Hillary Clinton - Telling The Story with Data

Yihe Chen, Daitong Li, Palmer Lao, Ziyue Shuai, Eric Zhang

Objective

We will apply various techniques in NLP, Social Network Analysis, ML, and Statistical Inference to analyze Hillary Clinton’s emails released in 2015 for embedded themes, ideas, and relationships.

Then, we will create a visual dynamic dashboard that will summarize these findings and allow for further exploration.

Data Source

Publicly available dataset (preprocessed and hosted on Kaggle):

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

Format - 4 SQL Tables:

1. Emails
2. Persons
3. Aliases
4. Email Recipients

Methods

Angles to consider:

- Text Summarization
- Social Network Analysis
- Time Series
- Similarity Analysis
- Visualization Tools and Techniques

EDA

- Number of emails sent and received
- Frequencies of emails
- Summarization of topics

Text Summarization

- Most work on the e-mails focuses on **topic modelling**
  - Each document is reduced to a distribution of topics and words within topics

- As part of our analysis, we will perform **extractive summarization**
  - Select a representative unit of text from each document
  - Our plan is to use TextRank

Social Network Analysis (SNA)

- **SNA**: investigate social structures through Network and Graph Theories.
- **Social Structure:** characterized by *nodes* (i.e. individual actors or things in network) and *edges* (i.e. friendship, kinship, collaborative, disease transmission etc.)
- Primary SNA for this project will focus on **visual exploration** using R and iGraph.
- **Clustering** algorithms such as logistic regression, K-NN may be applied to identify natural groupings of data and thus find the informal communities in the network.

Similarity Analysis

- For emails addressed to different people/organizations, produce dendrograms from a cluster observations analysis
- Use PCA to find a few factors to distinguish the texts sent to different receivers, to see if Hillary speaks in the same ways to certain circles of people
- Identify similarities and distinctive wording patterns in Hillary’s emails

Time Series

- Survival analysis of different recipients in term of time

The time of last email for one certain recipient will be modeled with survival model with the time since first email.

- Survival analysis of different recipients in term of email frequency

The time of last email for one certain recipient will be modeled with survival model with the email frequency since first email.

Visualization Tools and Techniques

R is the language of choice given that the team is most familiar with it. For visualization and presentation, RShiny works well, as it supports dynamic figures.

However, one may also integrate RShiny with D3.js, another popular visualization framework. This option is currently being explored.

Results

Pending

References

“Hillary Clinton’s Emails.” Kaggle. Last Updated 9-11-2015. Retrieved 9-19-2016. https://www.kaggle.com/kaggle/hillary-clinton-emails.

Thompson, James. “Integrating D3.JS into R Shiny.” 3-13-2016. http://myinspirationinformation.com/visualisation/d3-js/integrating-d3-js-into-r-shiny/

Questions?

Thank you for listening!