Data Science Report

Name: Syed Mohammad Ammar

Roll Number: 16k-3669

Section: GR1

# Data:

The data for this project was obtained from twitter via twitter streaming api tweepy. The data had been separated into two different datasets, based on two major Indian political party Congress and BJP information.

The dataset have variables such as:

* tweet
* country\_code
* geo\_location
* bounding\_box
* screen\_name
* favourites\_count
* followers\_count
* statuses\_count
* friends\_count
* listed\_count
* user\_described\_location
* created\_at
* utc\_offset

# Objective:

The Objective of this project was to analyze the Indian elections and determine whether twitter is a credible source for predicting results of elections. To predict the result. The models used was **Clustering** and **K-means.**

The Data Process

# Retrieving Data:

The data was retrieved from an external source API,the used api were tweepy and search api. The collected data is around 100k tweets for over a period of one week. The tweets were related to BJP and Congress .Also the data set is large so attached is the google drive link.

<https://drive.google.com/drive/folders/1pYNcOhoA64gNYRvFAVCkjE18YlEKYytY>

The data is of good quality, it is of right data type and in the right format.

# Data Cleaning:

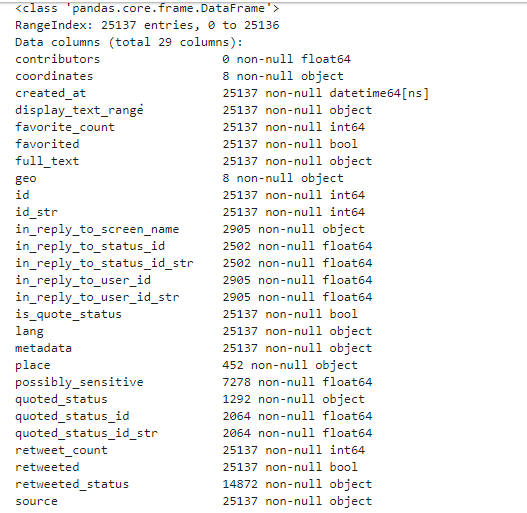
The data were of good quality and still contained some stop words and unnecessary characters.

## Removing the Stop words:

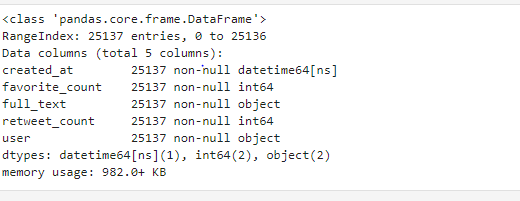
The stop words were treated by removing them with tweet\_clean and the unneccesarry columns were also seprated.



Before treatment:



After Treatment:



# Data Transformation:

The transformational activities were performed, by converting .json format into pandas dataframe to csv format.

# Building the Models:

## Model Selection:

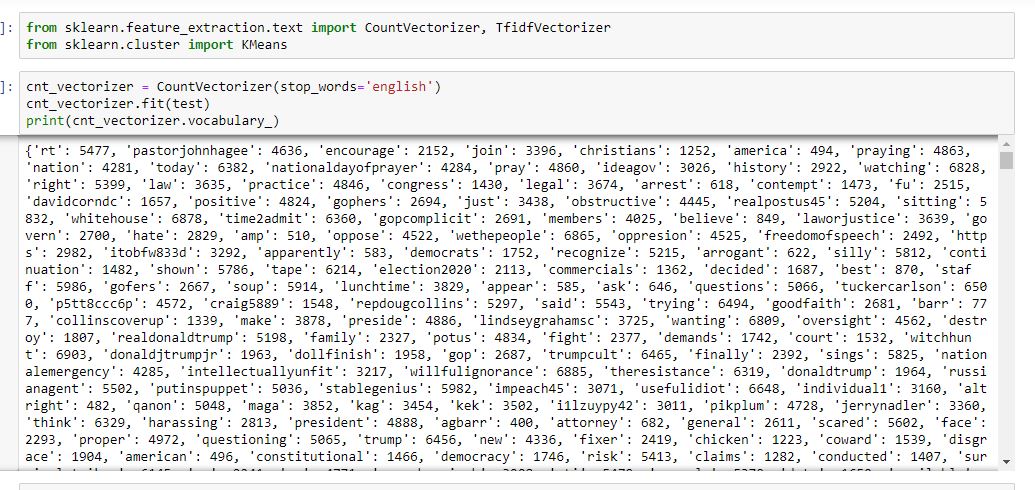
The models which are selected:

* K-means Clustering

## Model Execution of K-means:

For Congress:

Applying the Kmeans algorithm



Then the make cluster, which is represented using the bar graphs

