POLI 175 Challenge 2

Due 11:59PM Thursday March 11, 2021

General Instructions

In this assignment, we are providing you with a text data set. You'll work in teams of 3-4 to analyze the data using any methods we have learned in this class that you deem appropriate.

Please submit this assignment by uploading your write-up, html, and code files onto **Canvas** (navigate to the Assignments tab on our Canvas course website) before the due time. **Only one person from each group should upload the files.**

Introduction

You are being provided with a text data set pertaining to approximately 8000 emails sent and received by Hillary Clinton when she was Secretary of State. The emails were released by the U.S. State Department in 2015.

The instructors' choice to use these data for this assignment is not meant to express any particular political affiliation or intent.

Your goal for this assignment is the following.

ASSIGNMENT

1. **Employ any methods you have learned in this class (during the supervised and/or unsupervised units) to analyze the text data in any way you see fit.**
2. **Through your analyses, identify a key insight or discovery that you think is interesting and would not be easy to ascertain without the methods you employed.**

Your team must decide which methods to employ, how to analyze the data, and how to present your results. This assignment is designed to be open-ended.

Data

To aid in your endeavors, we are providing you with an already processed version of the email data, in `Clinton.csv`. This file contains email metadata as well as a document term matrix, containing unigrams, bigrams, and trigrams. It is up to your group to decide how to use the data, as well as how much of it to use. The structure of the dataset is as follows:

- Each row represents a unique email.
- Columns 1 to 9 contain metadata.
- Columns 10 to 3009 contain counts of the top 3000 frequent unigrams.
- Columns 3010 to 4009 contain counts of the top 1000 frequent bigrams.
- Columns 4010 to 4509 contain counts of the top 500 frequent trigrams.

In addition, the following is a link to the kaggle site that hosts the email data:
https://www.kaggle.com/kaggle/hillary-clinton-emails

You can access the raw email text data, if you should desire to do so, via that site. However, you are not required to use any other datasets other than Clinton.csv.

Guidelines

Each team must submit the following:

1. A 3-4 page standalone write-up of their discovery and how they made that discovery, submitted as a pdf, Word, or similar document file.
   - At the top of the submission, there should be a 2-4 sentence summary of what you discovered and why you think it is interesting and/or important. Be sure this is a clearly identifiable section of the submission.
   - The remainder of the submission should explain the methods used to make the discovery and any other models/methods that you tried along the way.
   - Tables and figures are encouraged to help display and visualize your insights, where appropriate.
   - This document should not include your actual code.
   - Bibliographies and appendices (which are not required) do not count towards the page limit.
2. Your html and code file(s).

An example of a past submission for this assignment, example_submission.pdf, is also provided. This is an example of a successful, interesting, and well-executed submission for this assignment.

Evaluation

Your grade for this challenge project will be determined by:

1. Succinct summary of your discovery and substantive reasoning of its novelty and/or importance. (25%)
2. Successful implementation of two different (supervised or unsupervised) learning methods that we have covered in class. (15%)

3. Concise justification for the choice of any methods you tried and employed to make the discovery. (30%)

4. Informative display and creative visualization of your results through tables, plots, and figures. (15%)

5. Clear documentation of your work and coding procedures. (15%)

More Background on the Emails

More information on the Clinton email controversy can be found at:
`https://en.wikipedia.org/wiki/Hillary_Clinton_email_controversy`

In addition, Kaggle's background on the email data is pasted below:

Kaggle's Background on the Data

Throughout 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. Some political experts and opponents maintain that Clinton's use of personal email accounts to conduct Secretary of State affairs is in violation of protocols and federal laws that ensure appropriate recordkeeping of government activity.

There have been a number of Freedom of Information lawsuits filed over the State Department's failure to fully release the emails sent and received on Clinton's private accounts. On Monday, August 31, the State Department released nearly 7,000 pages of Clinton's heavily redacted emails (its biggest release of emails to date).

The documents were released by the State Department as PDFs. We've cleaned and normalized the released documents and are hosting them for public analysis. Kaggle's choice to host this dataset is not meant to express any particular political affiliation or intent.