**AARTHI SHUNMUGAM – 10411286**

**REPORT: Hillary\_Clinton Email Analysis**

**Business Understanding:**

The aim of this assignment is to explore Hillary Clinton emails, identify the most popular senders and to plot the visualization graphs for the top senders. This involves identifying the top occurring words in the column MetadataFrom of the file (which contains all the senders). Also, Word-Cloud is generated for the top occurring words (most frequent words) from the contents (RawText).

**Data Understanding:**

The given csv file contains the contents of emails sent to Hillary Clinton by different senders during the time of election. It contains many attributes such as SenderPersonId(contains the Id of the senders), MetadataFrom(contains the name of the senders), RawText which represents the contents of emails and other contents. So, there are totally 22 attributes in the data. The most common stop-words are removed and the data is cleaned for unwanted spaces and special characters.

**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:**

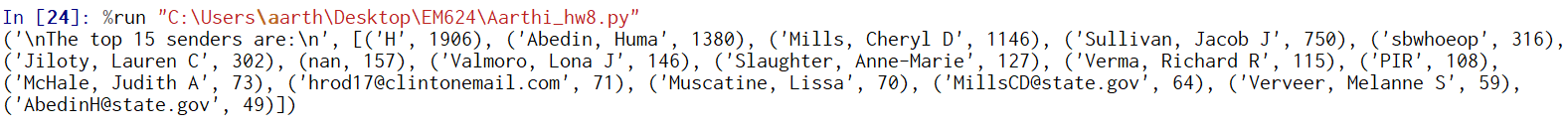
The data contains lines of text, which should be converted into list of words. This includes removing the end of line and creating a list of words from the file.

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

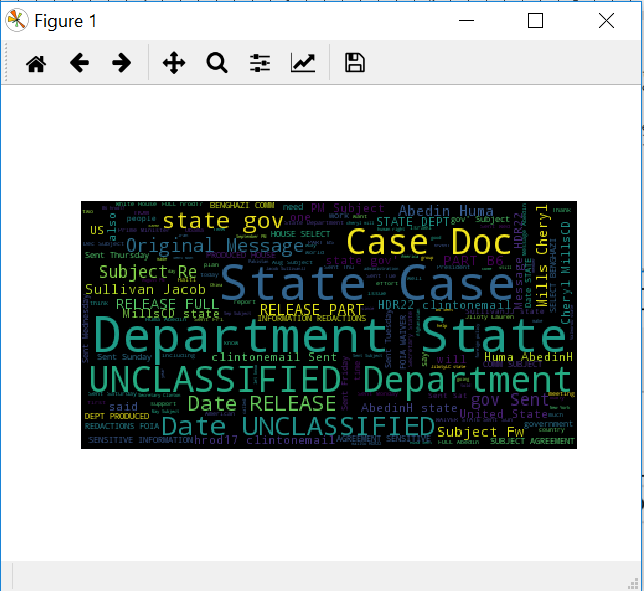
If there are repeated words in the word lists, those words are also removed, and the remaining words in the text are added to the Cleaned List.

**Practical Results:**

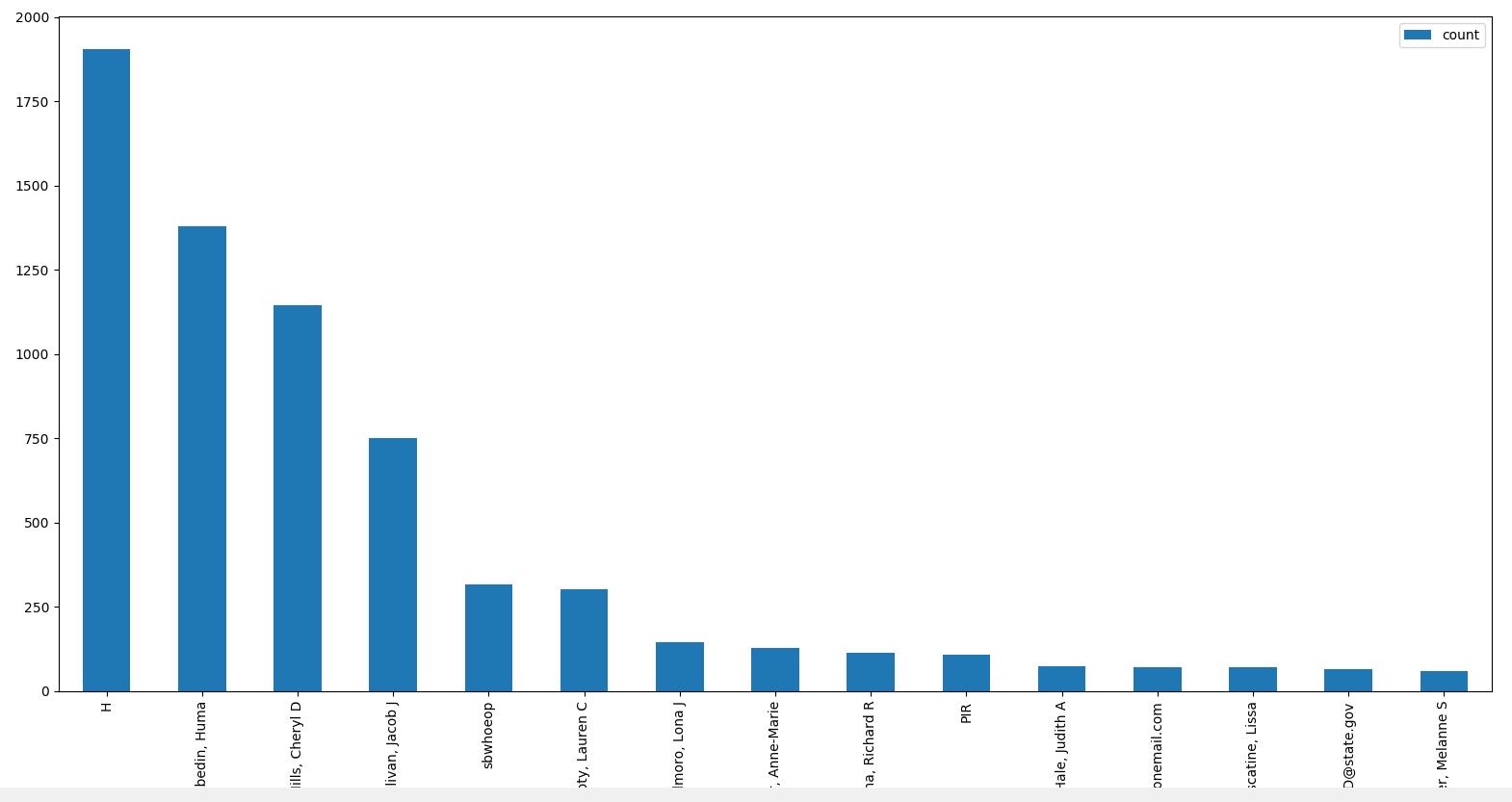
The top 15 senders (name of the top 15 persons, who has frequently sent emails to Hillary Clinton) are as follows:



The word cloud image for top words from H\_Clinton Emails file is obtained as below:



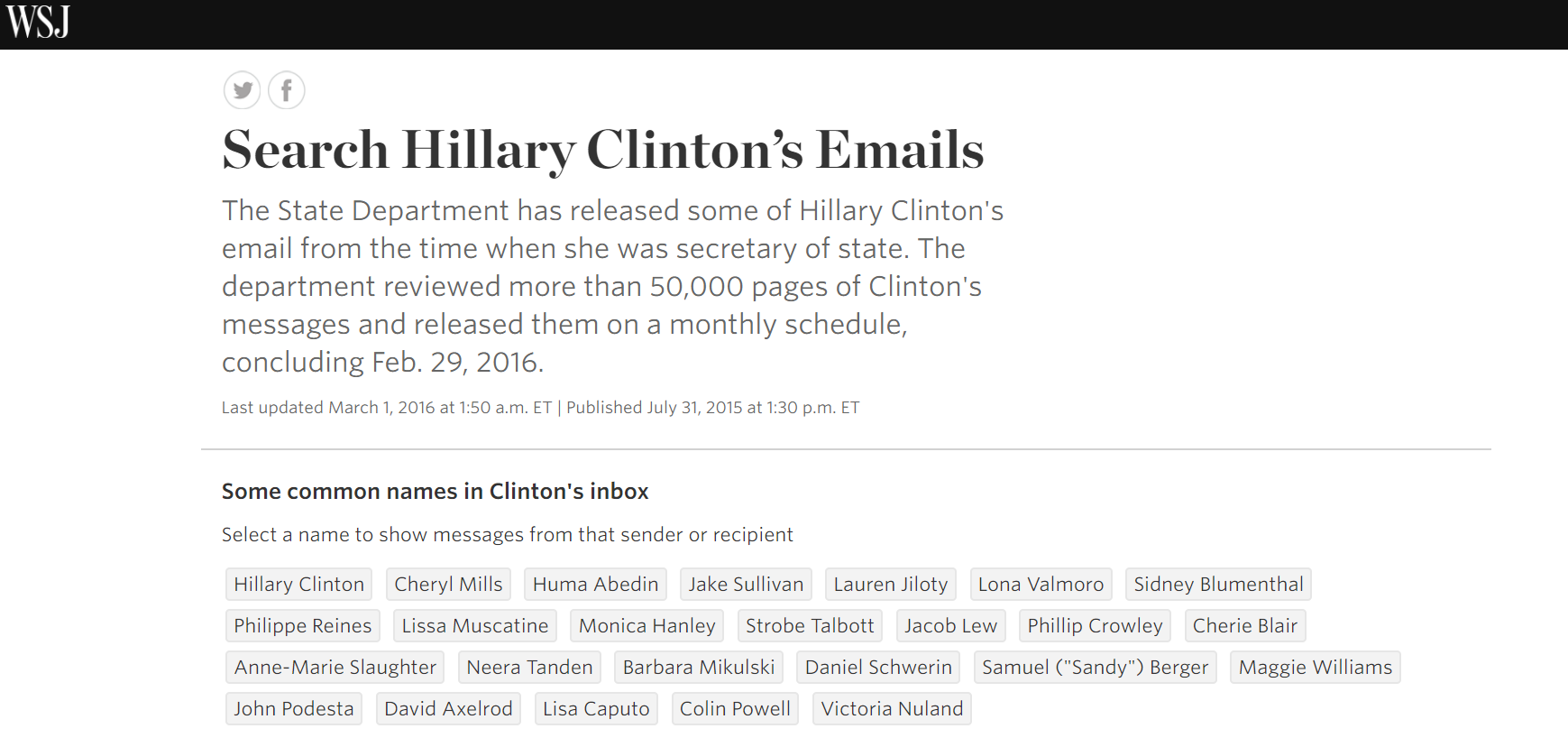
The visualization graph for the top senders is as follows:



**Conclusion and Findings:**

Using Google, I tried to find the most popular senders to Clinton’s email, and compare the names with my findings. I got the following names from Google that appeared to be the most popular senders.

Link: http://graphics.wsj.com/hillary-clinton-email-documents/



While comparing the findings from Google and my Visualization graph, it is clear that the most frequently occurring common names or the most popular senders are Cheryl Mills, Huma Abedin, Lauren Jiloty, Lona Valmoro, Jacob Lew, Anne-Marie Slaughter, Lisa Caputo.