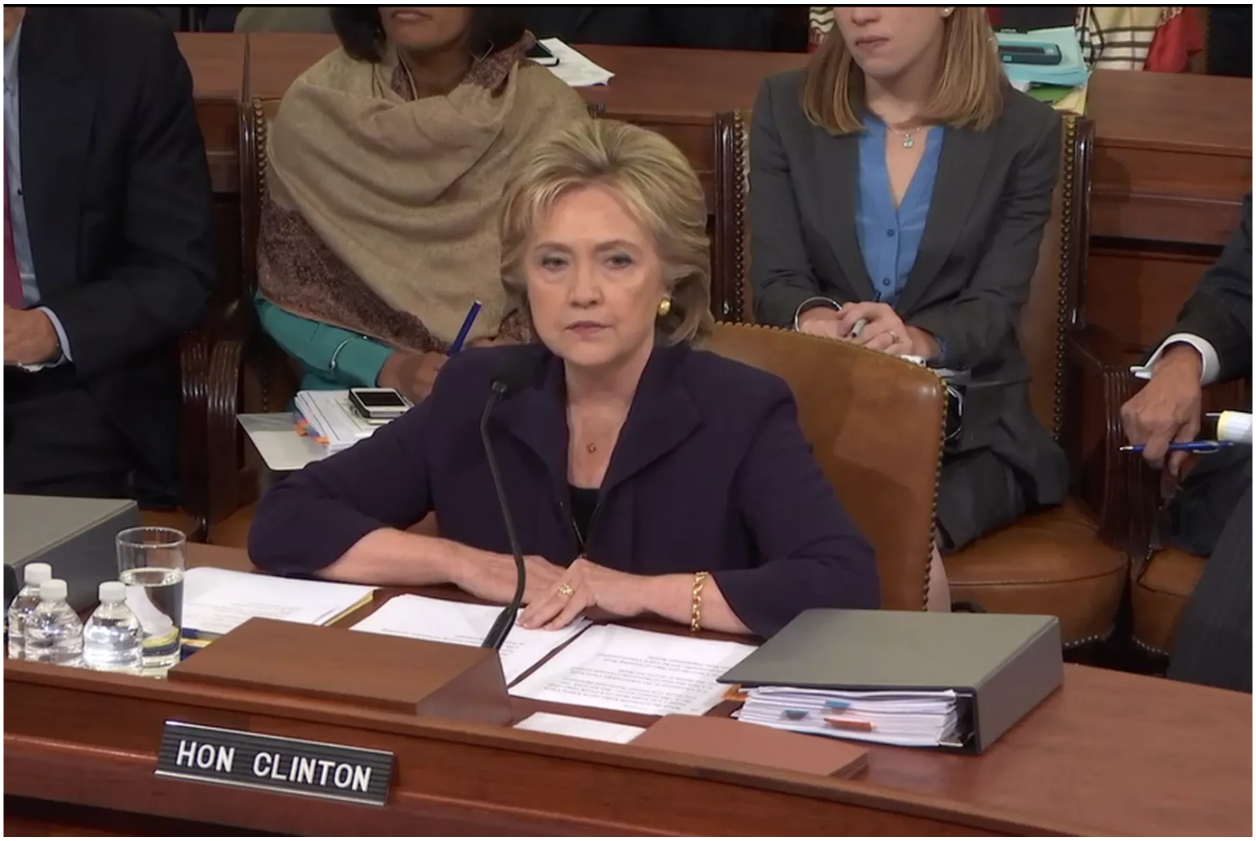
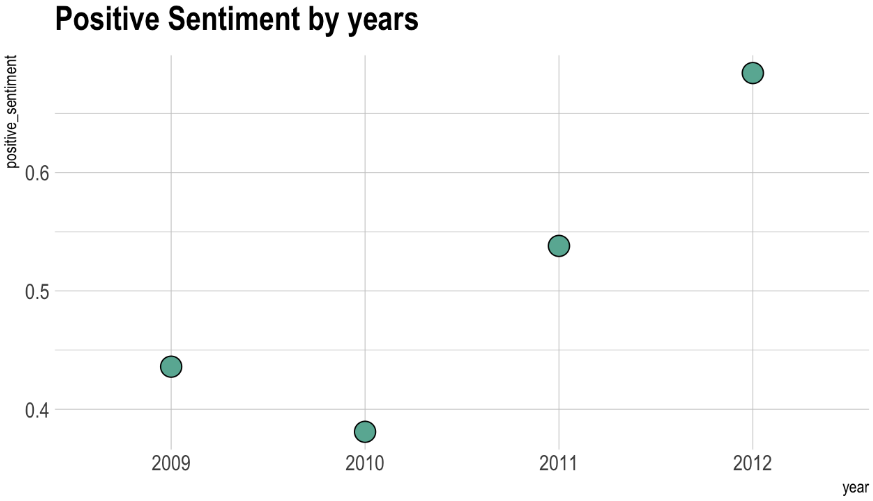
Challenge 2: Text Analysis of Hillary Clinton's Emails

By: Samuel Jin, Herbert Suarez, Genessa Marquez & James Ramos



Major Finding

Our group traces the sentiment of Hillary Clinton's emails from 2009 to 2012. By using K-means clustering and sentiment analysis, we find that the positive sentiment value for Hillary Clinton's emails decreased from 2009 to 2010 and increased from 2010-2012 in the midst of Haiti's devastating earthquake and the Benghazi scandal. Our study aims to analyze this time period and trace its textual context.

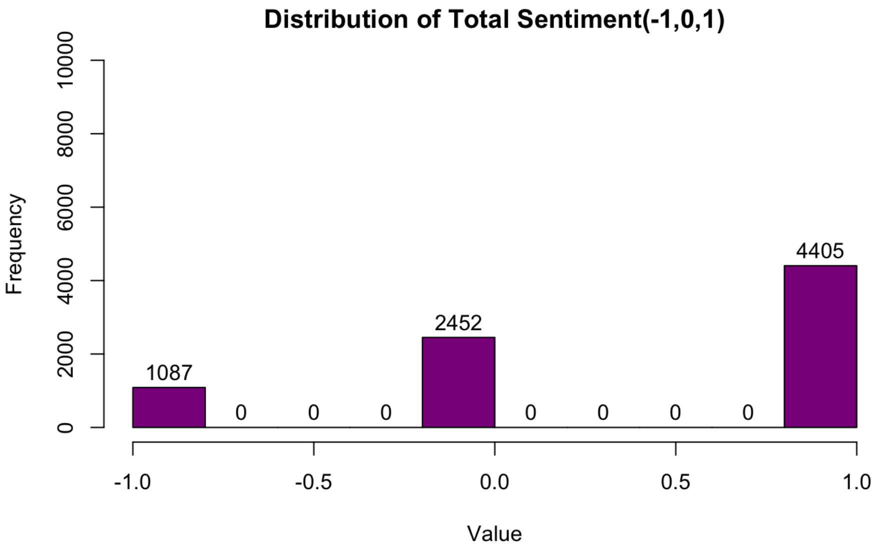


Methodology

We begin with a general sentiment analysis over all the emails. For a more straightforward analysis, we only extract the unigrams. In addition, our group creates our own dictionary to remove unnecessary words that are convaluting the quality of each cluster, while also using the positive and negative dictionaries. When our data is effectively cleaned for analysis, we fit several K-values such as 6, 15, and 20. By adjusting the K-values, we find different clusters and are able to see which K-value is the most optimal for our K-means clustering model. When using smaller values, we find that the topics covered are too broad and difficult to extract meaningful information. On the other hand, the larger K-values create a greater amount of clusters which compromise the distinctiveness of the clustered words. Our group finds the optimal k-value to be 20 because it strikes a balance between the exclusivity and vagueness of the K-means clustering.

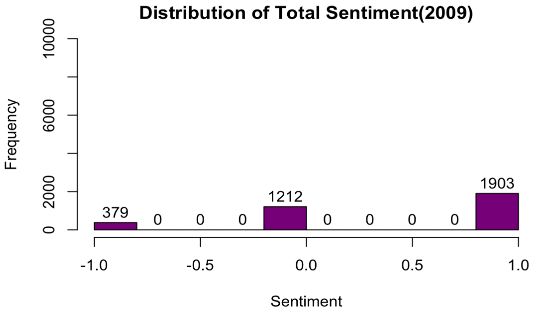
Sentiment Analysis by Year

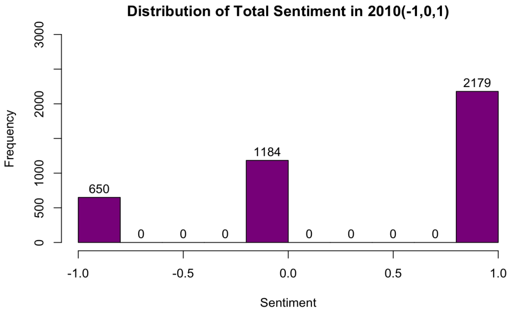
Using the cleaned unigrams and the positive and negative word dictionaries, our group finds that the majority of the emails have a positive sentiment. For further analysis, we examine the change of sentiment value each year and the most distinct and frequent words. We decided to not include the 2008 and 2014 emails because of the little amount of observations to analyze. Overall, we see that the emails are 55% positive sentiment words, 30% neutral and 14% negative.

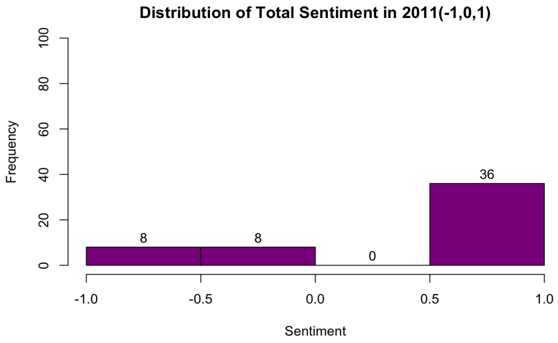


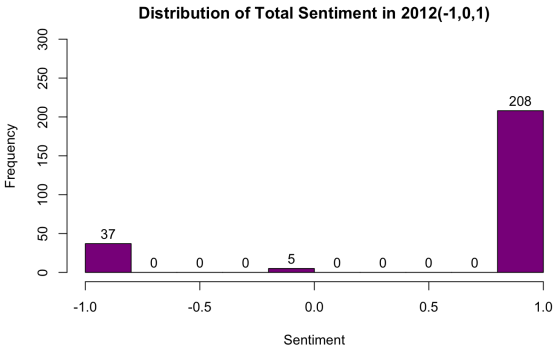
This discovery led us to subset the data into years, so that we could track the potential change in sentiment, as seen below:

Breakdown by Year









Clustering by Year

Given our discussion of sentiment analysis over time, we decided to conduct a cluster analysis to provide a better understanding of the topics of interest. Continuing on with the year as our level of analysis, we then clustered the emails by year such that every year had K number of clusters. Subsetting the data by year allows us to evaluate each set of clusters in order to give us a better idea of the content of the emails sent by Hillary Clinton. It should be noted that while our original cluster analysis had K set to 20, the following clusters were created with K set to 3. The purpose of setting K to a smaller number is to take into account the significant decrease in the number of observations we are left with when the matrix is filtered by year. Once we have the clusters, the functions five_frequent and five_distinctive provide the top five most frequent and distinctive words in each cluster. However, given the scope of this assignment we will be focusing on the top five most distinctive words for our analysis.

In 2009, the top 5 most distinctive words each cluster include:

```
[1,] "unclassified" "case"      "abedin"  "department" "abedinh"
[2,] "cheryl"       "mills"     "millscd" "fyi"        "cdm"
[3,] "his"          "president" "our"     "us"         "clinton"
```

When looking at the top words for the clusters in 2009, we see that there are keywords which give us insight on the subjects of conversation in her emails. More specifically, there are conversations about **Abedin**, the American political staffer who served as the vice chair to Hillary Clinton's 2016 campaign for President[1]. **Cheryl Mills** is another subject of interest who served as Senior Advisor and Counsel for Clinton's 2008 campaign. In her capacity as Counselor, she was a principal officer who served the Secretary as a special adviser on major foreign policy challenges[2]. Further, we ascertain that the words **His/President** are referring to Obama, following his 2008 presidential victory while "**We/Us**" talks about unity after a presidential campaign.

In 2010, the top 5 most distinctive words each cluster include:

```
     [,1]     [,2]           [,3]          [,4]        [,5]     
[1,] "cheryl" "mills"        "millscd"    "haiti"     "january"
[2,] "case"   "unclassified" "department" "gov"       "abedin" 
[3,] "his"    "news"         "new"        "president" "its"    
```

The subjects of conversation in 2010 appear to be similar since they include Cheryl Mills and talk about the president. There are new words that are introduced such as **Haiti/January/News.** It is revealed that Clinton's focus shifted in January of 2010, when a massive earthquake hit Haiti, resulting in an estimated 250,000 deaths. This catastrophic event left more than 5 million people displaced.[3] Given the magnitude of this advent, it makes sense that these words are found to be most distinctive in the 2010 clusters.

[1] Huma Abedin. (n.d.). Retrieved March 12, 2021, from https://ballotpedia.org/Huma_Abedin.

[2] Cheryl Mills (CLINTON ADVISOR). (n.d.). Retrieved March 12, 2021, from https://ballotpedia.org/Cheryl_Mills_(Clinton_advisor).

[3] Reid, K. (2020, February 27). 2010 Haiti Earthquake: Facts, FAQs, and how to help. Retrieved March 12, 2021, from https://www.worldvision.org/disaster-relief-news-stories/2010-haiti-earthquake-facts.

In 2011, the top 5 most distinctive words each cluster include:

```
    [,1]       [,2]      [,3]     [,4]      [,5]    
[1,] "tnc"      "stevens" "update" "tripoli" "mccain"
[2,] "sullivan" "jacob"   "march"  "X02"     "burns" 
[3,] "qaddafi"  "these"   "libya"  "sources" "libyan"
```

The most distinctive words from 2011 focus more on foreign policy and the conflict in Libya. In the first cluster we see mention of the word “**tnc”** which refers to the Transitional National Council of Libya, which was the “de facto government of Libya” around the time period “during and after” the state's civil war.[4] Moreover, “**tripoli**” appears in the same cluster, which is relevant to 2011 due to the battle of Tripoli in Libya. Interestingly, **“mccain”** appears in the same cluster. We can assume that the term refers to John McCain, who was the U.S. Republican senator of Arizona. McCain was an especially vocal supporter of the 2011 military intervention in Libya. The second cluster’s keywords included “**sullivan**”, referring to Laura Sullivan who is a correspondent and investigative reporter for National Public Radio.[5] Her relevance in the cluster is likely due to her work covering issues related to Tripoli. Cluster two’s words, including “**march**”, “**burns**” and “**jacob**” broadly refers to the issues regarding the uprising in Syria. March was when issues in Tripoli began to surface more and John F. Burns was one of the first people to help shed light on the situation. The mention of **“jacob”** refers to Jacob Zuma, the South African president at the time. Zuma tried to negotiate with Gaddafi early on in the conflict for a ceasefire on the rebels, but they did not come to an agreement[6]. The last cluster is most

[4] National transitional council – Libya. (n.d.). Retrieved March 12, 2021, from http://ntclibya.org/.

[5] Laura Sullivan. (n.d.). Retrieved March 12, 2021, from https://www.npr.org/people/4624985/laura-sullivan.

[6] Burns, John F. “Qaddafi and Zuma Meet but Reach No Agreement.” *The New York Times*, The New York Times, 30 May 2011, www.nytimes.com/2011/05/31/world/africa/31libya.html.

likely talking about the conflict that was earlier explained due to “**qaddafi**” and his regime being the main issue of the Battle of Tripoli in “**libya**.”

In 2012, the top 5 most distinctive words each cluster include:

```
[1,] "source"    "magariaf"   "libyan" "keib" "his"
[2,] "september" "department" "cheryl" "r"    "redactions"
[3,] "our"       "obama"      "attack" "she"  "what"
```

The emails in 2012 continue to focus on the growing tensions in Libya and foreign policy incidents.**“magariaf”**—a Libyan politician who dedicated himself to overthrowing the Libyan Arab Jamahiriya with violence—became a distinct word in this cluster. we can assume there has been an international conflict and subsequent US military intervention.[7] Another prominent figure mentioned is **“keib”,** the libyan prime minister during this time. Upon further research, we find these words to be related to the infamous Benghazi Scandal which explains **“source.” “September”** was the date that this insurgency took place.[8]

Conclusion

In this project, our group conducts exploratory text analysis to better understand the sentiment of Hillary Clinton’s released emails during her time as Secretary of State. By conducting sentiment analysis, we are able to compute the average sentiment values of her emails by year. We find that during 2009-2010, her emails consisted of connections with old campaign staff, her actions taken

[7] Zargoun, Taha. “Fighters Bulldoze Sufi Mosque in Central Tripoli.” *Reuters*, Thomson Reuters, 25 Aug. 2012, www.reuters.com/article/us-libya-islamists/fighters-bulldoze-sufi-mosque-in-central-tripoli-idUSBRE87O08Y20120825.

[8] “U.S. Vows to Hunt down Perpetrators of Benghazi Attack.” *CNN*, Cable News Network, 13 Sept. 2012, www.cnn.com/2012/09/12/world/africa/libya-us-ambassador-killed.

during the aftermath of President Obama's victory, and her response to a devastating earthquake in Haiti. After 2010, much of Hillary Clinton's emails have dealt with foreign policy, specifically in Libya. This was during the same time the infamous Benghazi scandal had emerged.