Data Mining Assignment – Clinton Emails

# Description

This dataset comes from Kaggle.com. Following is their description of the dataset (found https://www.kaggle.com/kaggle/hillary-clinton-emails):

“Throughout 2015, Hillary Clinton has been embroiled in [controversy](https://en.wikipedia.org/wiki/Hillary_Clinton_email_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. Hillary's campaign has provided their own four sentence summary of her email use [here](https://www.hillaryclinton.com/hillarys-emails-four-sentences/).

“There have been a number of Freedom of Information [lawsuits](https://en.wikipedia.org/wiki/Hillary_Clinton_email_controversy#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.”

# Dataset Introduction

# Learning Objectives

Learn how to perform the following data mining and analysis tasks in RapidMiner:

* Generate attributes
* Create scatterplots
* Create and interpret K-means clusters

# Before you Begin

* Complete all of the RapidMiner tutorials
* Complete the following chapters’ exercises “Data Mining for the Masses”, Dr. North, 1st edition. Available to purchase on Amazon or for free PDF download [here](http://docs.rapidminer.com/downloads/DataMiningForTheMasses.pdf).
  + Chapter 6: k-Means clustering

You will need the following datafiles for the RapidMiner analysis. The first two are straight from Kaggle.com:

* Persons.csv
* EmailReceivers.csv

And the second two have been processed specifically for RapidMiner and this assignment:

* Clinton\_emails\_for\_rapidminer.csv
* Clinton\_redaction.csv

# Context

You are interested to know who the most controversial people are in this corpus, based on how often they sent or received redacted emails.

# Questions

## Descriptive information on the dataset

* How many people are involved in this dataset?
  + Hint: count the number of rows in the “Persons” table
* Who sent the most emails? Show a chart of the top 10 senders
  + Hint: Do a `group aggregation` on the `Emails` table, then count the number of rows for each group, and then filter to the top ten for the count.
  + Hint: Join the “Persons” table to the “emails” table to get the name of the person. Don’t rely on the person name in the “Emails” table – it may have typos.
* Who received the most emails? Show a chart of the top 10 receivers
  + Hint: Join the “EmailReceivers” table to the “Emails” table using the “EmailId” column in the “EmailReceivers” table to get one row for each person who received a given email.

## Redaction analysis

The email dataset includes information about whether a released email was “released in full” or “released in part” – if it was released in part, then that means that part of the message was redacted, likely for security reasons. You will now investigate who are the most controversial persons.

A view of the dataset has been created for you which gives count for each person of how many of the emails they *sent* were redacted in part, as well as for how many emails they *received* were redacted in part.

* Generate two new attributes for each person. The two new attributes should be ratios of how many of the emails they *sent* and *received* were redacted. Name them as follows:
  + sender\_redacted\_ratio
  + receiver\_redacted\_ratio
  + Generate a scatterplot with `sender\_redacted\_ratio` on one axis and `receiver\_redacted\_ratio` on the other. Use a jitter so that points don’t overlap. Looking at the chart, how many naturally-occurring groups do you think might exist? Show a copy of your chart.
  + Create a subset of your dataframe keeping just person Name and your two ratio columns. Set Name to be your dataset’s `id` attribute (use the `Set Role` Operator). Perform a k-means cluster analysis on this subset, specifying 3 groups.
  + Create a scatterplot of your dataset with the new cluster\_id attribute. As before, set your two ratio attributes to be the x- and y-axes, and set the assigned cluster\_id to be color attribute for the scatterplot. Use a jitter. Include a picture of your chart. Are these the groups you expected to come out of the cluster analysis?
  + Include your centroid table. Which cluster is the “most redacted”? “Least redacted”? “Middle-grounders”?
  + Which cluster were the following people assigned to?
    - Hillary Clinton
    - Huma Abedin
    - Jake Sullivan
    - Philippe Reines
    - Sidney Blumentahl

# Deliverable

* The answers to the above questions, in .docx or .pdf format.
* A copy of the RapidMiner process(es) you used to obtain the answer(s).