Visualizing Hillary Clinton's Emails

Yihe Chen yc3076 | Palmer Lao pol2105 | Daitong Li dl2991 | Ziyue Shuai zs2285 | Eric Zhang ez2232

December 7, 2016

1 Introduction

In light of the recent U.S. presidential election race between candidates Donald Trump and Hillary Clinton, there have been large amounts of interest in the controversy involving Hillary's use of personal email accounts on non-government servers during her previous career as Secretary of State. After a number of Freedom of Information lawsuits were filed against the Department of State, the Department of State released on August 31, 2015 nearly 7,000 pages of Clinton's heavily redacted communications in PDF form.

Subsequent to this release, the data science competition website Kaggle released a sanitized version of the extracted content of the emails for public use and analysis. Our group was interested in analyzing and exploring this data given its connection to current events at the time.

2 Objectives

Our main objective with this project was to reduce the immense dimensionality and volume of the data so that the resulting output could be more easily digested by a data analyst or other interested party. We wanted to find patterns in the data that might be of interest, such as communities of receiving or sending parties that shared certain commonalities.

However, rather than simply report on the results of such analysis, we believed that a more effective way to communicate results and allow for the generation of new insights would be to

create an interactive visualization that would empower the user to both view our findings and explore the simplified data at will.

3 Data Source

4 Methodology

4.1 An Overview of our Summarization Procedure for one Document

4.2 The TextRank Algorithm

4.3 Cleaning the e-mails

5 Results

6 Validation

7 Conclusion