Enron Email Dataset

This dataset was collected and prepared by the CALO Project (A Cognitive Assistant that Learns and Organizes). It contains data from about 150 users, mostly senior management of Enron, organized into folders. The corpus contains a total of about 0.5M messages. This data was originally made public, and posted to the web, by the Federal Energy Regulatory Commission during its investigation.

The email dataset was later purchased by Leslie Kaelbling at MIT, and turned out to have a number of integrity problems. A number of folks at SRI, notably Melinda Gervasio, worked hard to correct these problems, and it is thanks to them (not me) that the dataset is available. The dataset here does not include attachments, and some messages have been deleted "as part of a redaction effort due to requests from affected employees". Invalid email addresses were converted to something of the form user@enron.com whenever possible (i.e., recipient is specified in some parse-able format like "Doe, John" or "Mary K. Smith") and to no_address@enron.com when no recipient was specified.

I get a number of questions about this corpus each week, which I am unable to answer, mostly because they deal with preparation issues and such that I just don't know about. If you ask me a question and I don't answer, please don't feel slighted.

I am distributing this dataset as a resource for researchers who are interested in improving current email tools, or understanding how email is currently used. This data is valuable; to my knowledge it is the only substantial collection of "real" email that is public. The reason other datasets are not public is because of privacy concerns. In using this dataset, please be sensitive to the privacy of the people involved (and remember that many of these people were certainly not involved in any of the actions which precipitated the investigation.)

- Prior versions of the dataset are **no longer being distributed.** If you are using the March 2, 2004 Version; the August 21, 2009 Version; or the April 2, 2011 Version of this dataset for your work, you are requested to replace it with the newer version of the dataset below, or make the the appropriate changes to your local copy.
- May 7, 2015 Version of dataset (about 1.7Gb, tarred and gzipped).

There are also several on-line databases that allow you to search the data, at UCB, and www.enron-mail.com

Research uses of the dataset

This is a partial and poorly maintained list. If I've left your work out, don't take it personally, and feel free to send me a pointer and/or description.

- A paper describing the Enron data was presented at the 2004 CEAS conference.
- Some experiments associated with this data are described on Ron Bekkerman's home page.
- A social-network analysis of the data, including "useful mappings between the MD5 digest of the email bodies and such things as authors, recipients, etc", is available from Andres Corrada-Emmanuel.
- A group from SIMS, UC Berkeley provides search, visualization, and some email that has been labeled with topic and sentiment labels
- Jitesh Shetty has put up a database of link-analysis results.
- A version of the dataset with all attachments is available from EDRM.
- Work at the University of Pennsylvania includes a query dataset for email search as well as a tool for generating spelling errors based on the Enron corpus.
- Kimmie Farrington and colleagues published a paper in 2011 that uses the Enron dataset as part of the test corpus for their work on crowdsourcing human vs. computer generated classification explanation: see Hutton, Amanda, Alexander Liu, and Cheryl Martin. "Crowdsourcing evaluations of classifier interpretability." In *Proceedings of the 2012 AAAI Spring Symposium on Wisdom of the Crowd*

- Parakweet has released an open source set of Enron sentence data, labeled for speech acts.
- A set of sentence level annotations (of what requires action or response from user) has been released by Charlie Oxborough.

William W. Cohen, MLD, CMU
Last modified: Fri May 8 09:52:31 EDT 2015