Dataset description

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. Hillary's campaign has provided their own four sentence summary of her email use here.

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.

Data Source

Table "Aliases"

1. **Id** - unique identifier for internal reference
2. **Alias** - text in the From/To email fields that refers to the person
3. **PersonId** - person that the alias refers to

Table "Email Receivers"

1. **Id** - unique identifier for internal reference
2. **EmailId** - Id of the email
3. **PersonId** - Id of the person that received the email

Table "Emails"

1. **Id** - unique identifier for internal reference
2. **DocNumber** - FOIA document number
3. **MetadataSubject** - Email SUBJECT field (from the FOIA metadata)
4. **MetadataTo** - Email TO field (from the FOIA metadata)
5. **MetadataFrom** - Email FROM field (from the FOIA metadata)
6. **SenderPersonId** - PersonId of the email sender (linking to Persons table)
7. **MetadataDateSent** - Date the email was sent (from the FOIA metadata)
8. **MetadataDateReleased** - Date the email was released (from the FOIA metadata)
9. **MetadataPdfLink** - Link to the original PDF document (from the FOIA metadata)
10. **MetadataCaseNumber** - Case number (from the FOIA metadata)
11. **MetadataDocumentClass** - Document class (from the FOIA metadata)
12. **ExtractedSubject** - Email SUBJECT field (extracted from the PDF)

13. **ExtractedTo** - Email TO field (extracted from the PDF)
14. **ExtractedFrom** - Email FROM field (extracted from the PDF)
15. **ExtractedCc** - Email CC field (extracted from the PDF)
16. **ExtractedDateSent** - Date the email was sent (extracted from the PDF)
17. **ExtractedCaseNumber** - Case number (extracted from the PDF)
18. **ExtractedDocNumber** - Doc number (extracted from the PDF)
19. **ExtractedDateReleased** - Date the email was released (extracted from the PDF)
20. **ExtractedReleaseInPartOrFull** - Whether the email was partially censored (extracted from the PDF)
21. **ExtractedBodyText** - Attempt to only pull out the text in the body that the email sender wrote (extracted from the PDF)
22. **RawText** - Raw email text (extracted from the PDF)

Table "Persons"

1. **Id** - unique identifier for internal reference
2. **Name** - person's name