Interim Report: Analysis of Hillary Clinton Mails

Debojyoti Dey
Roll No. 15511264

Nimisha Agarwal
Roll No. 15511267

November 9, 2016

1 Problem Statement

The motive of the project is to explore and analyse the released emails by US State Secretary Hillary Clinton with the following goals in mind:

Outline

- Extracting important topics from the emails. We should get a list of topics ranked according to their number of occurence.
- Getting an idea of US Foreign policy and relation with different nations. The nature of relation may vary over time. So we do a monthly/quarterly study to create a relational map with other countries using sentiment analysis.
- Persons linked to Clinton and Hillary's sentiment about them. The persons may be from sender/recipient list of the emails or mentioned frequently in e-mails.

2 Progress Made

2.1 Preprocessing dataset

As we are using bag of words model to represent a text(email body here), we need to normalize the text before we start tokenizing it. We normalize

date and time strings. A number of email bodies contain sytactic notations. We keep only the alphabetic words. We cast all words into small cases. We discard all numbers from the mail body. We create a stoplist i.e. set of words to be subtracted from the main text. Now we tokenize the main text and represent an email as vector of sentences, which are again vector of words. Now we are ready to create our vocabulary and represent each word by its position in it. The vocabulary is a Python dictionary.

2.2 Extract topics from emails

Now that we have vectorized our text, we can represent an email body as a bag of words form. We have a normalized tf-idf representation. Topic modelling tools like Gensim developed several ways to transform a document vector into small dimensional vector space. We tried Gensim's LDA based topic detection model to get a probability distribution over all the words in vocabulary. The predominant topics will be the words with "high" probability. Now we can categorize an email for some topic by checking the presence of topic words in it.

We want to try BigARTM package for topic modelling in future and compare its result with the one by Gensim.

3 Things to do

3.1 Sentiment analysis of an email

This part is to identify the following two things:

1. sentiment for a nation
2. sentiment for a person

Stanford has built a compositional Sentiment-treebank on which they train a Recursive Neural Tensor Network to detect sentiment(positive/negative/neutral) of an entire sentence, in stead of individual words. Thus we get sentiment for each sentence in the documents. To find sentiment for a particular nation, we need to identify all the sentences containing the country name or having a reference to it. Stanford corenlp also has a dereferencing implementation in Java. We may be able to use this as well to identify all the sentences having a link to a nation, explicit or implicit.

For sentiment about a person, we have to do the similar trick to get Hillary's sentiment towards him/her.

4 Conclusion

Our work, by no way, is unique or first. We have consulted several existing forums to get an idea about things that could be done. We have referred to the libraries have used or planning to do so. We may change some implementation from Java to Python. We hope to complete the project by end of semester.