ADA Project

Name: Daitong Li

UNI: dl2991

December 6, 2016

Cluster Analysis

The social network analysis in the previous section has discovered close relationship between Clinton and a group of her email receivers. The cluster analysis in this section focuses on classifying the receivers' emails in a meaningful way. It is useful to inspect the trends of the language Clinton has used when speaking with different people/groups of people. The clustering analysis is normally useful to categorize large sum of text into groups based on similarities. For exploratory purpose while analysing Clinton's emails, we would like to see if we can categorize based on the language used emails from different circles of people. Due to time limit, we only focuses on the top 10 accounts who have received most emails from Clinton.

Receivers

Using the emails datasets, there are 1906 number of people who have received emails from Hillary. We selected 10 receivers who have received highest numbers of emails from Clinton directly. These receivers are a mix of both work email addresses as well as personal addresses. For example Huma Mahmood Abedin, the vice chair of Hillary Clinton's 2016 presidential campaign, has received 331 emails through her work email address abedinh@state.gov and 31 through personal email. The top 10 email receivers are as below:

	Name	Email Type	Frequency	Word count (cleaned)
1	Huma Mahmood Abedin	Work	331	981
2	Cheryl D. Mills	Work	297	16145
3	Jacob Jeremiah Sullivan	Work	288	1696
4	Lauren Jiloty	Work	223	1993
5	Lona Valmoro	Work	129	8584
6	Philippe I. Reines	Personal	49	15114
7	Sidney Stone Blumenthal	Personal	47	2068
8	Cheryl D. Mills	Personal	35	1786
9	Monica R. Hanley	Work	33	13190
10	Huma Mahmood Abedin	Personal	31	6279

Other accounts which have received frequent emails from Clinton are Anne-Marie Slaughter, Richard Verma, Robert Russo, Lissa Muscatine (speech writer) etc. all under 30 emails.

If we break it into clusters, we would like to ask the questions below:

1. Are the people clustered into one group share common things?
2. Are the emails sent to private accounts separated from the emails sent to work account?

Here are a list of brief background of the 8 people involved:

1. Huma Mahmood Abedin (born July 28, 1976) is an American political staffer who was vice chair of Hillary Clinton's 2016 campaign for President of the United States. Prior to that, Abedin was deputy chief of staff to U.S. Secretary of State Hillary Clinton from 2009 to 2013.

2. Jacob Jeremiah Sullivan (born November 28, 1976) is an American policymaker and a senior policy advisor to Hillary Clinton's 2016 election campaign, with expertise in foreign policy. He was widely rumored to be a front-runner for the position of U.S. National Security Advisor under a Hillary Clinton administration, before she lost to Donald Trump in the 2016 Presidential Election

3. Cheryl D. Mills (born 1965[1][2]) is an American lawyer and corporate executive. She first came into public prominence while serving as deputy White House Counsel for President Bill Clinton, whom she defended during his 1999 impeachment trial.

4. Sidney Stone Blumenthal (born November 6, 1948) is an American journalist, activist, writer, and former political aide. He is a former aide to President Bill Clinton; a long-time confidant[1] to Hillary Clinton, formerly employed by the Clinton Foundation;[2] and a journalist, especially on American politics and foreign policy.

5. Philippe I. Reines (born November 25, 1969) is an American political consultant. He joined the Department of State as a Senior Advisor to Hillary Clinton when she became United States Secretary of State in January 2009, and in 2010 was promoted to the position of Deputy Assistant Secretary of State for Strategic Communications

6. Lauren Jiloty: Special Assistant to Secretary Hillary Rodham Clinton 2009-2011

7. Lona Valmoro is a former Special Assistant to Secretary of State Hillary Rodham Clinton. Lona served as senior advisor to Clinton during her two terms in the US Senate and played key roles in Clinton's two campaigns for the US Senate and her 2008 bid for the Democratic presidential nomination.

8. Monica Hanley, Clinton's ”confidential assistant” at the state department

Emails

First we cleaned and transformed the text by

1. Convert into lower case.

2. Remove problematic symbols, punctuations and numbers.

3. Remove stop words including 'date', 'can', 'we', 'will' etc.

4. Remove words that are typical in the emails including 'message','sent','original' etc and receivers' names.

    - We will have cleaned text like "... authorities theyve presenting president nightmare scenarios panetta persuaded renditions tool worth keeping rendition program began carefully monitored form clinton administration bush years transformed john radsan former cia lawyer called abomination many seven detainees misidentified abducted mistake many suspects alleged hideously tortured foreign governments panetta told worst rendition ..."

5. Convert all documents into a document-term matrix

    - The matrix has 10 rows representing each email receiver's account; and 20792 columns representing unique words that have appeared in all the email text
    - In each row, if a certain word appears N times the corresponding column would have 'N' , '0' if not.
    - The sparsity of the matrix is 74%, the non-/sparse entries are 53249/155201.

6. Compute Euclidean distance between each documents.

The distance between two documents is defined in high dimension, in our case is 20792 dimensional space. To illustrate for instance, if we have two documents Doc1 (house,libya, document) and Doc 2 (house, information, release, document). In the document-term matrix, the two documents is represented by word frequency eg. we can have Doc1=(1, 0, 1, 1), Doc2=(1, 1, 2, 1), the distance between Doc1 and Doc 2 will hence be $\sqrt{(1-1)^2+(0-1)^2+(1-2)^2+(1-1)^2}=\sqrt{2}$. A simple word cloud for all 10 documents can be seen below:

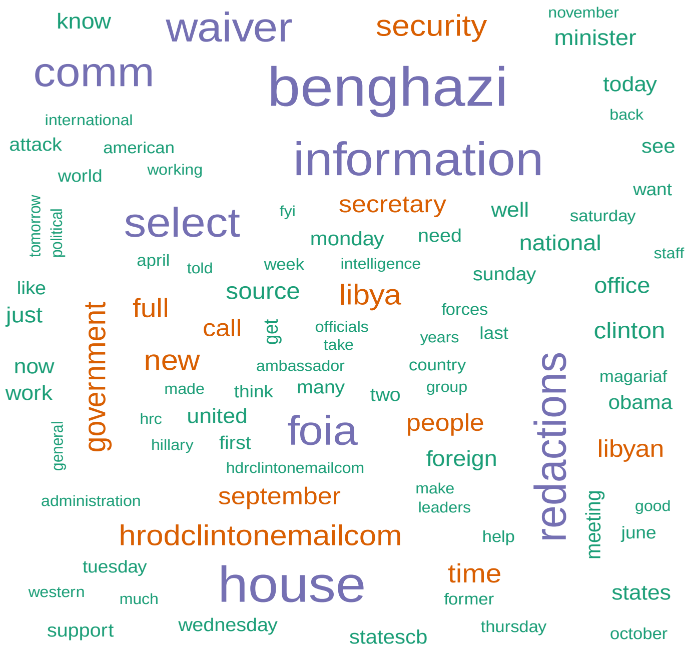

Figure 1

Clustering Results

After obtaining the matrix containing the distances between each documents, we tried two methods: hierarchical clustering and K-means.

For hierarchical clustering, we have used ward.D method or the Ward's minimum variance method. Below is a cluster dendrogram of the top 10 email accounts. The different colours represent grouping.

The height represents the distance between each document. For example, at distance of 800 we can separate the group in pink and others. The calculation of distance is explained in the section above.

We can also use K-means to determine the clusters. By measuring the within group sum of square, we can observe from the plot below that K=2 or K=3 might be sufficient number of clusters for this dataset. The cluster plots with K=2 and K=3 can be shown below. The x and y axis are represented by the first and second component from the PCA analysis. As we can see that the first component explains more than 95% of the variance in the text, allowing the clusters to be well separated.

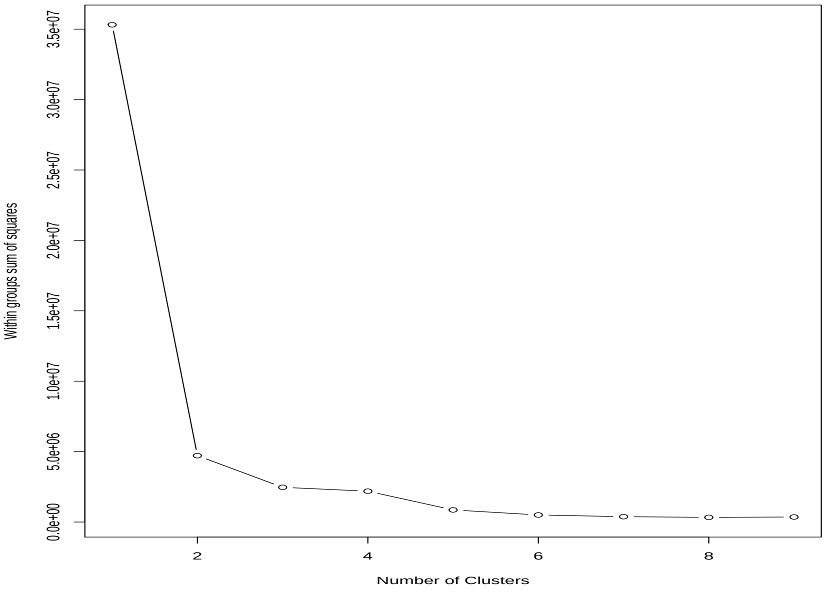

Figure 2

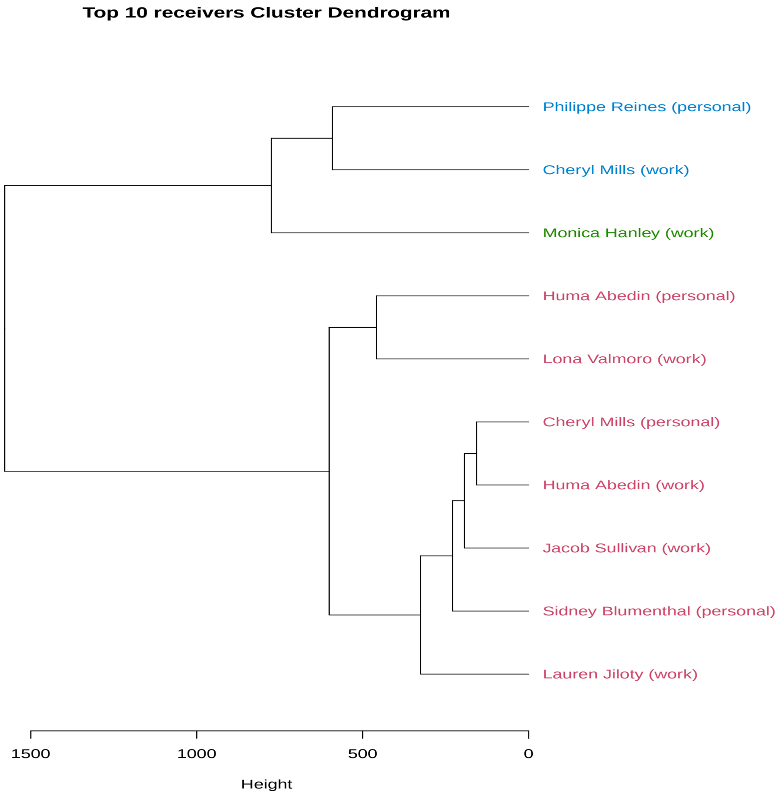

Figure 3

Limitations

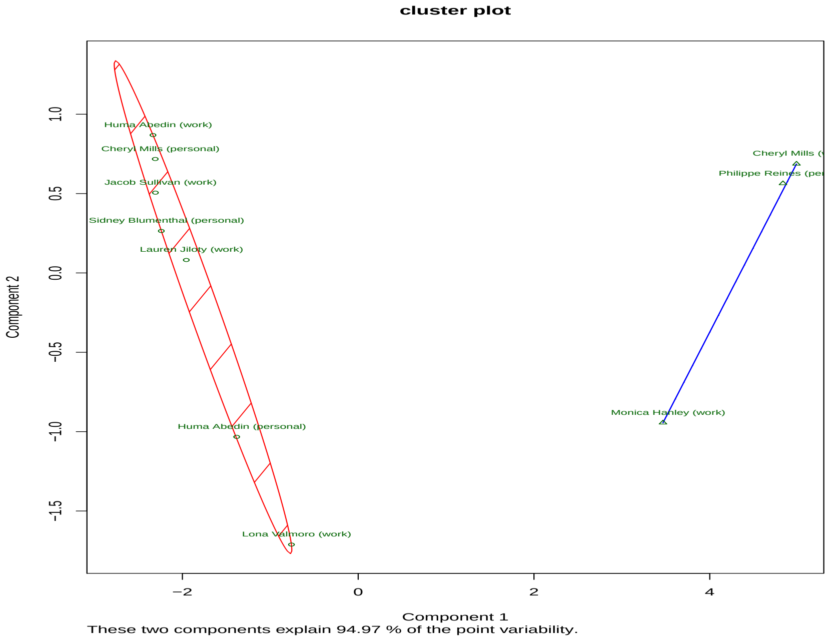

Figure 4

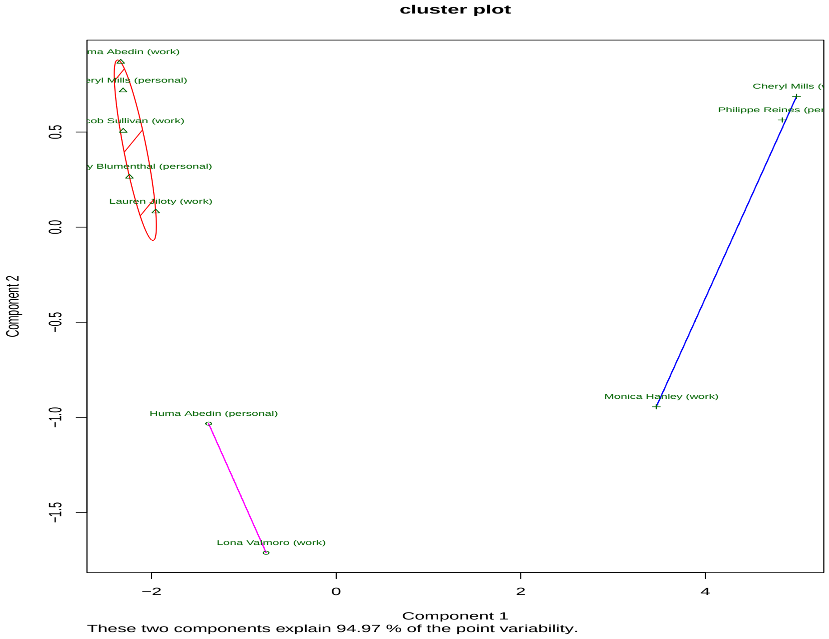

Figure 5