The 2001 Annotated (by Topic) Enron Email Data Set

By Dr. Michael W. Berry and Murray Browne
April 10, 2007

The 2001 Annotated (by Topic) Enron Email Data Set contains approximately 5,000 emails manually indexed into 32 topics. It is a subset of the original Enron email dataset of 1.5 million emails that was posted on the Federal Energy Regulatory Commission (FERC) site as a matter of public record during the investigation of the Enron Corporation. However, the original set suffered from document integrity problems and attempts were made to improve the quality of the data and remove some of the sensitive and private information. Dr. William Cohen of Carnegie Mellon University took the lead in distributing the improved corpus, which led to the distribution of 517, 431 emails of 150 Enron employees, from the years 1999-2002. This set of emails is still downloadable from www.cs.cmu.edu/~enron.

It is from this set of emails that we have created our dataset of approximately 5,000 emails covering the January to December 2001 timeframe. It is our understanding that the original emails are in the public domain.

Creation of the Annotated by Topic Enron Email Data Set

Inherently, manual indexing has some problems because human indexers do not always categorize information the same way. Assigning a topic to an abstract or in this case to an email, may shift from indexer to indexer, making manual indexing a collaborative art as much as an exact science. The two indexers, Murray Browne and Ben Signer, who worked on indexing the emails made every effort to be as consistent as possible.

Since October of 2004, Browne, a research associate for Dr. Berry, has been evaluating the Enron emails. He has read several books about the collapse of America's seventh largest company. The most influential and helpful book, in his opinion, was Brittany McLean and Peter Elkind's *The Smartest Guys in the Room: The Amazing Rise and Scandalous Fall of Enron* published in 2003. The book was an excellent background source about Enron and it included an account about the last year of the company's existence (2001) before it filed for bankruptcy on December 2, 2001. Browne worked with a student programmer to develop a tool call MailMiner that allowed him to search, organize and assign topics to emails. He worked on assigning topics several months before hiring a student Ben Signer (a junior in Economics) to assist him in manual indexing the collection. Before working on the collection, Signer was assigned to read *The Smartest Guys* and Browne worked with Signer on a daily basis in the summer of 2006 to assign topics to the emails.

The 32 topics are:

Calif_analysis 1 -- Executive summaries and analyses about the California situation. [1] (304 entries)

Calif_bankruptcy 2 -- Specifically mentioned financial difficulties of the utilities such as Southern California Edison (SoCal Edison) and Pacific Gas & Electric (PG & E). (36 entries)

Calif_utilities 3 -- General references to California utility companies: Edison, Pacific Gas & Electric, and the California Public Utility Commission (CPUC) which regulates them. (116)

Calif_crisis_legal 4 -- Articles about legal issues surrounding California energy crisis. (109)

Calif_enron 5 -- Enron business emails about the day to day operations of managing the California side of their business. (699)

Calif_federal 6 -- Emails about FERC (Federal Energy Regulatory Commission), U.S. Senate Hearings. (61)

Newsfeed_Calif 7 -- Long emails with a host of stories about California. These emails were news feeds from wire services such as Reuters and Dow Jones, which were widely circulated among Enron employees. (190)

Calif_legis 8 -- Emails about California legislature, bills in the California legislature or California Governor Gray Davis that are *not* related to the specifics such as bankruptcy or the energy crisis. (181)

Daily_business 9 -- As one might expect, the majority of the emails in this collection are emails about the regular day to day activities of a multinational energy company (i.e. " trade this share, buy these shares," etc.). Other daily business emails include setting up meetings, confirming meetings, and general announcements from human resources. These almost defy categorization by topic, but they do have a value. Researchers may decide to remove these emails to reduce the amount of noise in the collection and to improve their ability to detect topics. However keeping them in the collection provides an element of noise that gives the collection a "real life" quality. Either way by tagging such emails, the researcher has the option. (1595)

---

[1] *Historical Note about California Topics:* Deregulation of the California energy market led to rolling blackouts beginning in the summer of 2000 — a situation that Enron and other energy companies took advantage of financially. There were investigations at a state and national level about what happened in California.

Educational 10 -- This was a surprise topic that emerged later. It related to Enron's interns, scholarships or employees who are professors. Many of these emails center around the Head of the Research Group Vince Kaminski who taught at Rice University in Houston part time. (92)

EnronOnline 11 -- Enrononline is the electronic trading and information software tool that the Enron traders used. It was an invaluable asset to the company and gave them an edge on their competitors. Louise Kitchen was an early developer of the technology. (271)

Kitchen_daily 12 -- Daily emails to and from Louise Kitchen who developed Enrononline. This category includes questions to Kitchen about running EOL and trading information. (37)

Kitchen_fortune 13 -- Louise Kitchen was selected as one to the top corporate women in a *Fortune* magazine story (September 2001). (11)

Energy_newsfeed 14 -- Wire news feeds about various energy issues. Think of it as an electronic newsletter about energy that is circulation to a number of Enron employees. Usually these are lengthy emails. (332)

General_newsfeed 15 -- Long emails (wire feeds) with a host of general national and international stories. (48)

Downfall 16 -- Articles about Enron's demise. Messages from employees worrying about what is going on. This includes announcements from management about "not worrying about it." (158)

Downfall_newsfeed 17 -- Wire stories about Enron's demise. (48)

Broadband 18 -- Enron Broadband Services (EBS) Enron's failed Broadband (high speed cable to deliver entertainment) venture. (26)

Federal_gov 19 -- General information about Federal government that does not specifically mention California. (85)

FERC_DOE 20 -- General information about the Federal Energy Regulatory Commission/Department of Energy. (219)

College Football 21 -- Employee emails about college football more specifically a newsletter called TruOrange, which follows University of Texas football. (100)

Pro Football 22 -- Employee emails about professional football (The NFL), but these refer to fantasy pro football leagues, where the statistics of real players are used to play an online version of football. (6)

India_General 23 -- General information about the India energy issues.[2] (38)

India_Dabhol 24 -- Specific references to India Dabhol Power Company (DPC), the Maharastra State Electricity Board (MSEB), and the Indian province of Maharastra. (79)

Nine_eleven 25 -- The terrorist attack of September 11, 2001. Mostly newscasts and updates. (29)

Nine_Eleven_Analysis 26 -- Aftermath analysis (political and economic) resulting from the attack. (30)

Dynegy 27 -- This company was a competitor of Enron. They almost purchased Enron in Oct-Nov. 2001, but let Enron plummet into bankruptcy instead. (7)

Sempra 28 -- A utilities company that works with Enron. (16)

Duke 29 -- Emails about Duke Energy. (17)

El Paso 30 -- Emails about El Paso Energy/Pipeline Company. (34)

Pipelines 31 -- General pipeline management. Note that pipelines are important part in transporting energy from one place to another. Enron's original business was a pipeline business. (17)

World_energy 32 -- A general category about energy with one or more specific geographic locations (such as Asia, Africa) that is not about India. (25)

[2] *Historical Note about India topics.* Enron had a major investment in a business venture to build a power plant in Dahhol India in the province of Maharastra. Known as a $20 billion white elephant, Enron's involvement began in 1992 and still lingered on into the new millennium.