Final Project Report:
Hillary Clinton Email - Topic Modeling

Group Members:Dongheon Lee, Michael Lenart, Yifeng Song, Nicholas Venuti

Background

Problem Statement

According to Kaggle, "[t]hroughout 2015, Hillary Clinton has been embroiled in controversy over the use of personal email accounts on non-government servers during her time as the United States Secretary of State." The question has been raised whether this practice violated any laws, specially the Federal Records Act which requires that federal employees retain copies of all communication. A Clinton spokesman claims that no laws were violated because when she emailed colleagues about official government business, she used their official government email accounts. We wanted to use topic modeling along with k-means clustering to examine whether Secretary Clinton's claim that the Federal Records Act can be substantiated by the emails.

It is important to know if a high-ranking government official violated federal law, especially one running for President. Since Secretary Clinton is running for President, it is of interest to the US electorate whether she violated any laws and whether she has been truthful and is a person who can be trusted. Also, if Secretary Clinton's claim is accurate, it would be important for her and her campaign to use this information to clear her name and restore trust.

Status of Problem

There have been reviews of the emails using topic modeling, but we did not find any examination of the emails using topic modeling to evaluate the claim about how Secretary Clinton handled emails containing official government business.

Objectives

Our objective was to use a combination of topic modeling and clustering to examine whether Secretary Clinton's claim that all of her emails containing State Department business

included a State Department colleagues official email, which should include state.gov in the email address.

Evaluation of Objective

In order to assess the success of our project, we two team members manually scored emails that our model showed should contain topics from State Department official emails but which did not contain state.gov. After scoring the emails, we compared the two members score for each of the emails and calculated Cohen's Kappa coefficient which is defined as:

$$\kappa = \frac{1 - P_o}{1 - P_e}$$

Where:

$P_o = \ Relative\ observed\ agreement\ amount\ raters$

$P_e = \ Hypothetical\ probability\ of\ chance\ agreement, which\ is\ defined\ as:$

$P_e \equiv P_A(+) * P_B(+) + P_A(-) * P_B(-)$

Where:

$P_A(+) = Proportion\ A\ marked\ positive$

$P_B(+) = Proportion\ B\ marked\ positive$

$P_A(-) = Proportion\ A\ marked\ negative$

$P_B(-) = Proportion\ B\ marked\ negative$

Through the use of Cohen's Kappa we were able to obtain a more accurate measure of agreement between our two graders, as Cohen's Kappa takes into account chance agreements.

Related Work

From a Kaggle standpoint, it appears no one has used topic modeling to examine how Secretary Clinton's used her personal email account or to assess her claim that all emails with State Department business have been retained. We have reviewed other submissions on the Kaggle website and many of those use visualization such as word clouds or maps to see the most used words or who Secretary Clinton emailed most often or which countries she emailed most often. We noted a couple submissions using topic modeling, but the output appears to be a list of topics rather than an analysis of the topics. We found an example of the use of LDA to find 50 topics in Clinton emails, but no additional analysis was done using the topics. As far as we have been able to determine, no one has used machine learning to determine if Clinton

violated laws regarding retention of communication. Outside of Kaggle, it has been stated that these emails have been manually vetted, but no results have been published on the matter other than political sound bytes.

More generally speaking, in regards to actual email classification, a wide array of techniques using Latent Dirichlet Allocation ("LDA") in concert with other computational methodologies have been studied. These techniques involve using supervised learning alongside techniques such as graph theory (Joty, 2009) to identify thematic indicators, however, as this was an exploratory task, a supervised learning methodology would not have been helpful. Another adaptation using Author-Topic relations presented by Geng et al seemed much more promising, as their proposed work identified a methodology to identify author-document relations, and document-topic relations (Geng, 2008). While their proposed methodology seemed to be ideal for the task at hand, their algorithms would have had to be manually reconstructed in R, as there are no packages currently available to do this.

To accomplish our goal, we used available R packages to accomplish topic modeling though LDA and NLP, as well as k-means clustering.

It appears that the Clinton emails have only been classified and vetted by hand. We automated this process using machine learning approaches to bridge two gaps: first, it brings automation and machine learning to a task that thus far lacks such an approach, and two it provides a different methodology for data mining that has yet to be explored.

Approach

Approach Taken

We performed several data preparation and pre-processing steps which eventually resulted in a Document-Term Matrix. After creating a data frame using the emails which were released by the State Department, we sorted the data frame putting all emails including state.gov at the top of the data frame. The emails not containing state.gov were further sorted into those between Secretary Clinton / Sidney Blumenthal and the rest. We sorted the emails in this fashion so we could see how document probabilities compare between these groups. According to the New York Times, Sidney Blumenthal is a trusted advisor to Secretary Clinton who frequently corresponded with her and was an unofficial source of intelligence. As such, we selected Sidney Blumenthal as an interesting marker and likely source of emails containing State Department business which may not not include a state.gov email address.

Our approach of applying LDA and k-means clustering to the corpus of sorted emails allowed us to quantitatively evaluate the similarity of topics between the clusters. We hypothesized that clusters which contain a low proportion of non-state.gov emails were likely to contain official business not sent using state.gov email accounts; we used a threshold of 10% of

non-state.gov emails to state.gov emails and looked at all the emails in the clusters at or below this threshold. As noted below, there were some challenges with this but they were related to how the email data was collected, not to our approach.

The emails were manually annotated as good or bad by two independent scorers; the guideline applied was to mark as good any emails that were incorrectly classified as not being state.gov emails such as state.goy or emails missing an email address where at least one person on the email is known to work at the State Department, such as Jacob Sullivan, Huma Abedin, Cheryl Miller, etc. Other emails were read and if they contained State Department related business, they were marked as bad. The scoring achieved through the manual annotation was used to calculate Cohen's Kappa to determine the amount of agreement not due to chance in which the manual annotation was conducted.

Data Source & Preprocessing

Our data was obtained from Kaggle (www.kaggle.com) and is being used in an active competition on their website. Kaggle obtained the first set of Secretary Clinton emails that were released by the State Department. This dataset contains 7,945 observations and 22 variables. We limited the analysis to just one variable - "RawText" - as this contains all email data which was released. The other variables represent an attempt to extract particular pieces of information from "RawText" but these contain many missing values, so these variables are not very useful. It is important to note that the emails were released by the Department of State as PDF documents; these PDF documents were used to create the dataset we obtained from Kaggle which are stored in a csv flat file. Due to the fact that the dataset was created from PDF files, there are errors present in some emails, which will potentially affect the reliability of our text mining results.

First we searched all raw texts of emails for the substring "state.gov" or "stategov", and marked them as the "official" emails. Then the non "official" email group was searched to find if the substring "Sid" (but not a word like "Side" that contains "Sid") exists in the text body. In this way the non "official" emails were divided into two groups. Next, the 7,945 rows of emails were rearranged such that the first part are all "official" emails (1 - 6549), the second part are all non "official" emails without "Sid" in it (6550 - 7586), and the third part are non "official" emails with "Sid" in it (7587 - 7945). This will make it easier to analyze the topics associated with each email groups after the topic modeling.

Since the raw email data contains the texts such as the names and email addresses of senders and recipients, the date / time, the words like "Sent:", "To:" or Fwd:", and the string patterns regarding the classification and release information, the first step of cleaning the texts was to replace all such strings by a " " (space) character. Particularly, in each email all texts showing up before the first occurrence of "Subject:" (if it is present) are not the actual contents of the email so they could be replaced by " ". And every "\n" string was also replaced by " ". Those operations were handled by regular expressions in R. The reason why " " is used in the

initial cleaning procedures instead of directly removing the useless texts is that it could avoid unexpectedly concatenating two words that are separated by the texts to be removed. Lastly, all of the big chunks of upper case texts (string length >15) were replaced by " ", as they are unlikely to be in the email body. This completed the initial text cleaning step.

Next natural language processing (NLP) techniques were used to further clean the emails. First the emails were converted into the "VCorpus" format in R, and all characters except the uppercase and lowercase letters were replaced by " ". Using the built-in stop word dictionary, the corpus of emails was searched through and any stop word encountered were removed. To achieve better text cleaning performance, an additional list of stop words in English (downloaded from http://xpo6.com/list-of-english-stop-words/) was read in and used in the second round of stop words removal. Also, all "words" starting with a lowercase letter that only contained 1 or 2 characters were removed as they are generally not meaningful words. Then the entire corpus of emails was stemmed, in order to reduce the word (term) space for topic modeling. Because the raw email texts were quite messy and they contained a lot of character combinations that are not real English words, an English dictionary containing 109, 582 words (downloaded from http://www-01.sil.org/linguistics/wordlists/english/) was loaded into R and stemmed. Then the words in the entire corpus was compared against this dictionary of stemmed English words, and any word that does not belong to the dictionary was removed from the corpus. Finally the document-term frequency matrix (DTM) was computed.

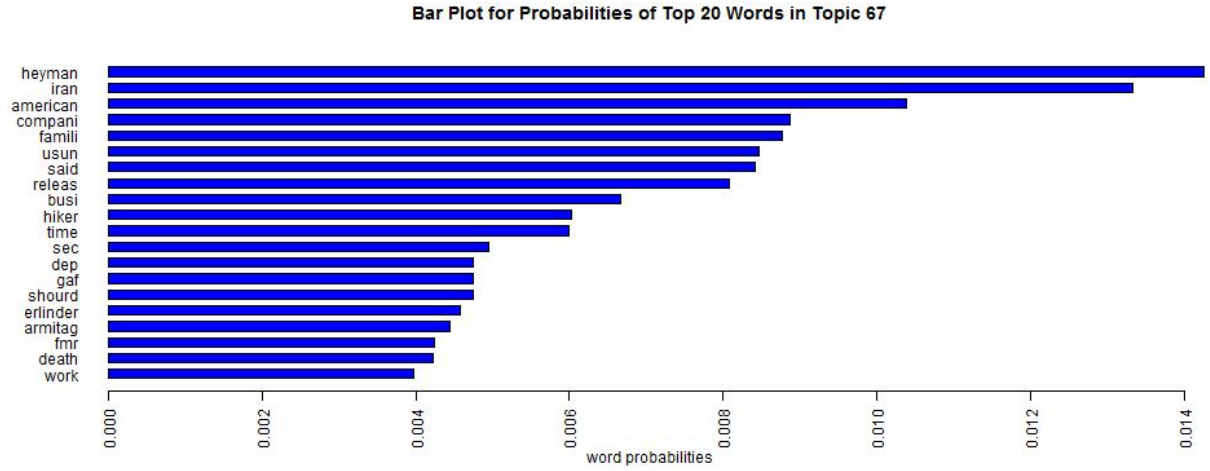
Up to this point, all words that started with the uppercase letter were still preserved in their original format and were not yet checked against the stop words list and the stemmed English dictionary, since there might be some important people or country names among those words, which could be the key words for recognizing a topic after the LDA modeling. So if they were converted into pure lowercase words and then checked against the dictionary, they would be taken as non English words and get deleted. With the DTM at hand whose column names are all distinct words in the corpus sorted in the alphabetical order, all words starting with the uppercase letter could be converted into lower cases. Then there were duplicated column names in the matrix (for instance, "Side" and "side" were different words originally but they became the same word after the uppercase letter was converted into lowercase). These duplicated column names were all marked, so that they could be collapsed into a single column in the DTM in the subsequent step. In order to collapse the columns in DTM, the DTM (simple_triplet_matrix) needed to be transformed into a normal matrix in R first. Due to the enormous size of the DTM, a direct transformation was impossible as this operation would exceed the memory limit of a PC. So the DTM was split into 3 normal matrices, and the column combinations were then completed within each matrix, by collapsing the column vectors which have the same column names into one column (add those column vectors together). Then the 3 normal matrices were transformed back to simple_triplet_matrix format respectively and stacked together as a single DTM.

There were still stop words within the words that originally started with an uppercase letter. So the list of the column names of the DTM were iterated over, and any column whose

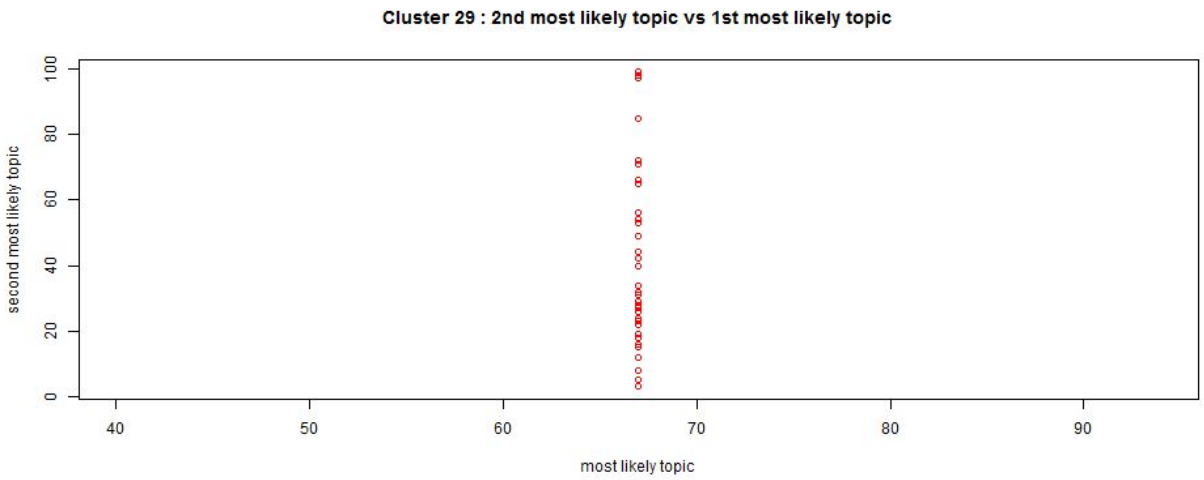
name was a stop word was dropped. Next the column sums of the DTM was computed, and the words whose frequency (column sums) in the corpus is among the top 2000 were inspected, and all “words” that are not actually words or not meaningful words were dropped from DTM. The final step for data preprocessing was to drop the emails that were essentially empty (of which row sums are zero). The dimension of the ultimate DTM (ready for LDA analysis) is 7,888 × 26,661.

Analysis

After preprocessing the data, we performed an LDA analysis using 100 topics over the corpus of emails, and topic probabilities were assigned to the emails. A screenshot of the top 20 words for Topic 67 from our shiny app is shown below:



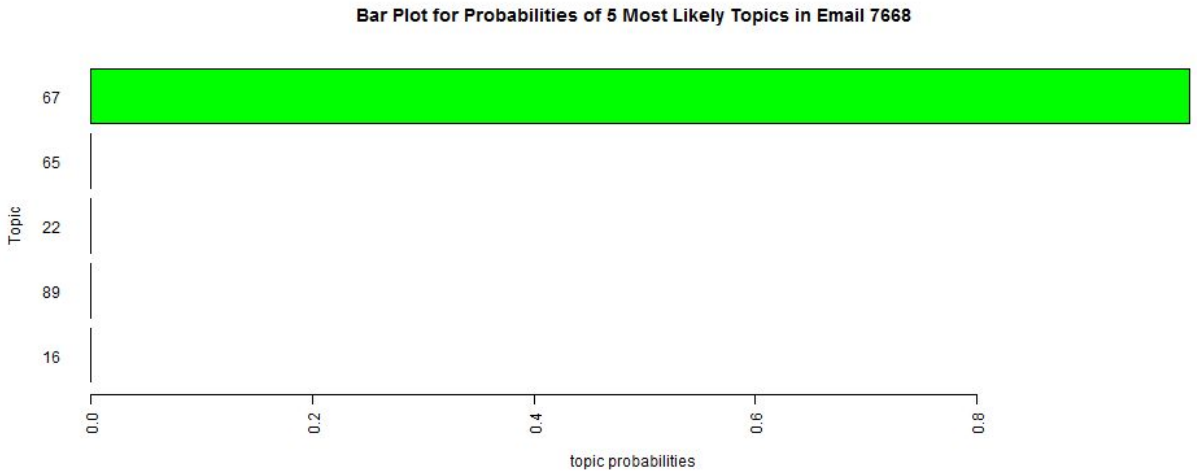
With these newly assigned topic probabilities the emails were clustered into 50 subsets using k-means clustering. A screenshot from our shiny app plotting the second vs. first probable topics for Cluster 29 is depicted below along with the emails within this cluster:



Emails contained in the cluster

256, 892, 893, 894, 984, 987, 1014, 1175, 1263, 1496, 2170, 2619, 2623, 2625, 2631, 2636, 2637, 2639, 2663, 2792, 2943, 3082, 3523, 3717, 3797, 3820, 3821, 3823, 3824, 3855, 4015, 4031, 4143, 4230, 4365, 4484, 4571, 4671, 4783, 4784, 4786, 4787, 4791, 4795, 4833, 4845, 4852, 4858, 4899, 5084, 5229, 5243, 5306, 5378, 5520, 5627, 5758, 5801, 5813, 5986, 6095, 6182, 6330, 6480, 7138, 7214, 7525, 7668, 7860, 7888

A screenshot of the probabilities associated with the top five topics in Cluster 29 are also listed:



After clustering the counts of “state.gov” emails, non “state.gov” emails containing “ sid ”, and non “state.gov” emails not containing “ sid ” were calculated, along with the ratio of non “state.gov” emails containing “ sid ”, and non “state.gov” emails not containing “ sid “ over the ratio of “state.gov” emails.

Using the aforementioned ratio, a threshold value of 0.10 was chosen and all clusters falling into this range were flagged as potential “hot clusters” due to the fact that these clusters would likely contain very offical dense topics, and the non “state.gov” emails containing “ sid ”, and non “state.gov” emails not containing “ sid “ within these topics would have a higher probability of containing offical information. Through this process we flagged 110 emails to be reviewed. The following clusters fell below the threshold:

1 2 11 12 13 20 29 36 37 39 45 49 50

These clusters contained 110 emails which had the potential to contain State Department business by no state.gov email address. All 110 emails in these clusters were manually annotated/scored as good or bad based on the criteria described above under Approach Taken.

Evaluation

Out of the 110 emails we manually reviewed, our reviewers agreed on 96 emails (87% of total). The probability of match by chance was 60.36%, which yielded us a Cohen’s Kappa of 0.6789 which represents the probability of agreement not due to chance. The calculation is depicted below:

$P_A(+)$	74	$P_A(-)$	36
$P_B(+)$	88	$P_B(-)$	22
$P_e =$	0.6036	$P_o =$	0.8727
$\kappa =$	0.6789		

This implies that our methodology of manual annotation was substantially robust. Overall, we were able to uncover 22 emails that could be deemed as “non-compliant.”

Of these 22 emails, our approach identified 18 emails which were sent from Sidney Blumenthal (“Sid”) to Secretary Clinton. Some of these emails contained “intelligence” topics that he gathered in his travels and meetings with various people, such as Joschka Fischer, former Foreign Minister and Vice Chancellor of Germany (1998-2005) where they discussed a policy approach to Iran which advocated for an “iron fist in velvet glove” approach which would couple harsh sanctions with the possibility for negotiations. It is interesting to note that this approach is the one adopted by the US Administration, although identifying the reason for this is

beyond the scope of our work. Only 5 of the Sid emails were not unique, and as such contained 13 different topics, 8 of which concerned Libya, including the attack at Benghazi. Only 1 of the emails appeared to be of a truly personal nature.

Despite the sucess of our approach, there were some issues and one area in which our approach broke down was in the identification of non “state.gov” emails. There were a myriad of instances in which it appears that the pdf ripper that produced the raw data inaccurately pulled the “state.gov” email address as “state.goy”. Furthermore, there were many instances in which the email address would not pull at all. However, upon review of the names in the subject line it was apparent that that information was sent to a “state.gov” email, the pdf ripper just did not adequetely capture the information.

With these issues in mind, our method appears to have greatly outstripped the known manual methodology utilized by the U.S. Government, at least those known to the U.S. public. Our methodology allowed us to quickly parse the nearly 8,000 documents down to 110 emails that needed to be manually vetted. Furthermore, it appears we were able to identify findings similar to those presented in the press through using mostly automated means.

Conclusions and Recommendations

Some of the important things we learned in doing this project are very practical lessons. For instance, we learned that working with a relatively large dataset can require a lot of preprocessing, including multiple passes, in order to remove as many non-meaningful words/terms as possible from the corpus. For example, in the context of our analysis, we did not initially consider that it would be helpful to even remove date related words such as day of week, month, etc. as they did not help the model to determine topics.

One really important lesson we learned is to not underestimate the challenges of working with real world data. As mentioned above, the emails were released as PDF files which were converted to as CSV file by Kaggle, and naturally errors were introduced into the data; it would have been worth spending even more time upfront exploring the data, discovering and handling these errors upfront. It might have saved us some time in the stemming of the email corpus. Additionally, it might have allowed the LDA and k-means clustering to achieve better results and would have saved time in the manual annotation/scoring step.

In terms of evaluating our model by manually scoring the emails, we learned that very clear instructions and standards about the scoring process are necessary in order to achieve worthwhile results.

In summary, we feel that the approach we took using LDA to create topics from the emails and k-means clustering to combine the topics was ultimately successful and we were able to quantitatively find emails with similar topic probabilities. In the end, we used these

techniques to identify 18 emails from a source outside the State Department who sent emails to Secretary Clinton, not all of which were forwarded to someone with a state.gov. This means that strictly speaking, Secretary Clinton's claim that all emails containing State Department business have been retained, does not hold up. Whether these emails are sensitive or important enough to represent a violation of the Federal Records Act will be for courts or some other body to decide, but we think we identified some emails they would want to consider.

For future work, it would be interesting to attempt this approach in a parallel environment, such as Spark, to see if better results could be achieved. Additionally, it would interesting to add more emails, once available, to see if that would improve performance of the LDA. Additionally, it would be interesting to have assistance from a State Department/foreign affairs subject-matter expert who could assist in better identifying topics. Lastly, it would be interesting to attempt the Author-Topic relations presented by Geng et al, as this may yield even more accurate LDA results, and in turn better clustering.

References

Joty, Shafiq, Giuseppe Carenini, Gabriel Murray, and Raymond Ng. "Finding Topics in Emails: Is LDA enough?." In *NIPS-2009 workshop on applications for topic models: text and beyond*. 2009.

Geng, Liqiang, Hao Wang, Xin Wang, and Larry Korba. "Adapting LDA model to discover author-topic relations for email analysis." (2008): 337-346.

https://www.kaggle.com/c/hillary-clinton-emails

http://www.harvest.ai/blog/2015/10/12/topicmodelingfoiadata

http://www.newsweek.com/hillary-clinton-email-private-server-top-secret-362822

http://www.nytimes.com/2015/09/02/us/politics/emails-show-how-hillary-clinton-valued-input-from-sidney-blumenthal.html?_r=0