The Enron Corpus: Obstacles Before Opportunities

Richard Careaga

2016-06-11

Introduction

Enron Corporation was an energy trading operation whose principal business was buying and selling electric and gas futures. It failed spectacularly in 2001 and both regulatory agencies and the bankruptcy courts conducted extensive investigations.

One of those agencies, the Federal Energy Regulatory Commission (FERC), focused on Enron's market activities in California's day-ahead energy market. FERC was investigating whether Enron had manipulated market prices to its advantage through its trading strategies.

As part of its investigation, FERC identified executives and employees whose emails might provide evidence of this wrongdoing. There were 150 email accounts that FERC required Enron to provide all records for. Due to some account name changes and duplicate accounts, this represents 148 individuals. The term for these people is *custodians*.

Fraud

In the late 1990s, California substantially restructured its regulation of electric utilities. Vertically integrated utilities were required, with certain exceptions, to divest generating plants and all electricity purchases and sales were required to be executed through a central exchange, with market prices set on a daily basis.

As a legal concept, *fraud* has a variety of meanings. Enron's activities of concern to FERC were regulatory violations somewhat removed from the concept of making misrepresentations with the intent that the recipient reasonably rely on them and suffer harm as a result.

By contrast, the way in which Enron reported its financial results to investors by failing to disclose its use of certain accounting devices to make its financial position look healthier than it actually was *did* qualify.

The Securities and Exchange Commission (**SEC**) was responsible for this aspect of the investigation of Enron and its outside accountants. For its purposes, the definition of fraud was *a misstatement of material fact or the omission of a statement of fact necessary to make any other statement not misleading in the circumstances under which it was made.*

The Enron corpus derives from FERC's selection of custodians. Those custodians were believed to have information related to regulatory violations in its trading operation. Some of them may also have been participating in the areas of concern to the SEC. However, the Enron corpus is primarily related to the trading activities, not the accounting activities, that were problematic.

Understanding the purpose for which the Enron corpus was collected is key to developing an approach to analysis.

Another Era and the Fundamental Character of Email

Around the turn of the century, email was a relatively new feature of corporate life, regarded as a convenience for reaching large numbers with a single message and an alternative to leaving notes on colleagues' desks or circulating newsletters. It was not the sole channel of communication.

Because email became popular first for personal use, its stylistic conventions radically differ from traditional business written communication. Reviewing email is much more akin to eavesdropping than it is to reading. Email style, and content, depend critically on the relative status of sender and recipient, the specialized vocabulary, the slang, jargon and shorthand of the participants.

Metadata Connects Pieces, But Only If ...

Enron used two email systems, IBM Notes and Microsoft Outlook. Each uses proprietary formats rich in metadata. The form in which FERC accepted the email it demanded was not in these native formats. As a result, much of the metadata is lost. The results are essentially a printout in editable form.

The two systems require different janitorial treatment. Both systems' results are essentially ASCII, rather than UTF-8. Diacritical marks have been obliterated.

The files are organized by custodian. The same email, therefore, may appear in the custodian's sent folder, one recipient's inbox and another recipient's archive. Some custodians are packrats, saving everything, and others are pruners, keeping only select emails.

During the 2000s, many corporate IT departments imposed quotas on email storage, which makes finding emails with relevant content something of a crapshoot.

90% of Everything Is Dreck

Of the approximately 500,000 emails in the Enron corpus, approximately half are duplicates. A large part of the remainder consists of newsletters, bulletins, fantasy football matters and other emails addressed to a large audience with only one or a few Enron recipients. Say, 125,000. The remaining 125,000 have roughly 75,000 addressed to large groups ("be advised that the Houston gym hours will be changing") or to individuals on routine matters ("your approval for expense report #1234 is overdue"). Another 25,000 deal with scheduling of meetings, transmission of periodic reports, and circulation of form documents covering derivative trading with counterparties.

Of the remaining emails, approximately 15,000 involve correspondence from a sender to a custodian who never sends a reply or an original email to the sender. This leaves about 35,000 emails among senders and receivers who engage in some degree of reciprocal correspondence. This is where to begin.

Don't Look for Much from the Big Shots

CEO Ken Lay's administrative assistant handled much of his email. CFO Andy Fastow's email is not included. COO Jeff Skilling is included, but his volume is small.

Keywords Aren't Much Help

First, it is essential to have a well-thought-out file of stop words. Second, NLP approaches based on written composition are of little help in the misspelled, ungrammatical, free-flowing, implicitly rich world of email.

Emails Are Not the Universe

The business of Enron was trading, and the tool of trading is and was the Bloomberg terminal with its instant messaging feature. The conduct of a trading floor, more than most other large corporate enterprises, is effectuated face-to-face and by telephone. Don't look for meeting agendas and minutes.

The Best Use of the Enron Corpus Is Tracing Relationships

Starting with penpals, people who regularly exchange one-on-one emails, it is possible to infer the functional organization of an organization like Enron. Think of it like whispering. Going beyond that, avoid the spoke-and-wheel representations that you so often see. Most of those connections end up being broadcast email.

Don't Neglect Time Series

In looking at a subset, pay attention to how volume varies with time. As others have noted, a sudden drop in email volume within the senior group, such as occurred in May 2001, can indicate a situation in which decisionmakers are meeting personally. When trouble looms, people clam up.

Avoid the Echo Chamber

Email streams in which an initial email goes out to a group with responses coming back quoting the original email, forwards and re-forwards and responses, amplify whatever NLP content can be gleaned. Parse each email to determine its original content load and delete that from replies and forwards, as well as standard disclaimers.

The Real Value Proposition

The purpose of examining a huge body of email is not to find a smoking gun. It is to understand the business, its language and its people. While you may sometimes run across braggadocio crowing over putting one over, those are rare and seldom determinative.

How I Know This

After being let go from my former position of Senior Vice President and Associate General Counsel for Washington Mutual Bank by its new owner, I looked to the Enron corpus to understand how the emails I wrote and received in connection with Washington Mutual's securitization business would be examined and what questions I should be prepared to answer in the several lawsuits in which I was a defendant in purported class-actions seeking many more billions in damages than I had lying around. Through a combination of rusty bash skills, MySQL, R and Python, I spent two years analyzing the Enron corpus. Scrubing the data was tedious but relatively simple; the keyword searching and natural language processing was where I spent the most time. During my depositions, I learned that plaintiff's counsels' reliance on keyword searches had diverted it from many potentially more fruitful lines of questioning.