First Capstone Project | Springboard
Author: Amir Benbouza

Analysis: FOIA Release of Email Corpus Stored on Private Email Server of Former Secretary of State and NY Sen. Hillary Rodham Clinton

Q: What is the problem you want to solve?

A: Initially, analysis of the email corpus belonging to a private mail server (clintonmail.com) of Presidential Candidate and Senator H. R. Clinton released by the US State Department as part of a FOIA request. If successful, the project will move to analyze another 29,000 emails obtained through a different channel, which belong to the same private email server.

Q: Who is your client and why do they care about this problem? In other words, what will your client DO or DECIDE based on your analysis that they wouldn't have otherwise?

A: Sen. H. R. Clinton used a private email server to conduct official State Department business as well as private business such as campaigning for the 2016 Presidential Elections, bypassing the IT requirements of someone in her position as head of the State Department to use the official mail server of said department. My client is the Federal Government as well as the American people, who would like to know whether sensitive or classified information was transferred or leaked through this private email server, as well as whether there were any occurrences of legally dubious or suspect exchanges via this mail server.

They care about this problem in order to set a precedent for future candidates and Secretaries of State (and other departments) implementing a BYOD policy, endangering transparency and accountability while conducting governmental duties, which may interfere with their private affairs conducted on the same server, leading to a suspicion or even accusations of corruption. Based on my analysis, the client will obtain a public report on the exchanges communicated through this server, and be reassured once and for all that transparency and accountability were preserved. It also serves as a warning shot to any future representative in government who uses their own private servers (or BYOD) for governmental affairs as well as other, possibly conflicting, matters, such as election campaigns.

Q: What data are you going to use for this? How will you acquire this data? In brief, outline your approach to solving this problem (knowing that this might change later).

A: Details (First Dataset):

I have obtained nearly 7,000 emails ETL'd from clintonmail.com, released by the State Department as part of a FOIA request. The dataset was obtained from Kaggle.com. As for the second dataset, it contains nearly 29,000 emails from the same server. While obtained through a third-party, it was nevertheless released by the State Department under another FOIA request and both datasets are "UNCLASSIFIED". I suspect the second dataset is only available to certain media organizations and news outlets.

The data is available in two formats, CSV files, and an SQLite database containing the data in the files. The SQLite data is not relational; I had to add Foreign Keys myself. Furthermore, the data is not exactly clean. Many people in the database are identified via different aliases (I intend to fix this issue). Additionally, some fields are missing due to redaction by the state department. What I intend to do is verify the missing fields manually.

History of the Datasets:

> *First Dataset (+7000):* The State Department released 7000+ PDF printouts as well as metadata of emails belonging to clintonmail.com, a private mail server of the Senator, related to State Department daily affairs. The dataset was made available on Kaggle.com, and some members cleaned up the data, OCR'd the PDFs and ETL'd the data into CSV files, namely: Emails, EmailsReceivers, Persons, Aliases. They also stored a duplicate of the same data into an SQLite DB file to make queries easier. Note that the members did not trust the metadata released along with the 7,000 PDF files, and ran their own ETL on the PDFs. The SQLite DB contains both the Kaggle' members metadata as well as the State Department's metadata.)
>
> *Second Dataset (+29,000):* The State Department further released thousands of other emails from the same server as PDF files. My analysis of this dataset is not complete, but I suspect I'll OCR the files into text files (either a pdf2txt UNIX utility or OmniPage or Readiris), then run a Naive Bayes or Deep Neural Network algorithm to classify the emails into "RE:", "FW:", "Press Release", etc. I'll also identify (probably using a Python PDF library) who sent what to whom on which date, creating an Entity Database (Entity Extraction, followed by generating a Social Network graph for Analysis), as well as timeline visualizations of exchanges between various entities. I also plan on running a sentiment analysis for each email exchange that took place, for each party. Last, I plan on exporting the data to a Maltego / Casefile/IBM Analyst i2 file for further analysis using Patreva's / IBM i2's transforms. Last but not least I intend to run a few data mining algorithms to automatically find relationships in the dataset that are not so obvious.

I may use Amazon's Mechanical Turk as a form of Supervised Learning to classify the emails after clusterification via DNN. I may also use a GPUPU to parse the dataset.

My toolset will include: Python (general scripting, statistical analysis), sed and awk (data cleanup and wrangling), NavicatPremium (interfacing with SQLite), SQLite (store the files in one file database), SQL (to query the database, statistical analysis), Python NLTK (NLP), Python Seaborn (visualization), etc.

Q: What are your deliverables? Typically, this would include code, along with a paper and/or a slide deck.

A: Deliverables, The Datasets, The Datasets' derivatives (text' corpus, relationships, social graph, etc.), New Databases (further ETL to generate new intelligence from available data), Source code, Report with the most interesting questions and patterns answered and observed by my analysis noted above, as well as my methodology, A Slide Deck.