IST 652 - SCRIPTING FOR DATA ANALYSIS

FINAL PROJECT REPORT

SCHOOL OF INFORMATION STUDIES

Srinivas reddy pachika

spachika@syr.edu

SYRACUSE UNIVERSITY

DONALD TRUMP VS BERNIE SANDERS   
SENTIMENT ANALYSIS

**INDEX**

1. **OBJECTIVE**
2. **DATA DESCRIPTION**
3. **ANALYSIS QUESTIONS**
4. **CODE DESCRIPTION**

**OBJECTIVE**

The objective of this project is to Analyze the sentiment of Twitter users towards Donald Trump and Bernie Sanders by collecting data from twitter API and compared it with the results derived by performing analysis on raw data collected from “The pulse of the nation” survey.

**DATA DESCRIPTION**

Data for this project is mainly collected from two sources

* TWITTER API
* PULSEOFTHENATION.COM
* First data is mined from twitter Api by using #Donaldtrump and #Berniesanders
* A total of 1000 tweets are collected as first set of data, 500 tweets from each hashtag
* Second data which is downloaded from pulseofthention.com is in the csv format and has 801 rows and 24 columns

**ANALYSIS QUESTIONS**

I have mainly analyzed four data questions two of which are to explore the tweets collected and other two are focused in Analyzing the sentiment of Donald trump and Bernie Sanders.

1. What are the top frequent words used in tweets with #Donaldtrump and #Berniesanders?
2. What are the top 10 hashtags used in tweets?
3. How is the sentiment for Donald Trump and Bernie Sanders?
4. Is the sentiment Analyzed for Donald Trump from twitter in accordance with the results from analysis of Raw data collected from pulseofthenation.com?

**CODE DESCRIPTION**

**Step-1: Authentication**

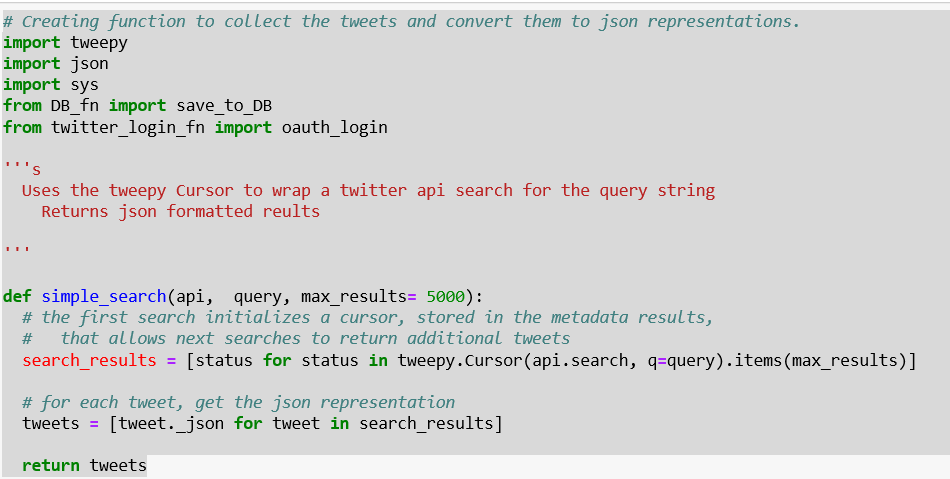
First, step here was to create authentication for twitter Api. I have used tweepy package

For both creating authentication and collecting tweets. As the below snippet of code shows I have created Authentication for twitter API using Authentication tokens from twitter App.

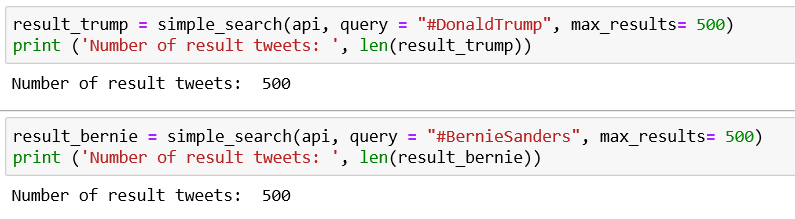


**Step – 2: Collecting Tweets from Twitter Api**

* For collecting the tweets from Twitter Api, I first created a function which takes arguments as query and maximum number of results and returns the collected tweets as Json objects.

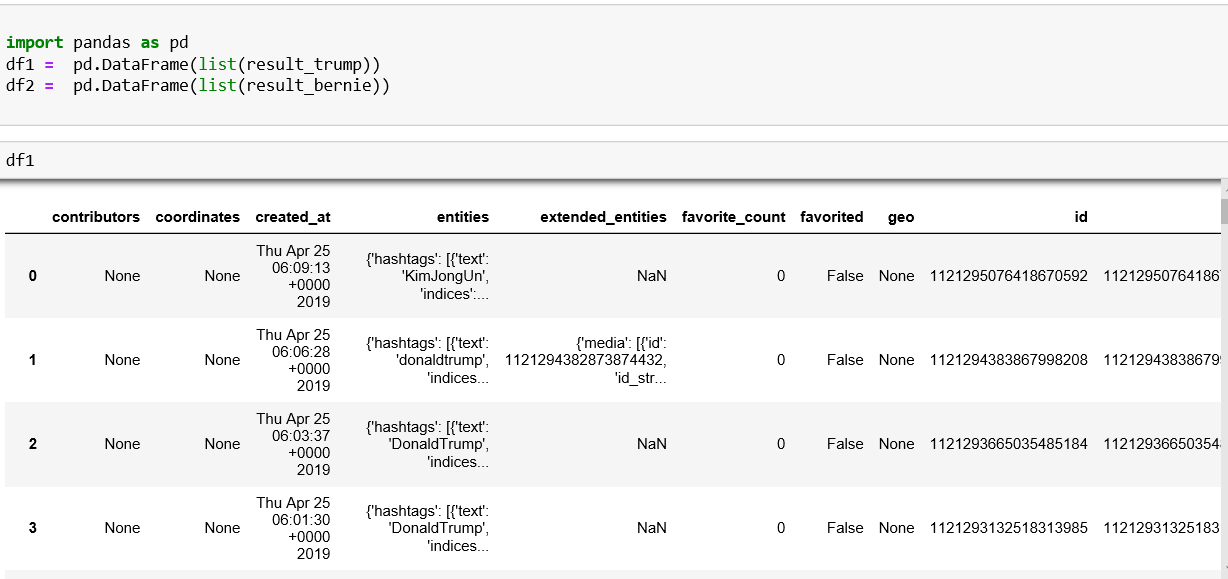


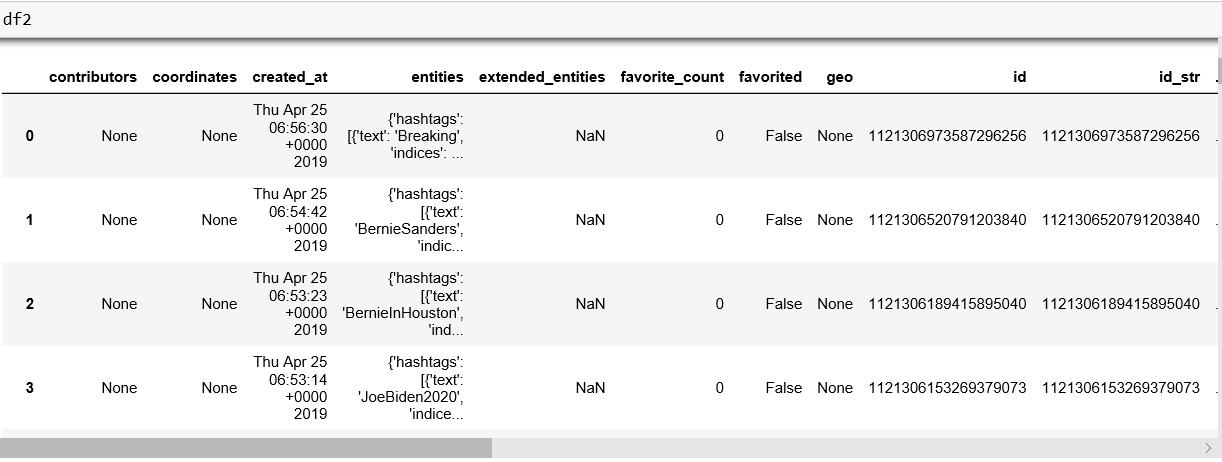
* Then I have used This function “simple\_search” to get tweets 500 tweets from each query using #donaldtrump and #berniesanders.



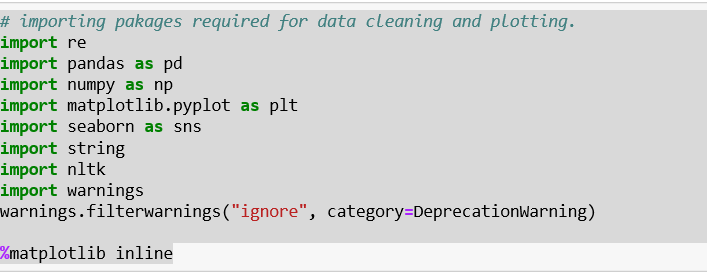
**Step – 3: Importing Tweets into Data Frame and Performing Analysis**

* I have imported collected tweets into a pandas data frame

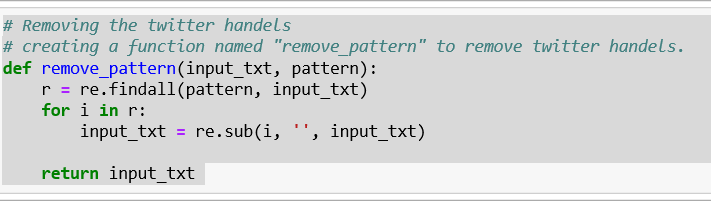




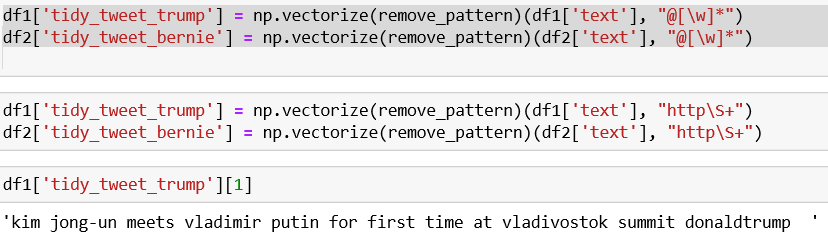
* Importing Required Packages



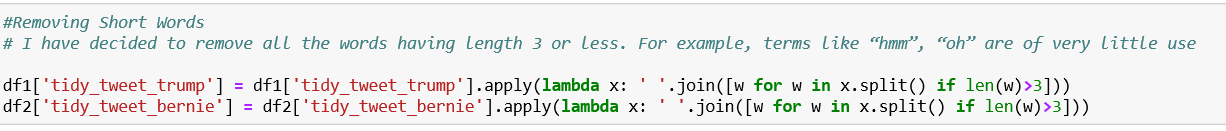
* Then I created a function called “remove\_patter” to clean the tweets which takes text and regular expressions pattern as arguments



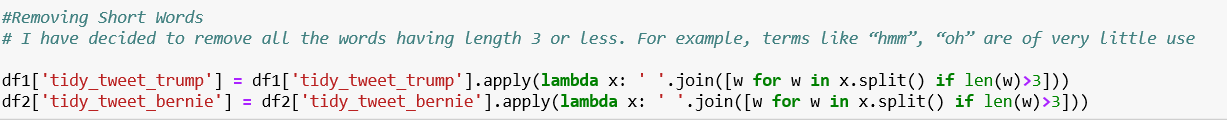
* Removing Twitter Handles and URL from tweets using “remove\_pattern” function we created



* Removing Punctuations, Numbers, and Special Characters
* punctuations, numbers and special characters do not help much. It is better to remove them from the text just as we removed the twitter handles.

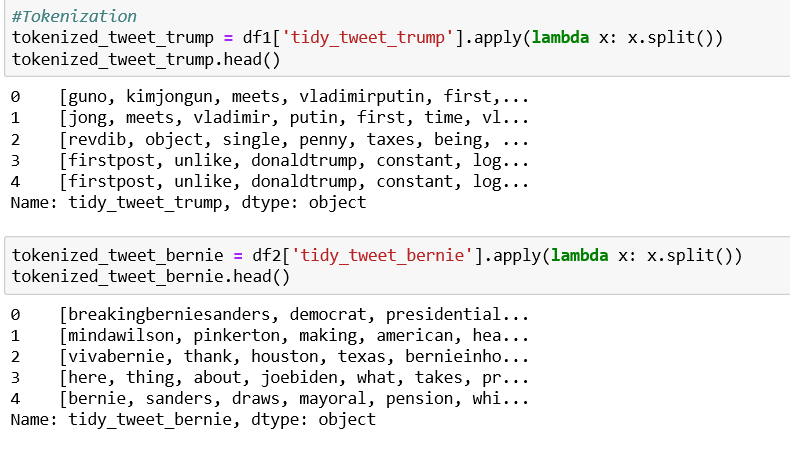


* Removing Short Words
* I have decided to remove all the words having length 3 or less. For example, terms like “hmm”, “oh” are of very little use and hence I got rid of them.



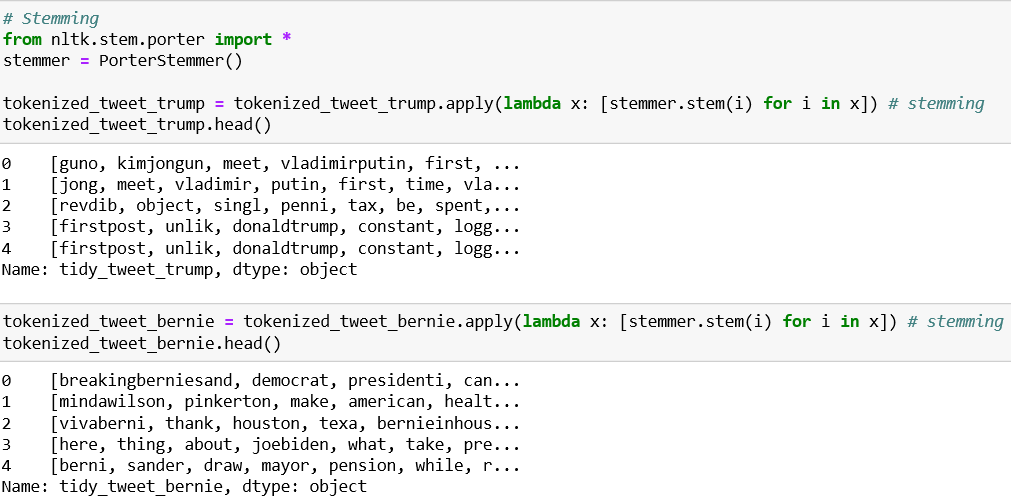
**Tokenization**

* Now we will tokenize all the cleaned tweets in our dataset. Tokens are individual terms or words, and tokenization is the process of splitting a string of text into tokens.

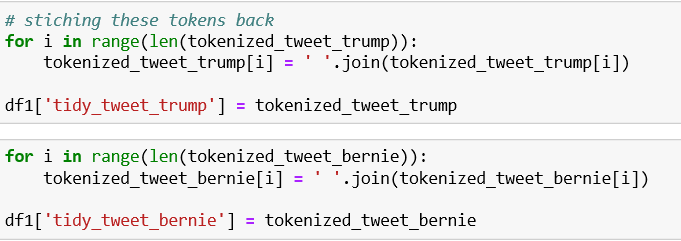


**Stemming**

* Stemming is a rule-based process of stripping the suffixes (“ing”, “ly”, “es”, “s” etc) from a word.



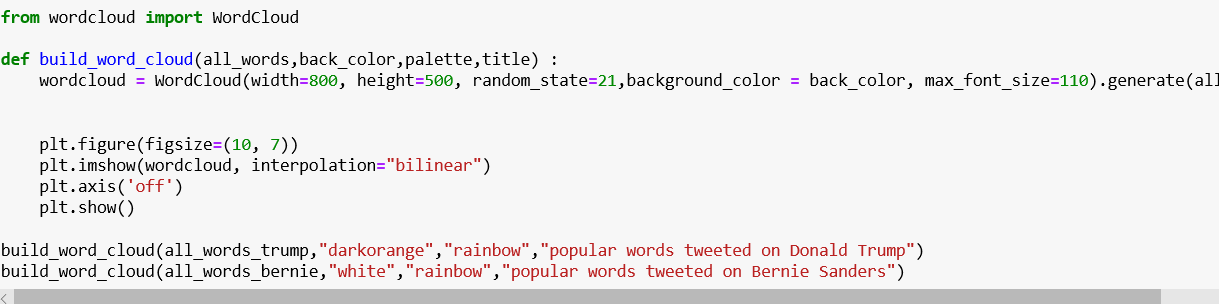
* Stitching back Tokens

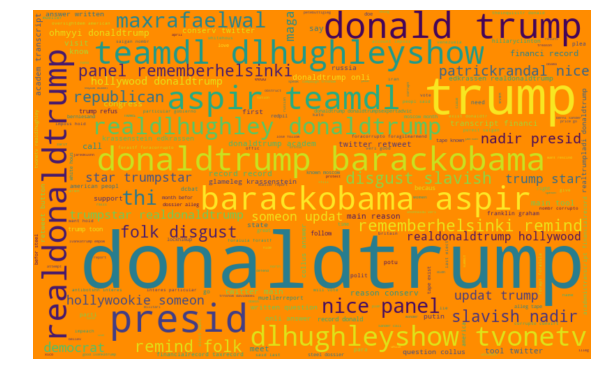


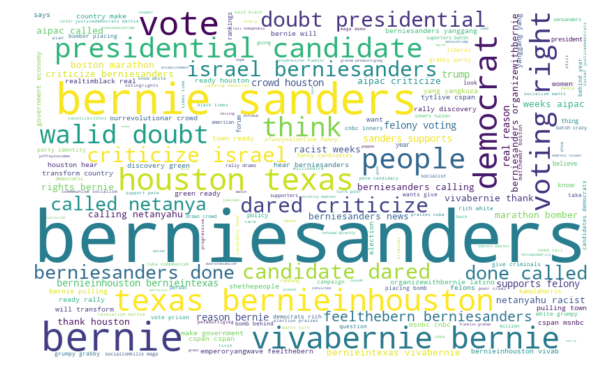
**Answering Question 1 (What are the top frequent words used in tweets with #Donaldtrump and #Berniesanders?)**

**Understanding the Frequent words used in the tweets: Word Cloud**

Now I want to see how well the given sentiments are distributed across the train dataset. One way to accomplish this task is by understanding the common words by plotting word clouds. A word cloud is a visualization wherein the most frequent words appear in large size and the less frequent words appear in smaller sizes.

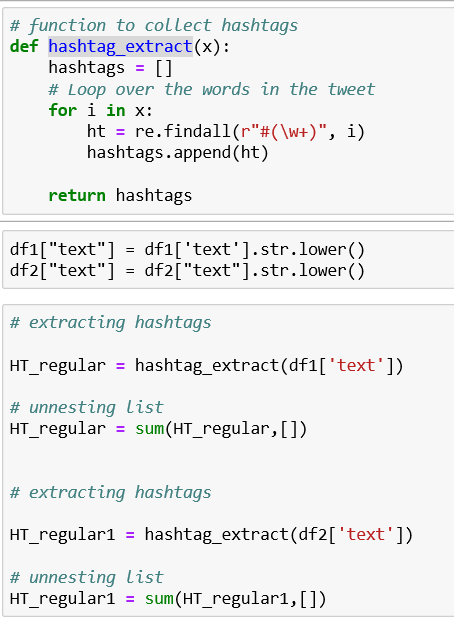




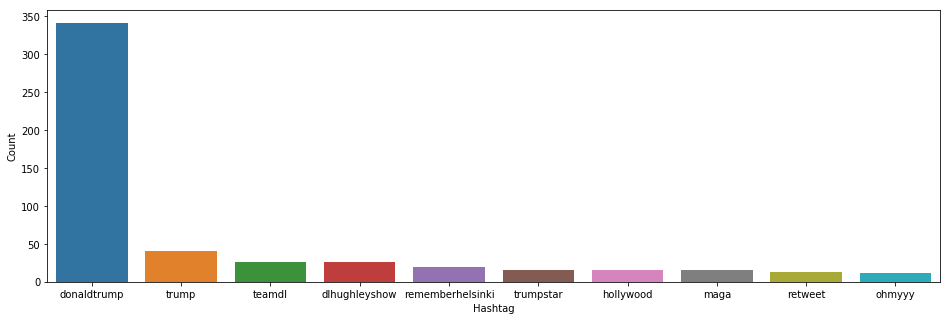


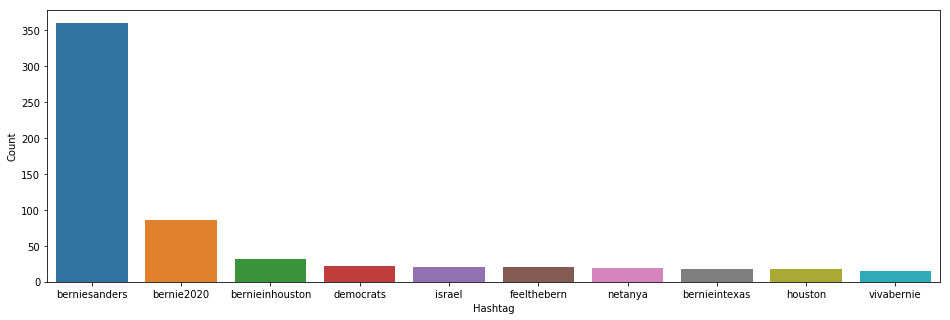
**Answering Question 2 (What are the top 10 hashtags used in tweets?)**

First, I have Created a function called “hashtag\_extract” which takes regular expression to search for all hashtags in the collected tweets, Then I have collected hashtags from collected tweets using this function and then using an inbuilt function “FreqDist” of NLTK I have retrieved top 10 hashtags used in tweets.



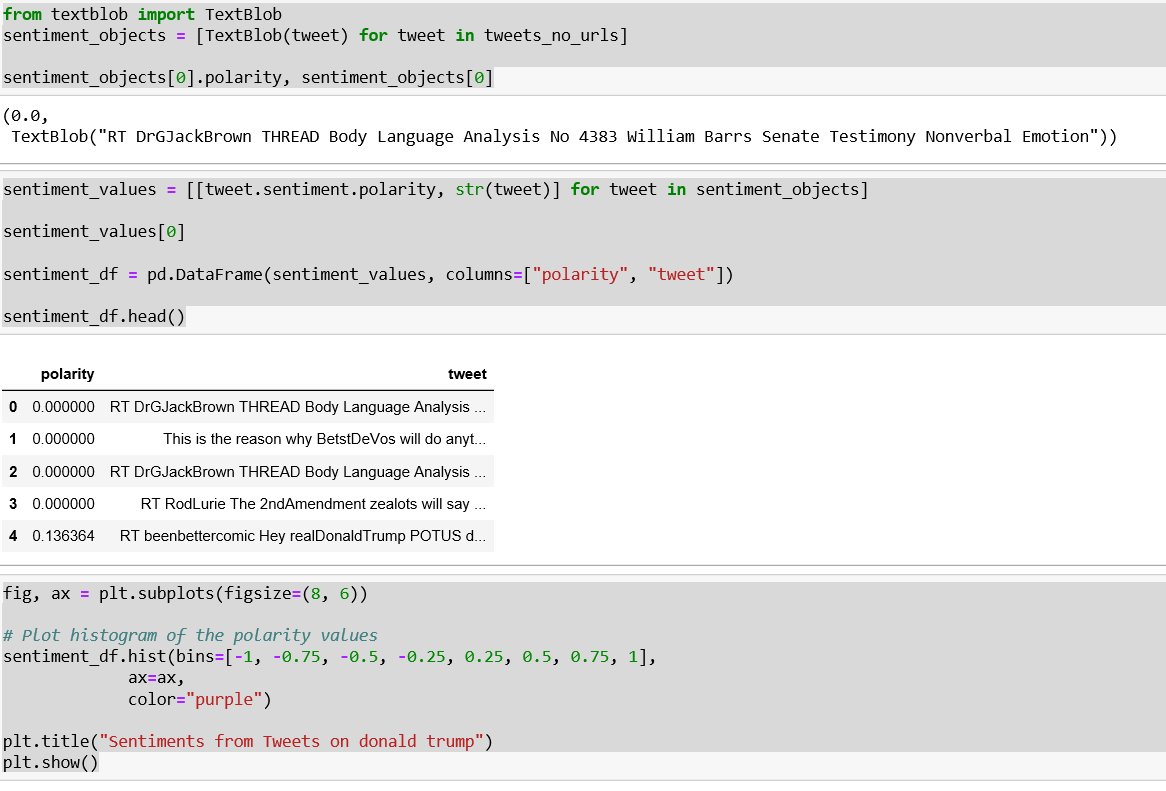


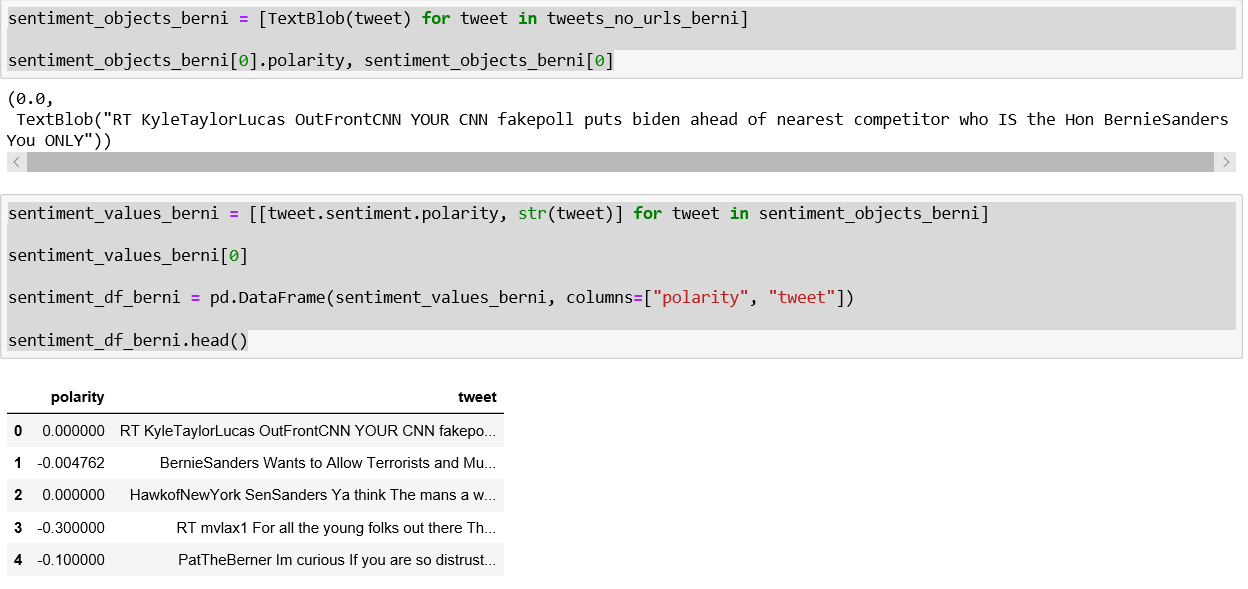


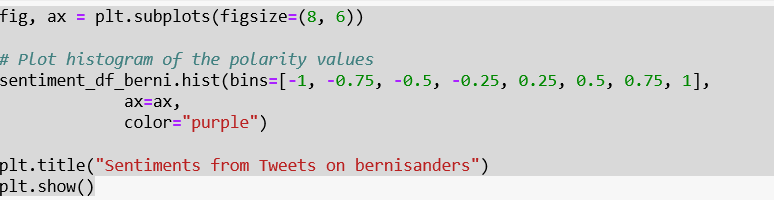


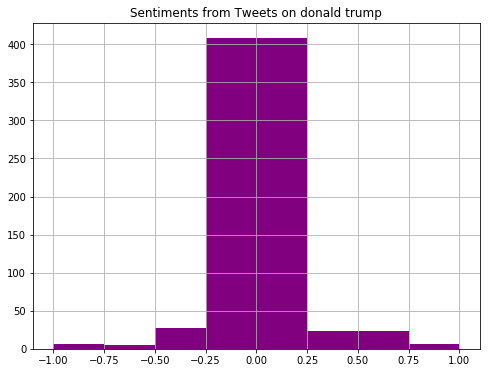
**Answering Question 3 (How is the sentiment for Donald Trump and Bernie Sanders?)**

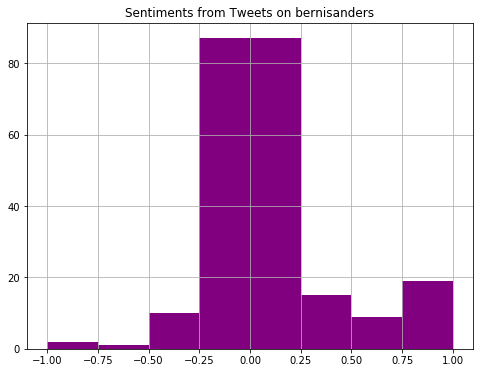
Then As I final step of Analysis I have used TextBlob package to score polarity for all the collected tweets and then imported them into a data frame and then plotted the values





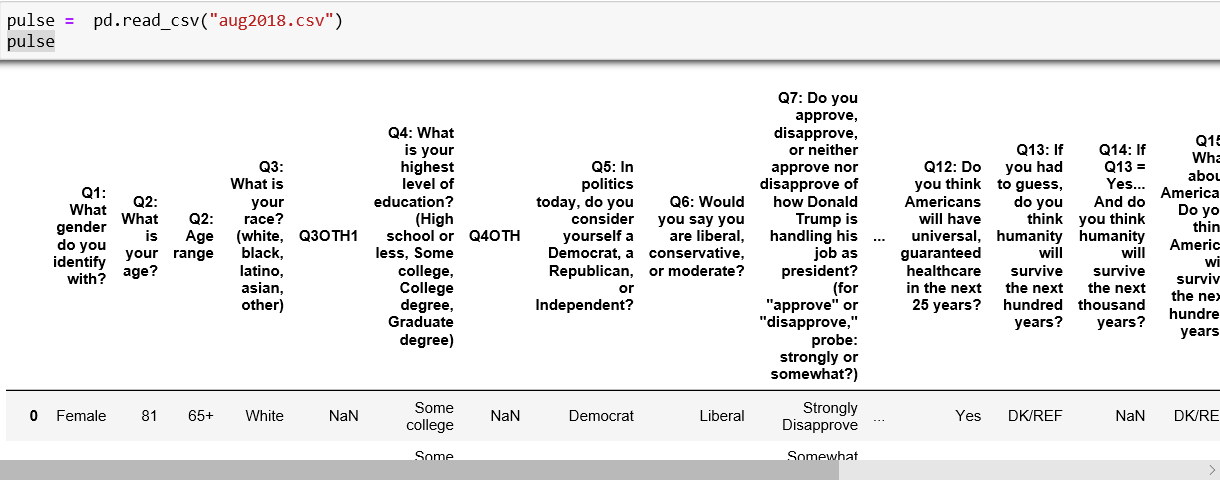




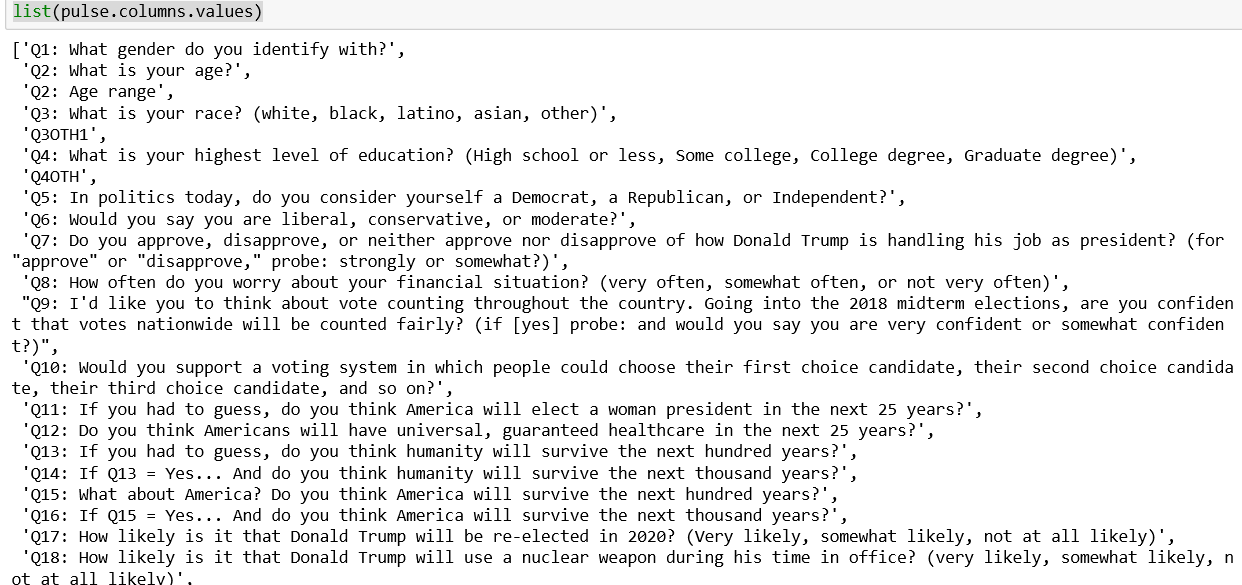


**Analysis on Raw Data Collected from Pulse of the nation.**

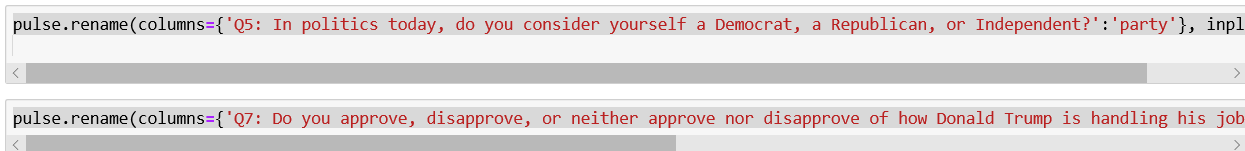
Imported data into padas data frame using read.csv function



Printing out column names



We are Interested in only column 5 and column 7. But these column names are very big. In fact, they are the questions asked in the survey. In the raw data the questions asked in survey are given as column names. So, we are now renaming only column 5 and column 7 names which are our desired fields for analysis.



Then as a final step I have used value\_counts function to