1. PRE-PROCESSING OF TWEETS

A Python script was written which with the help of Twitter Streaming API which collects the most recent raw tweets with keywords such as ‘Zika’, ‘Zika Virus’ ,‘Aedes’ in a text file and then each tweet is converted into JSON (JavaScript Object Notation) for easy manipulation and handling of data. Pandas, an open source library for data manipulation in Python is then used to store the data in a data frame with columns such as Twitter ID, created-at, text, favourite-counts etc.

A total of 4751 tweets were collected and after removing the re-tweets, we were left with 1471 unique tweets. The original tweet contains many elements other than the original text such as hashtags, external links, user mentions etc. Thus, for proper analysis of the tweet, from the text we separated 1) stop words such as ‘a’, ‘an’, ‘the’ etc. 2) user mentions 3) hashtags 4) URL’s or external website links 5) special characters such as emoticons. This process of segregation left us with the tweet containing only the main words. A special type of analytical methodology called the word-clouds was then used which when given an array of words, gives us insight into what words have the highest frequency and are important for the analysis. Therefore, world-clouds were generated for main text, hashtags and user-mentions.

The training tweets were then given class label according to the 3 classes – 1) Tweets related to fight and prevention against Zika. 2) Tweets related to cure for Zika. 3) Tweets related to damage caused by the Zika virus, mainly the infected areas and the death caused. Word-clouds were also generated for each of the three classes.