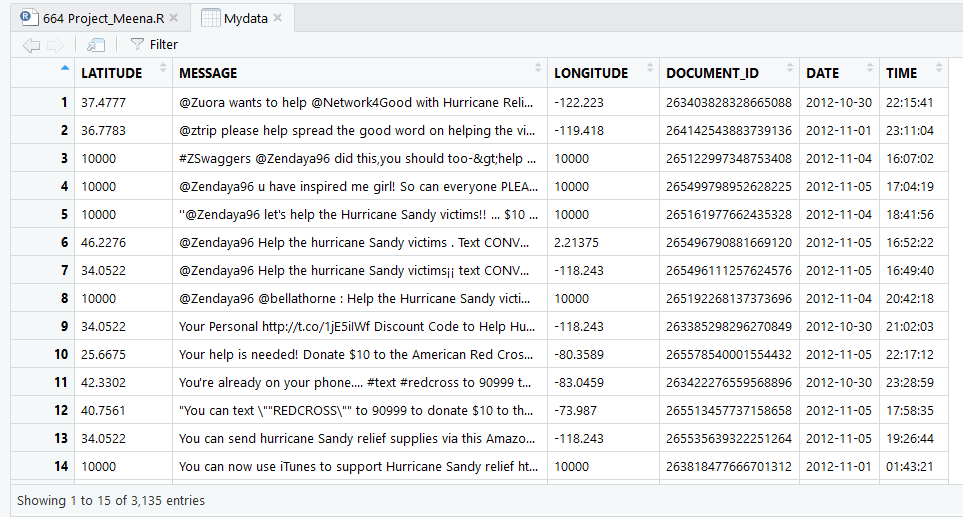
The target of this step is to download and then convert the given (<http://ist.gmu.edu/~hpurohit/courses/ait690-proj-data-spring17.json>) JSON file into Comma separated values format. I have used R programming language to convert the downloaded JSON file into CSV format using jsonlite package.



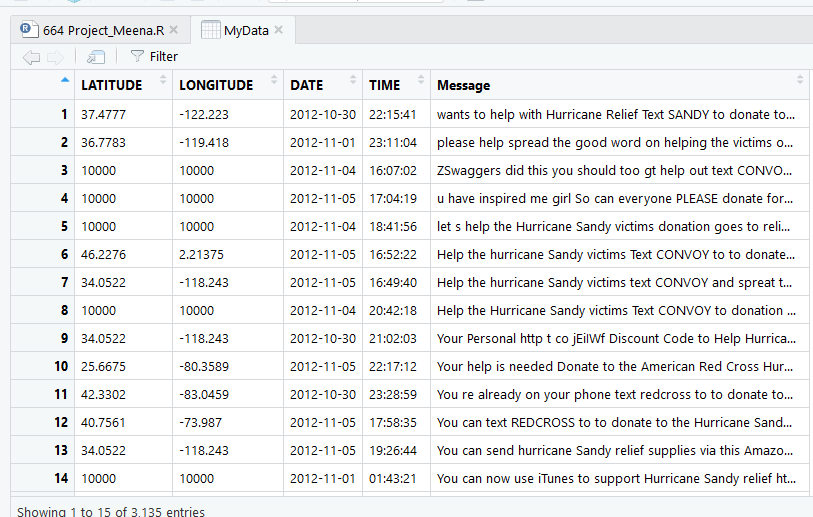


1. **Data Preprocessing**

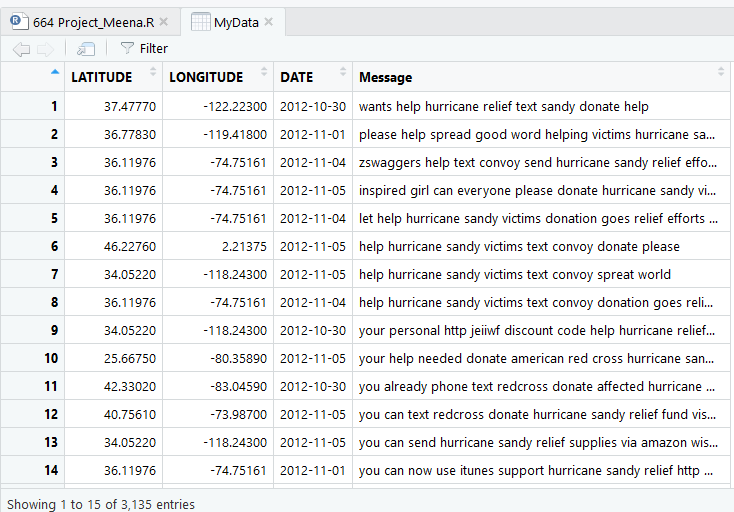
After converting to CSV, we could observe that Data and time are not separated and there were 3135 tweets. So, I have segregated the DataTime into two separate columns using the Split string function.



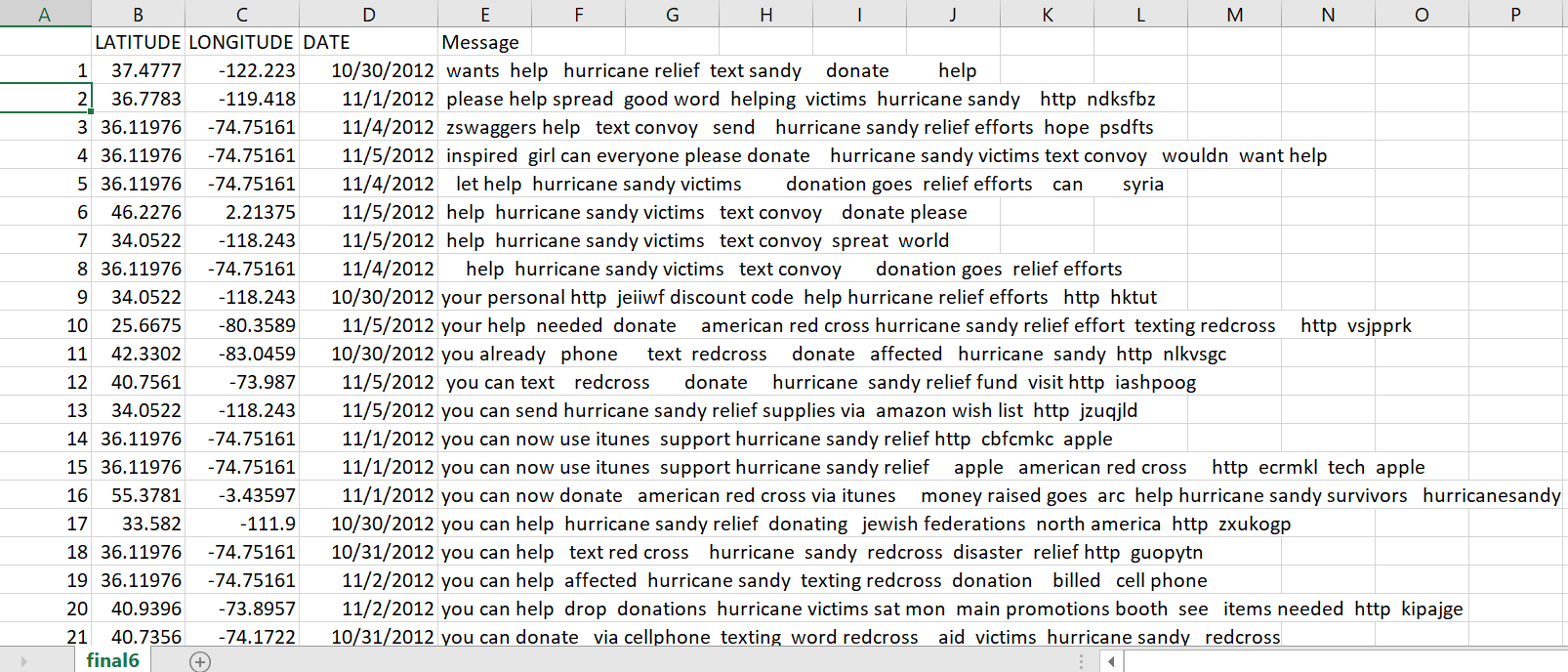
And then, cleaned up the Message column, it had many Alpha numeric values and Numbers, I used Gsub function to remove the unwanted data from the column. And removed user names, symbols, digits and changed the case to lower letters and used NLP and stopwords libraries to remove stop words.



Then, I replaced the 10000 values in latitude and longitude with the mean of respective columns.

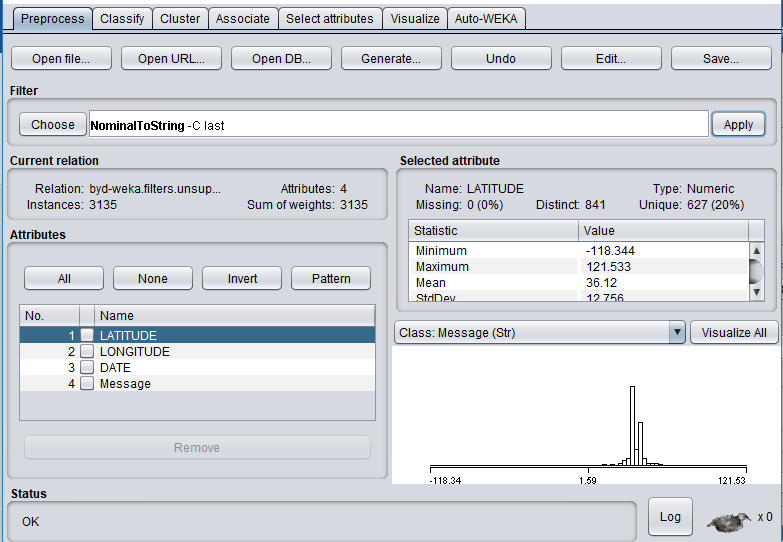


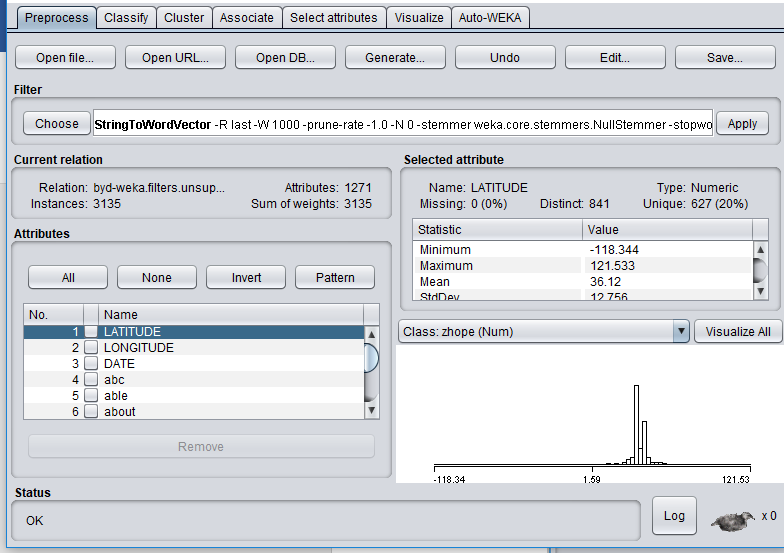
This is the CVS file we get after Data Preprocessing is completed.



1. **Mining Tool Preparation**

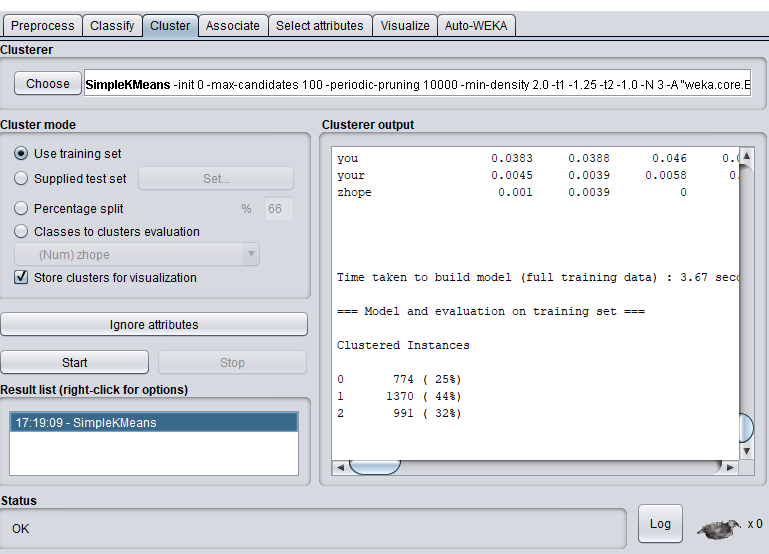
For Milestone 3, we first the convert the csv file into ARFF and then load the cleaned dataset into WEKA. Initially the message column is Nominal, I have converted it into string using the filter “NominalToString”, then converted it into wordvector using “StringToWordVector” Filter.



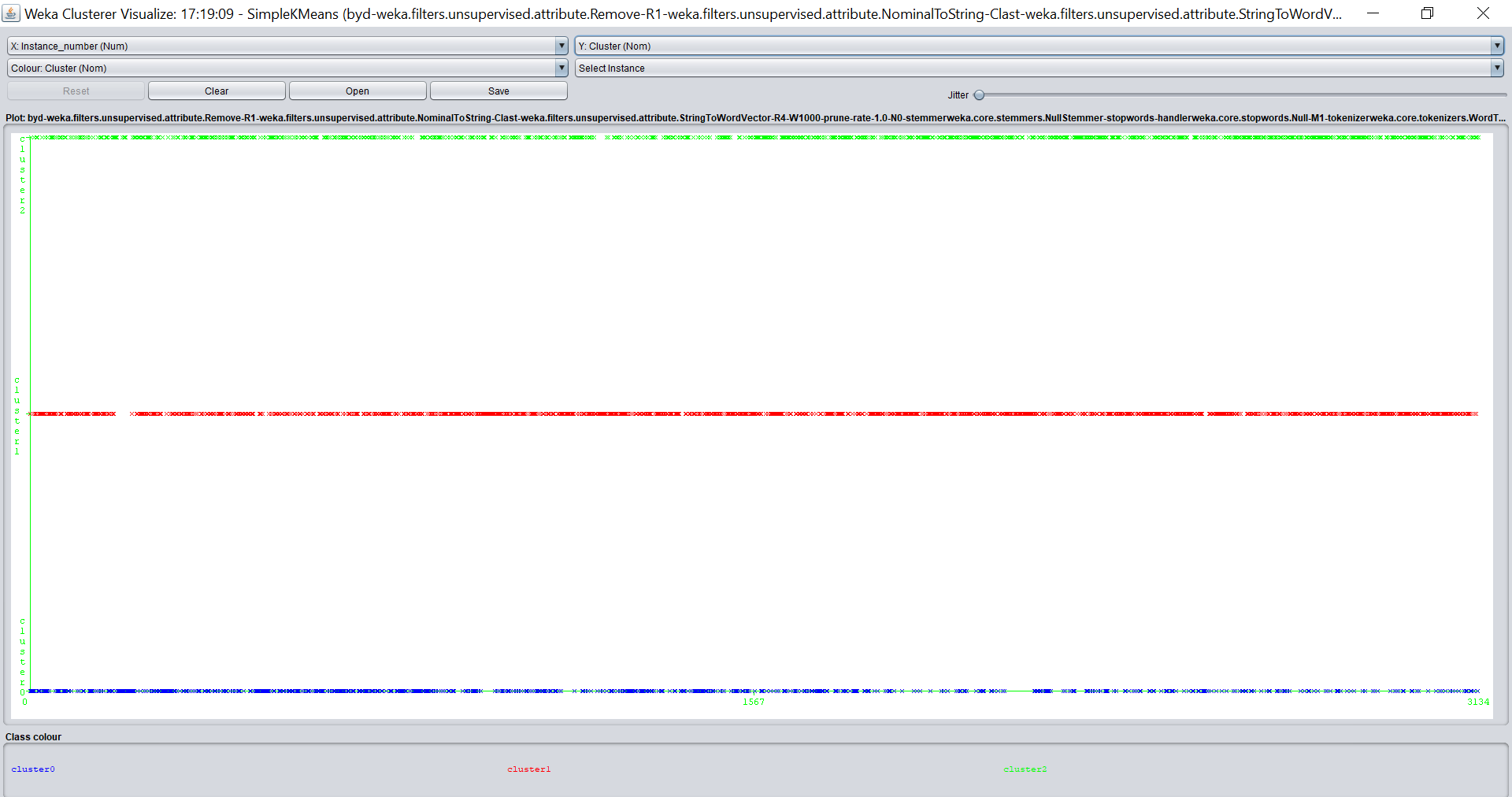


1. **Clustering Analysis**

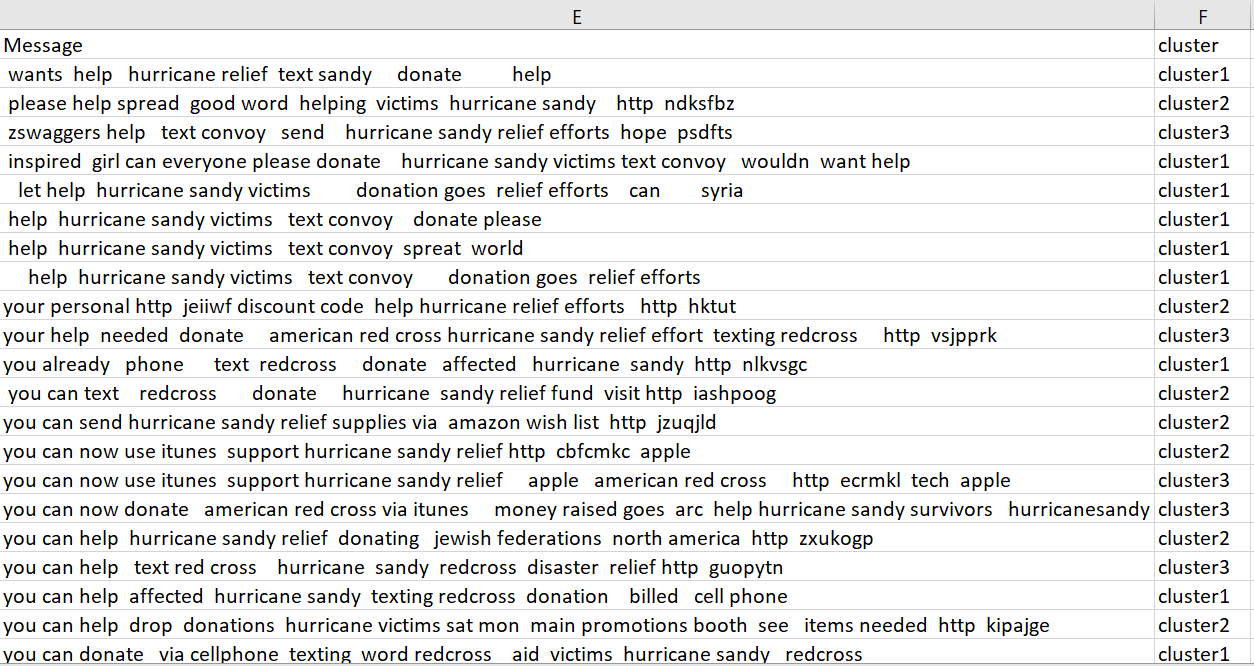
Here, we cluster the message column using Simple K means clustering method with minimum of 3 clusters.



When we view the visualization for cluster assignments, we observe that the cluster distance between intra-cluster is minimized and inter-cluster is maximized.



**Cleaned Spreadsheet**



1. **Visualization**

For this step, I saved three text files for each specific to produce word cloud. I have used Semantic Word Cloud Visualization tool for the same. I have generated top 100 and top 10 words for each cluster.

**Cluster 0**

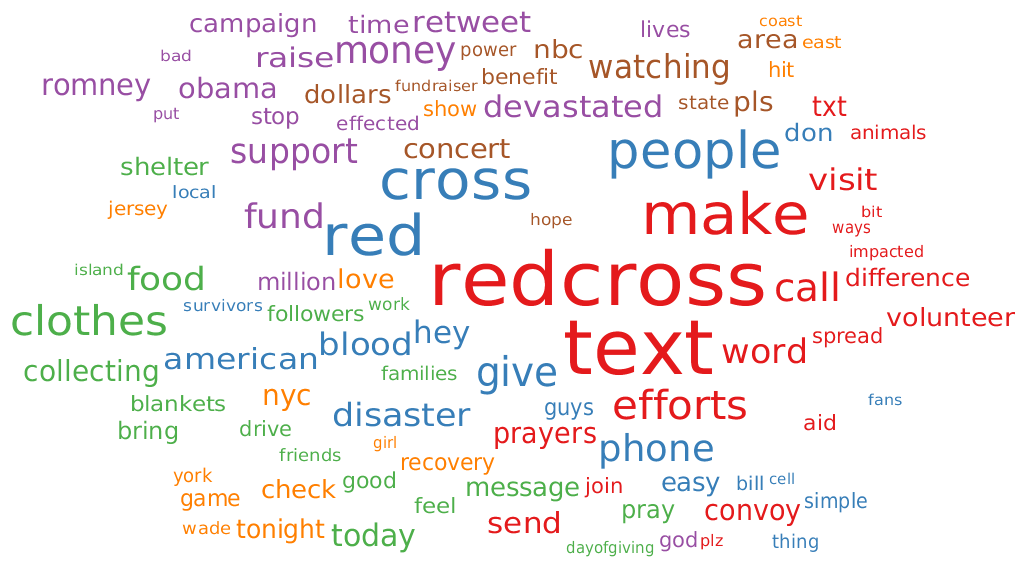
**Data:** Hurricane, Sandy, text, victims, red cross, make, sandy help, prayers, disaster, message, recovery, campaign, clothes, jersey, retweet, devastated

**Information:** A hurricane named Sandy has hit in New York and New jersey regions. There are many people who are affected by this hurricane. Many people need help.

**Knowledge:** People use social networking to relay information about the hurricane. Organizations like redcross are helping the victims with food, clothes and funds. News channels like NBC made efforts to cover the incident and broadcast the information.

**Wisdom:** The outcome for this cluster is related occurrence of the hurricane.

**Top 100 words for Cluster0**



**Cluster 1:**

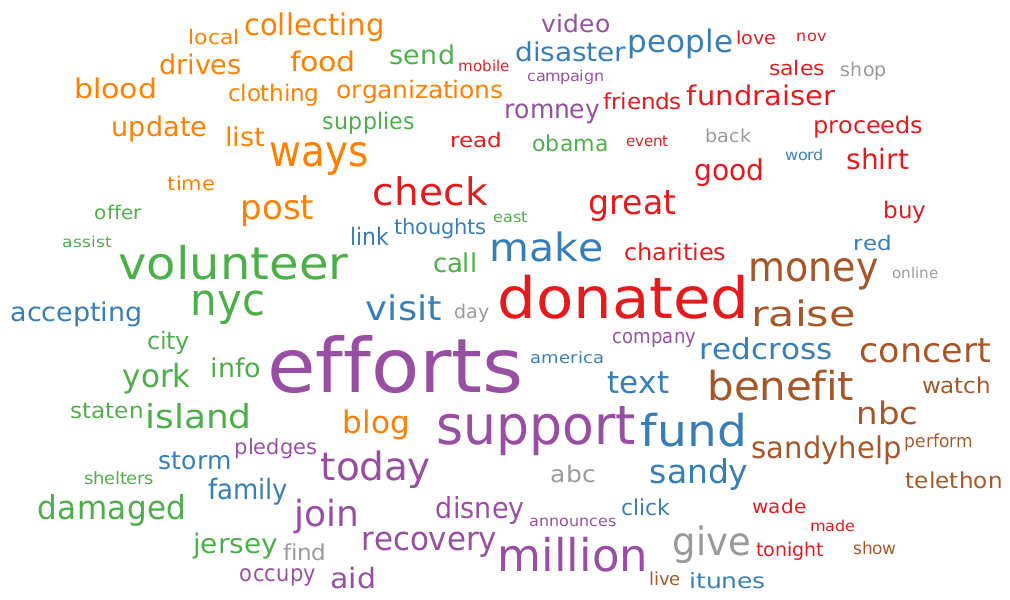
**Data:** efforts, disaster, NYC, new jersey, sandy, support, volunteer, money, raise, fund, million, fund, recovery, benefit, damaged, charities, organizations

**Information:** New York city and New Jersey have been affected by the hurricane and millions of lives have been lost due to the hurricane sandy. People and animal’s lives have been compromised.

**Knowledge:** People needed help in terms of blood, first aid, food, clothing. There are many organizations providing help.

**Wisdom:** The outcome of this cluster indicates more about affects caused by the hurricane and loses experienced by people.

**Top 100 words for Cluster0**



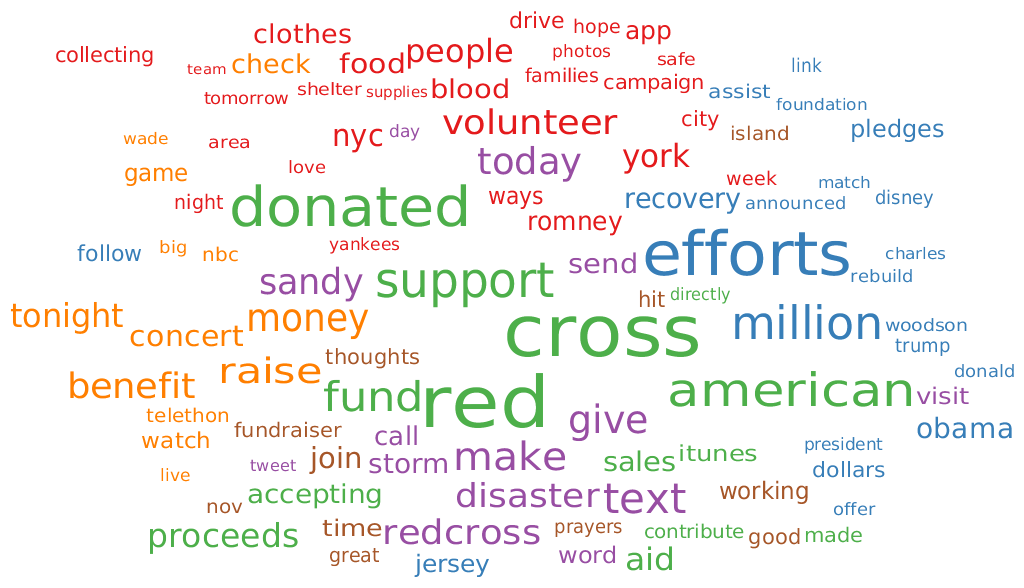
**Cluster 2:**

**Data:**  American, prayers, benefit, concert, raise, support, fund, red cross, volunteer, money, pledges, make, thoughts, assist, campaign

**Information:**  Many of the organizations are coming forward to help the hurricane victims like red cross. They are organizing benefit shows to help them. Everyone is in their prayers.

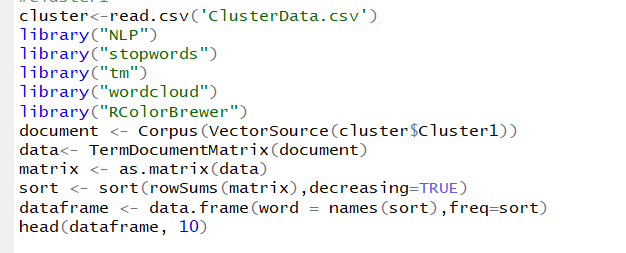
**Knowledge:** Many fund-raising shows are being organized to help the victims.

**Wisdom:** How people are raising funds for the victims and people praying for their health and losses.



**b**. In this step, we visualize top 10 words in each cluster using R.

**Code I used to get top 10 words is:**



**Top 10 words in Cluster 0 are:** This cluster provides insight about the occurrence of Hurricane Sandy.

word freq

help help 521

text text 270

redcross redcross 267

please please 175

can can 115

affected affected 100

make make 88

red red 87

cross cross 83

relief relief 81

**Top 10 words in Cluster 1 are:** This cluster is about organizations like red cross helping affected victims.

word freq

relief relief 577

red red 390

cross cross 387

efforts efforts 163

help help 146

american american 100

please please 88

the the 73

million million 71

affected affected 69

**Top 10 words for Cluster 2 are:** This cluster provides insights about efforts being taken to report the incident.

word freq

relief relief 743

help help 412

efforts efforts 236

can can 163

affected affected 107

please please 104

how how 92

the the 86

new new 74

via via 72