AIT 664: INFORMATION: REPRESENTATION, PROCESSING, AND VISUALIZATION

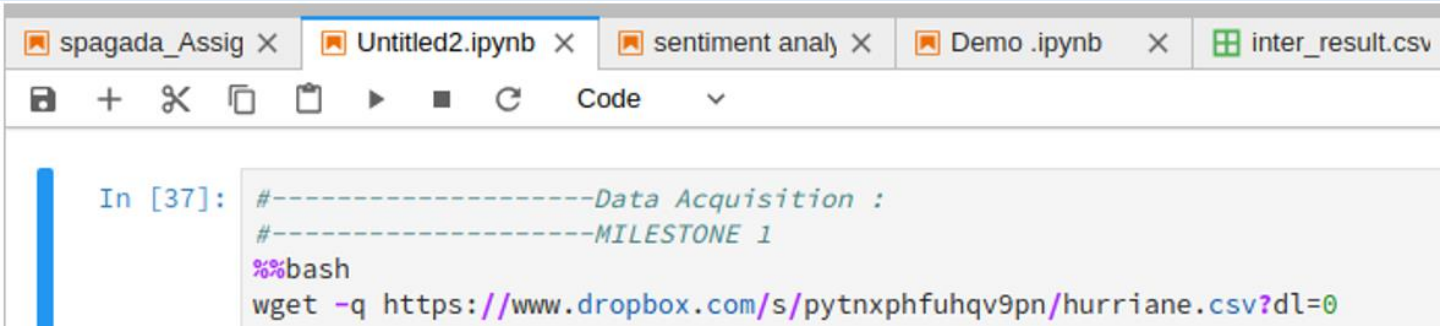
SRAVYA PAGADALA - G01161487

1. Objective

The objective of this project is to gain hands-on experience to process data, to extract information, and discover patterns using data mining method and identify the information and patterns that are helpful to facilitate decision making from tweets on Hurricane Harvey.

2. Milestone 1 : Data Acquisition

Programmatically downloading the project data file on Hurricane Disaster Management.



Downloading programmatically is achieved using the function "**wget"**. Upon observations ,the downloaded file is a CSV format file, and in a table format, Containing the tweets and the time of their creation is in a tabular form like below:

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

1. TWEET_TEXT: Content of the tweet in the form of text (String).
2. CREATION_TIME: Consists of the time of occurrence of each tweet including the day of the week, month, date, time and year respectively.

3. Milestone 2 : Data Preprocessing

The following steps are implemented in R for preprocessing milestone on the downloaded csv data file from Milestone 1 output as these steps can they can help in improving your performance:

- **Read the csv file:**

  Using read() function in R the csv is read and loaded into a Dataframe.

- **Drop NA :**

  Dropped rows from a Data Frame with missing values on a given variable using the function DropNA().

- **Remove Time :**

  Discarded time part in the DATETIME field by writing a function in R.

- **Split Time column :**

  The month and day are sorted into different columns and is split using R function with a regular expression.

- **Remove unnecessary patterns :**

  Patterns like "_url_" and "_URL_" are removed from the message string.

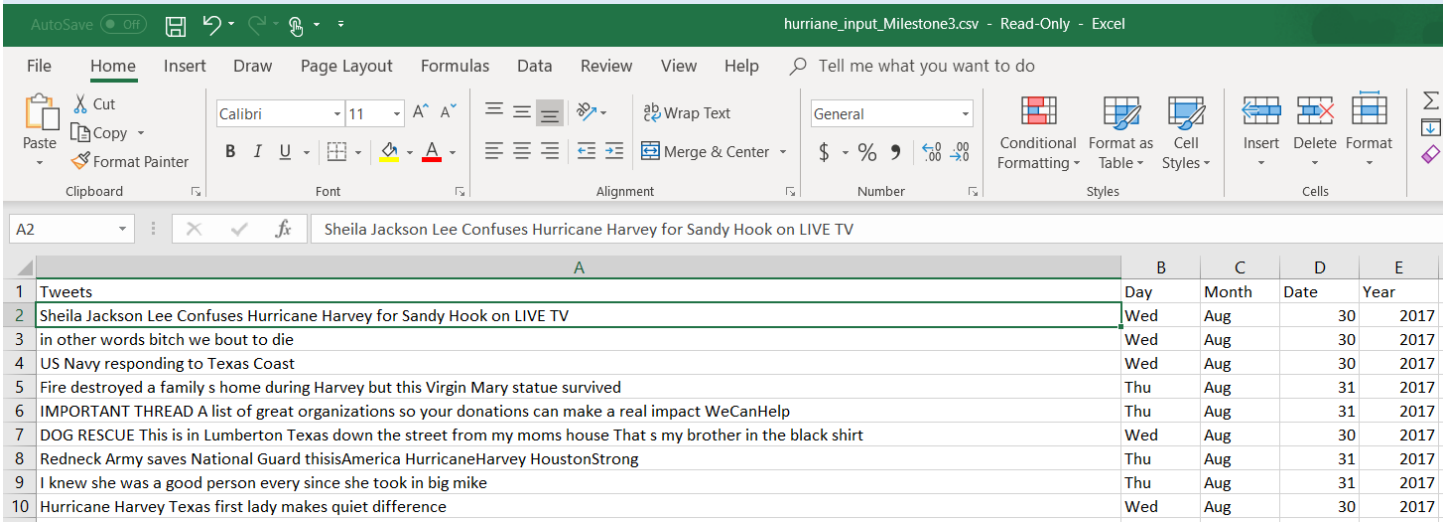
- **Removed stop words :**

  In computing, stop words are words which are filtered out before processing of natural language data (text).

- **Perform stemming :**

  Stemming is the process of reducing the words(generally modified or derived) to their word stem or root form. The objective of stemming is to reduce related words to the same stem even if the stem is not a dictionary word. Porter Stemmer()function is used in R and before this the string is tokenized into words.

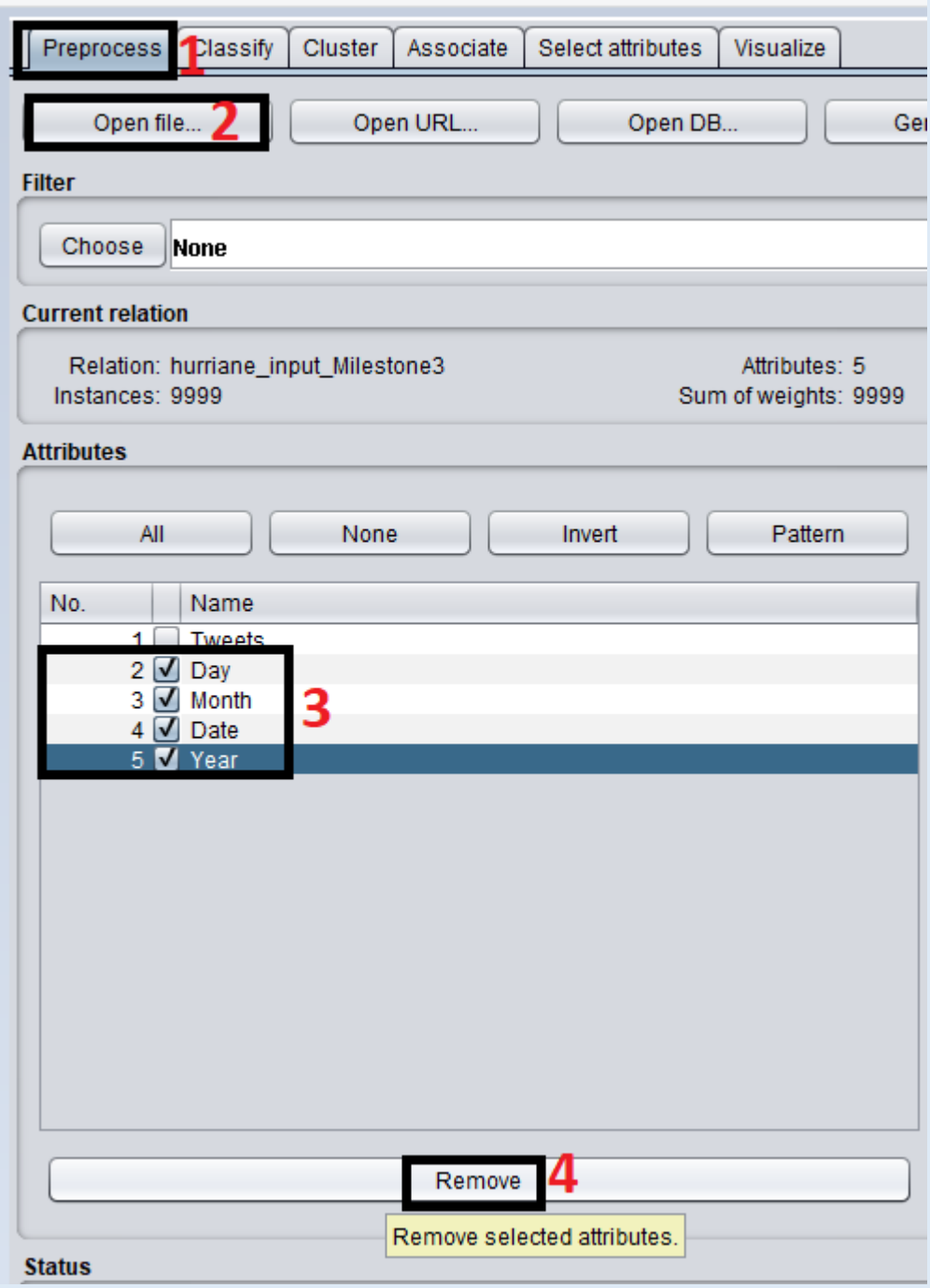
- **Save Output :**The output is stored in a CSV file and it looks like the following after preprocessing.



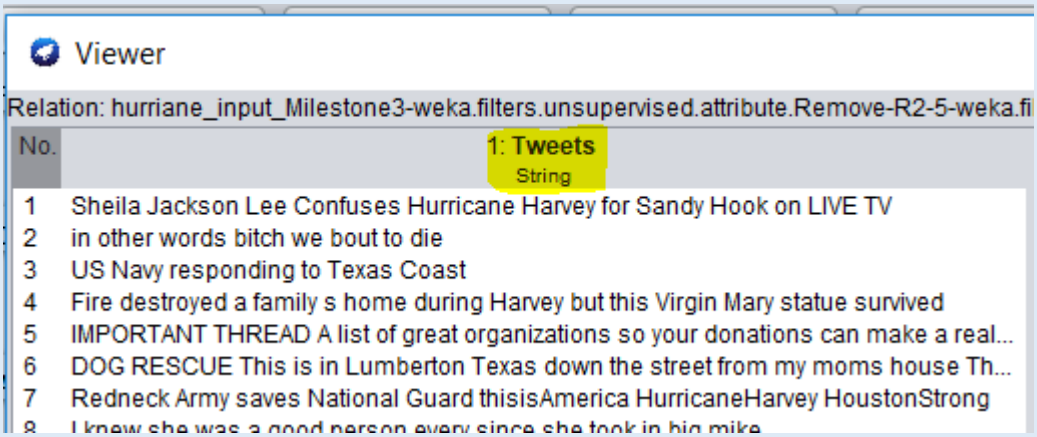
	A	B	C	D	E
1	Tweets	Day	Month	Date	Year
2	Sheila Jackson Lee Confuses Hurricane Harvey for Sandy Hook on LIVE TV	Wed	Aug	30	2017
3	in other words bitch we bout to die	Wed	Aug	30	2017
4	US Navy responding to Texas Coast	Wed	Aug	30	2017
5	Fire destroyed a family s home during Harvey but this Virgin Mary statue survived	Thu	Aug	31	2017
6	IMPORTANT THREAD A list of great organizations so your donations can make a real impact WeCanHelp	Thu	Aug	31	2017
7	DOG RESCUE This is in Lumberton Texas down the street from my moms house That s my brother in the black shirt	Wed	Aug	30	2017
8	Redneck Army saves National Guard thisisAmerica HurricaneHarvey HoustonStrong	Thu	Aug	31	2017
9	I knew she was a good person every since she took in big mike	Thu	Aug	31	2017
10	Hurricane Harvey Texas first lady makes quiet difference	Wed	Aug	30	2017

4. Milestone 3 : Mining Tool Preparation

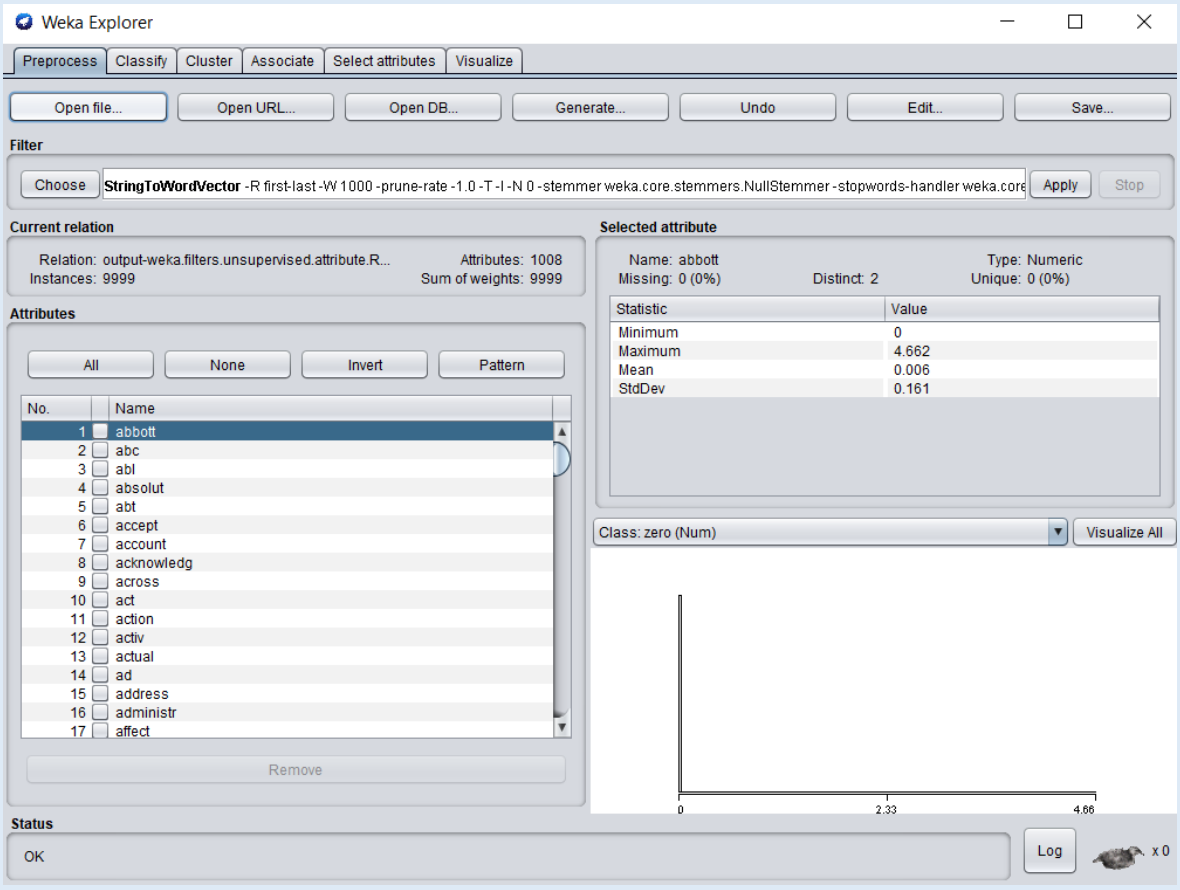
Weka GUI tool is installed and output of milestone 2 is the input file for Weka tool. Once the file is loaded in Weka remove the fields/attributes that are not needed for k-NN algorithm implementation.



The words are clustered basing on the keywords which have the highest number of frequencies and not on the time or date of the generation of tweet. Since the TWEET message is in Nominal type, first step is to convert “NominalToString” filter in the first tab of WEKA i.e. Preprocess. After the above filter the tweet attribute is converted from Nominal to String.

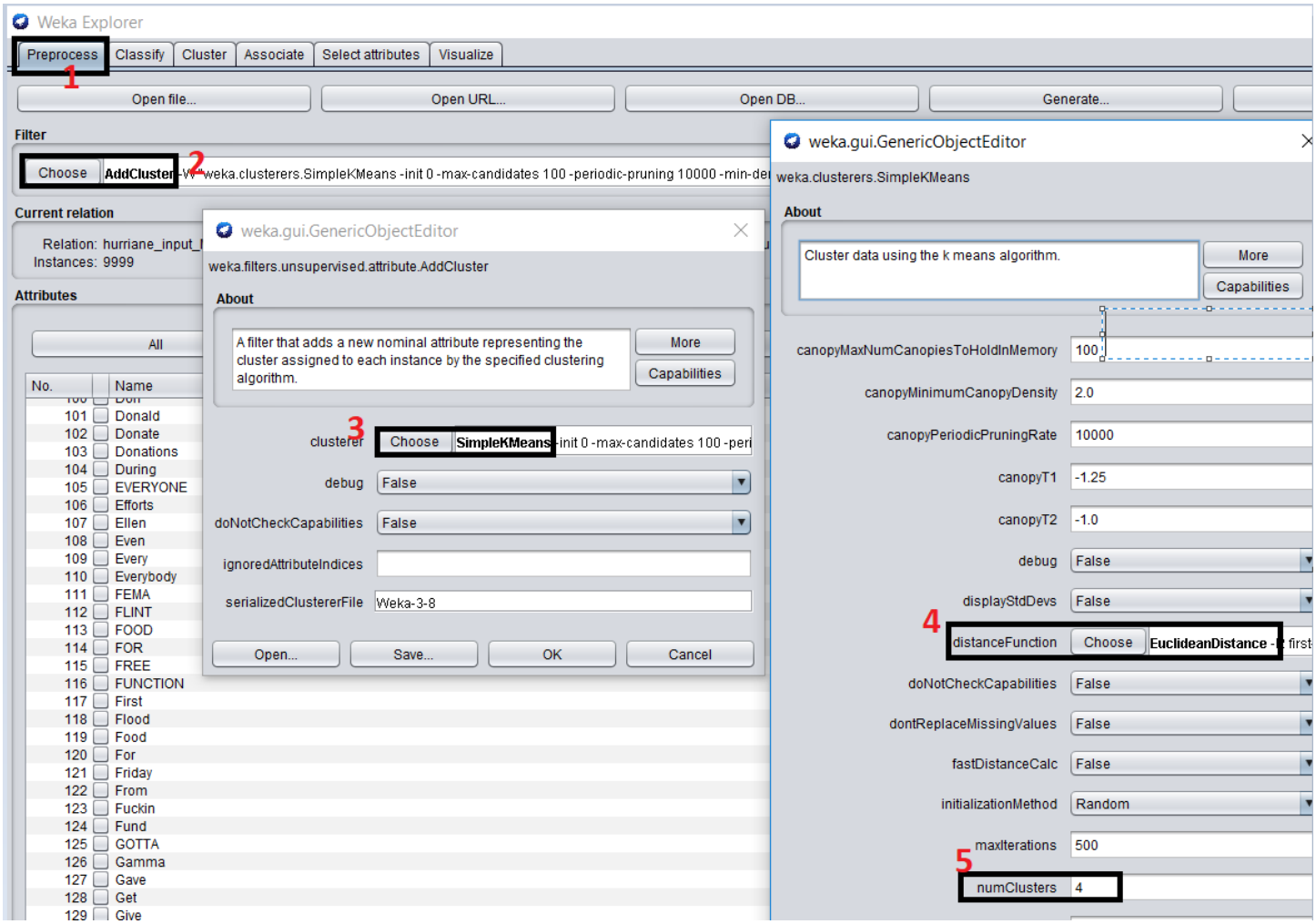


Then once the tweet text is converted into String format , then 'Filter' to apply "StringToWordVector", for transforming MESSAGE string into a vector of words, which become part of attribute set.



5. Milestone 4 : Clustering Analysis Implementation

Data mining is achieved by clustering the documents on the basis of frequent words on the Tweets attribute. For this, SimpleKMeans clustering technique is used with 4 clusters and the results are obtained as follows:



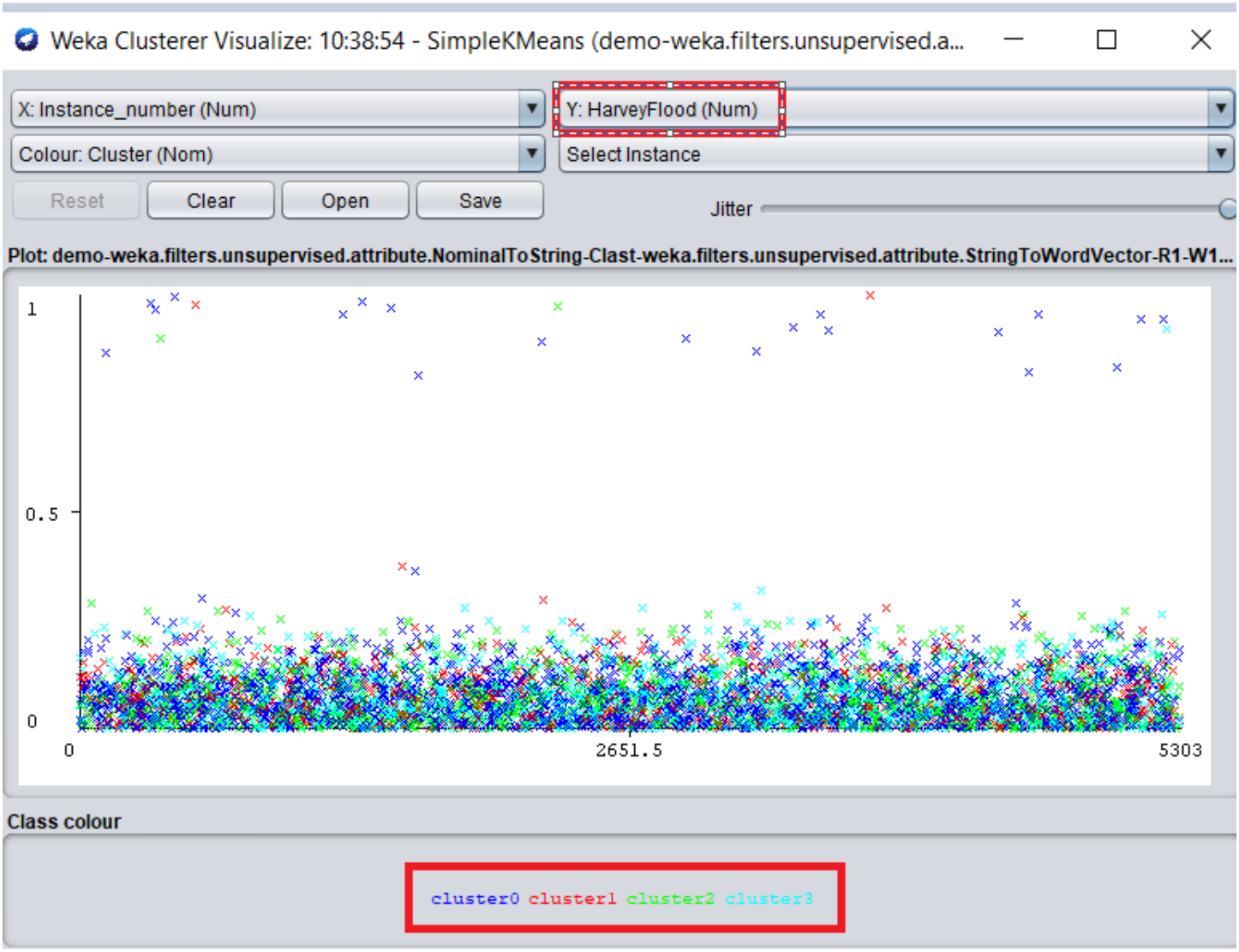
Using SimpleKMeans with 4 clusters I generated the results as shown with information regarding cluster assignments.
From this visualization we can observe that the word “Harvey flood” is most frequently occurring in Cluster0 as blue is the most dominant color observed. Similarly, this behavior can be observed for every attribute and the corresponding cluster can be obtained.

```
Time taken to build model (full training data) : 13.38 seconds

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

Clustered Instances

0       2048 ( 39%)
1        766 ( 14%)
2       1239 ( 23%)
3       1251 ( 24%)
```



6. Milestone 5 : Visualization

Before clustering the word cloud of tweets is shown below:

This shows that the tweets during the days 29, 30 and 31 of August 2017 mostly concerned about the hurricane Harvey and the context revolved around the donating the needy and tweets about help and relief for people affected by Hurricane Harvey.

Interpretations of observations in the DIKW framework :

Data: The input csv containing the tweet messages and creation time is the data. These bits of pieces values are still raw and comes in the form of raw observations and measurements.

Information: After pre-processing using R, the output of milestone 2 can be considered as Information as this is created by analyzing relationships and connections between the data. It's basically some sort of awareness from given data.

Knowledge: This is understanding the patterns of information such that it's intent is to be useful such as most frequent words from tweets.

Wisdom: The application of knowledge is Wisdom, thus information shared on Twitter by both affected population (e.g., requesting assistance, warning) and those outside the impact zone (e.g. providing assistance) would help first responders, decision makers, and the public to understand the situation first-hand and also contribute to situational awareness during disasters.

Cluster wise visualization using Voyant tool.

Cluster0:Below diagram shows the word cloud from cluster 0

Inferences from above word cloud:

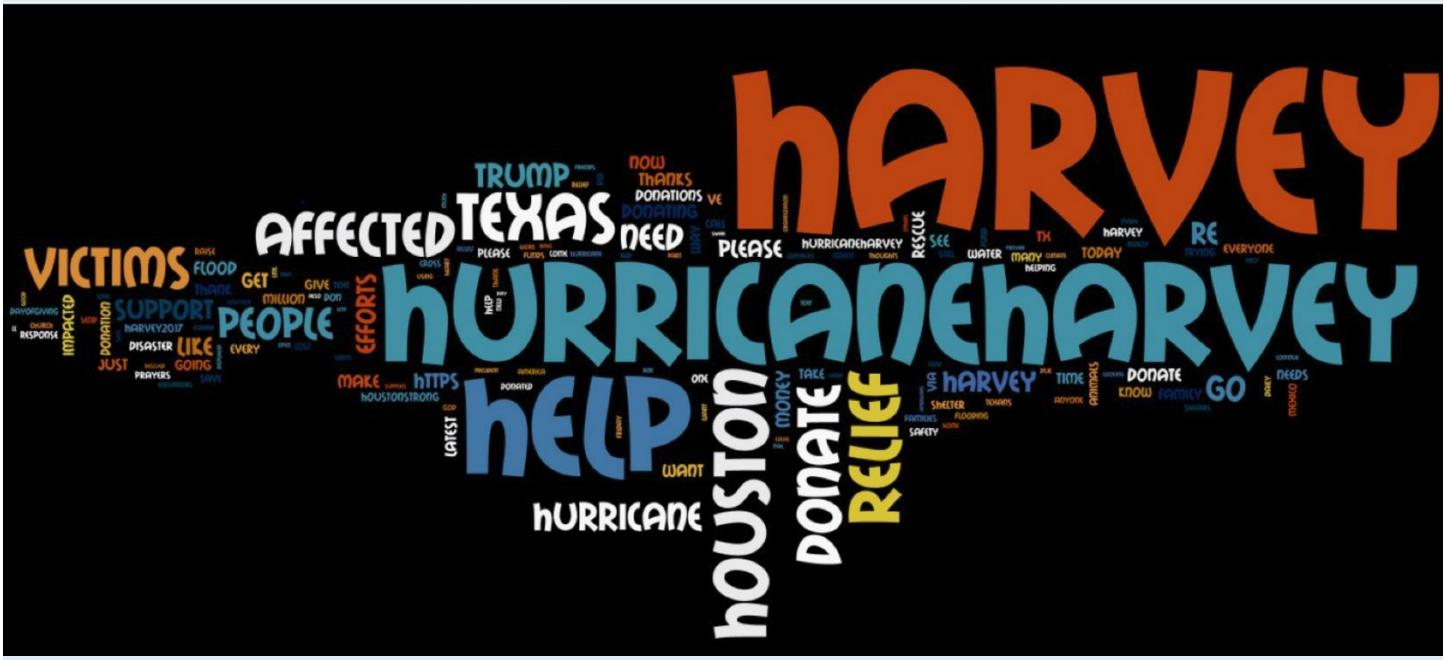
Most frequent words in the corpus: harvey (671); hurricaneharvey (439); houston (228); texas (220); help (116); hurricane (95); people (92); trump (78); like (73); victims (66)

Cluster 1:Above diagram shows the word cloud from cluster 1

Inferences from above word cloud:

Most frequent words in the corpus: harvey (312); hurricaneharvey (163); houston (133); texas (118); help (99); hurricane (87); victims (85); people (70); need (40); water (35)

Cluster2 :Below diagram shows the word cloud from cluster 2



Inferences from above word cloud:

Most frequent words in the corpus: harvey (481); hurricaneharvey (276); help (265); relief (142); donate (141); houston (140); texas (123); affected (107); victims (100); hurricane (87)

Cluster3 :Below diagram shows the word cloud from cluster 3

Inferences from above word cloud:

Average Words Per Sentence: 17688.0

Most frequent words in the corpus: harvey (1265); hurricane (1265); relief (205); victims (184); help (179); houston (142); trump (133); texas (127); affected (99); donate (71)

From the above word clouds , we can easily understand the most frequent words in each cluster and thus can take required measures that would help first responders, decision makers, and the public to understand the situation first-hand and also contribute to situational awareness during disasters.

7. References:

1) On Identifying Disaster-Related Tweets:Matching-based or Learning-based?Hien To, Sumeet Agrawal, Seon Ho Kim, Cyrus Shahabi Integrated Media Systems Center, University of Southern California, Los Angeles, USA

https://www.computer.org/csdl/proceedings/bigmm/2017/6549/00/07966769.pdf

2)WEKA tutorial

https://www.slideshare.net/butest/weka-tutorial

3) Voyant tutorial

http://docs.voyant-tools.org/start/