**AARTHI SHUNMUGAM – 10411286**

**REPORT: POLITICAL DEBATE TEXT ANLYSIS**

**Business Understanding:**

The aim of this assignment is to analyze two text files and compare them. This involves identifying the common words in both the files, to identify the words that are most popularly used in the political debate from both the files, their sentiment values (that is positivity, negativity and objectivity), also representing their word clouds as a figure.

**Data Understanding:**

Given two text files, which contains the data extracted from popular news websites- New York Times and Wall Street Journal. This data is about the Democratic Political debate that was held in Brooklyn on 4/14/16. Since the text didn’t contain any non- relevant information, not much data cleaning is required. Also, it contains all the relevant information required to process the data.

**Data Preparation:**

Though the non- relevant information from the data has been removed, the data should be cleaned (that is the special characters should be removed).

**Data Cleaning:**

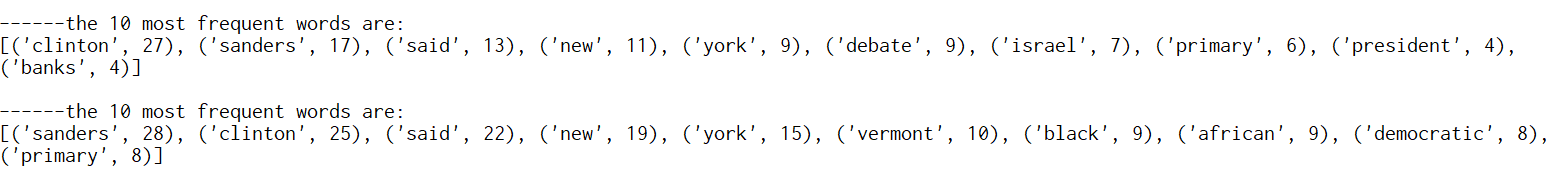
The data contains lines of text, which should be converted into list of words. This includes removing the end of line and creating a list of words from both the files.

The words in the created list from both the files are compared with the Stop-Words list and the stop-words are removed from both the list of words.

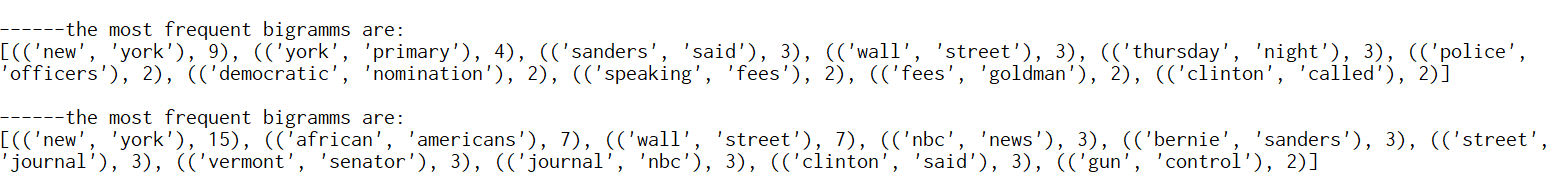
If there are repeated words in the two top 10 words lists, those words are also removed, and they are added to the stop-words list.

**Practical Results:**

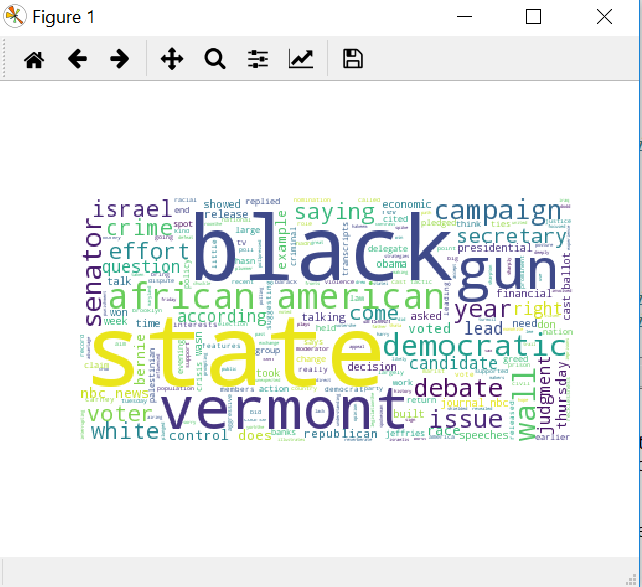
The 10 most frequent words identified from both the New York Times file and Wall Street Journal Files are:



The 10 most frequent bigrams (the words appearing together more than 2 times in the whole text of both the New York Times file and Wall Street Journal Files are:



The word cloud image for the NYT file is below:



**Conclusion:**

From the sentimental analysis it is clear that the both the files has some negativity in it, which means there was some negative arguments passed during the political debate. The sentiment score by running both the text in sentiment analyzer is as follows.

Sentiment Analysis of NYT file is -24.9 depicting overall a negative tone of the article. Sentiment Analysis of WSJ file is -11.4 depicting overall a negative tone of the article