Outline of eDiscovery paper

Richard Careaga

February 8, 2019

Introduction

Goal

The goal of this paper is to illustrate a combination of machine learning, natural language processing and graph analysis techniques applied to corporate email to identify potential witnessess in litigation.

Background

> In times of political turmoil, events can move from impossible to inevitable without even passing through improbable.

Anatole Kalesky

Enron Corp. and its affiliates were engaged in energy-related businesses, as described in its Annual Report on Form 10-K for the year ended December 31, 2000

```
*    the transportation of natural gas through pipelines to
markets throughout the United States;

*    the generation, transmission and distribution of
electricity to markets in the northwestern United States;

*    the marketing of natural gas, electricity and other
commodities and related risk management and finance services
worldwide;

*    the development, construction and operation of power
plants, pipelines and other energy related assets worldwide;

*    the delivery and management of energy commodities and
capabilities to end-use retail customers in the industrial
and commercial business sectors; and

*    the development of an intelligent network platform to
provide bandwidth management services and the delivery of
high bandwidth communication applications.

As of December 31, 2000, Enron employed approximately
20,600 persons.
```

For that year it had operating revenues of $100,789 million, according to the same report.

On December 2, 2001, Enron filed for bankruptcy protection.

In less than a year, Enron underwent a complete reversal of fortune as its business strategies ran afoul of applicable regulations. The Federal Energy Regulatory Commission (**FERC**) was one of those regulators.

FERC became aware of irregularities in the California wholesale electricity market prices. An orientation to the issues is provided by the testimony before FERC provides a concise summary.[1]

Following Enron's bankruptcy, FERC began an intense investigation, including the email records of 149 Enron employees. A preliminary staff report issued six months later.

Motivating Data

FERC obtained approximately 500,000 emails. Copies of these were acquired by Leslie Kaelbling of MIT and published by William W. Cohen of Carnegie Mellon University. It is one of the largest publicly available datasets of corporate email and is referred to as the Enron Corpus. The term *corpus* is used in natural language processing to denote a collection of related text.

At the time, electronic record examination (*ediscovery*) in litigation was in a primative state. It was not uncommon, for example, for paper copies of email to be offered. These would typically be read by teams of freelance attorneys looking for keywords. Advanced technology included scanning with optical character recognition and some proprietary software options to organize emails and capture the status of review.

Much of the focus was directed to keyword searches, sometimes called the *smoking gun* approach. Brute force examination misses opportunities to understand the social networks that reflect how the organization operates, what their concerns are and which part of the corpus should receive priority. To do that the corpus must be distilled and analyzed.

Analysis

Data acquisition

I obtained a copy of the 2009 version of the corpus. It contains copies of emails of a private nature that involve three users have since requested to be redacted. I have removed those 27 emails.

Conversion

Each email was a plaintext file[2] Each user had a directory tree similar to the one below.[3]

#[Typical user data tree] #(https://s3-us-west-2.amazonaws.com/dslabs.nostromo/dtree.jpg)

While tedious, traversing the directory tree, parsing the emails and loading them into an SQL database, was accomplished with a combination of Python and Perl scripting and standard bash tools. I do not reproduce that process here as it has little bearing on the main topic of this paper.

Data structure

While the emails were not in native format, the plain text versions contained nine principal segments, as shown in the figure below

Of those, the following were extracted:

- sender

[1] The short version, which I can relate as a former California regulatory official from personal knowledge, is that public electric utilities were losing a large share of industrial customer to self-generation. Many businesses found it cheaper to generate on than to pay tariff rates. Foreseeably, residential and business customers without the option to self-generate would bear the entire cost of utility fixed assets, and rates would increase. The solution was to require the utilities to sell their generation plants and buy power on a new public market on a "day-ahead" basis, tomorrow's estimated demand. Although much thought was devoted to the dangers of participants gaming the system to sell or buy at discounts for market, insuffient consideration was given to multi-participant cooperation.

[2] Most had been generated by Microsoft Outlook, but some older emails were producted in IBM Notes, which created some character encoding issues.

[3] This user had 10 directories with 3048 files (the directory tree has been pruned to omit spurious detail) containing 12,147 lines and 69,226 words.

- date
- primary recipient(s)
- cc recipient(s)
- subject line
- message body
- file name from directory tree

Data augmentation

The imported data and is augmented fields are as follows:

Field	Type	Description
body	mediumtext	entire email less metadata
payload	mediumtext	the original message in email chain
hash	varchar(250)	an md5 hash of payload
sender	varchar(250)	email address of sender
tos	text	direct recipient(s) email address(es)
mid	varchar(250)	message identification metadata
ccs	text	copied recipient(s) email address(es)
date	datetime	date metadata
subj	varchar(500)	subject line of email
tosctn	mediumint(9)	number of direct recipients
ccsctn	mediumint(9)	number of copied recipients
source	varchar(250)	path of directory from which extracted

Deduplication

The payload field hash, an md5 encoded message digest[^In theory, it is possible that two non-identical sequences of bytes be encoded identically, the probability is low enough to make an md5 digest usable as a checksum verification, its purpose here.] was used as a primary key to assure the uniqueness of each record. Approximately half of the corpus consisted of duplicates, such as the original message in the sender's sent file and one or more copies in the recipient's inbox, at a minimum. Multiple recipients and recipients who used email folders as a filing system were another source of duplicate messages.

Text isolation

For natural language processing (**NLP**) purposes, treating the `payload` rather than the `body` as the unit of analysis avoided an "echo chamber" effect of threads quoting and requoting the original message, multipling the frequency of the words it contained.

Prioritization

Analysis of emails in an arbitrary manner delays reducing its collective informational value in preliminary analysis. Prioritizing always leaves open the option of reviewing the set-asides later.

The first filter applied was to eliminate all email from external addresses that were not also recipients from internal addresses. Spam, newsletters and the like have low information potential. This filter reduced the remaining half of the corpus by half again, leaving approximately 125,000 emails.

A second filter for internal email was used to eliminate broadcast messages and high frequency administrative messages. Indicia of broadcast messages were large numbers of recipients, high frequency, paucity of return correspondence and keyword in context screening. Administrative messages to single recipients were identified

by frequency, lack of return correspondence and high frequency words. Many of these were nagging emails concerning the lack of approval of expense reports, for example. This filter again halved the corpus, to approximately 75,000 emails.

The third filter eliminated internal broadcast emails. This filter reduced the email count to approximately 24,000.

The final filter limited the dataset to emails sent before Enron's December 2, 2001 bankruptcy. This filter reduced the email count to approximately 13,500, about 2.7% of the original total.[4]

Resulting dataset

For purposes of this paper, the SQL database has been serialized to an `Rda` file of approximately 10MB size, which can be brought into an `R` session with the following

```
# syntax to load Rda from S3

con <- url("https://s3-us-west-2.amazonaws.com/dslabs.nostromo/enron.Rda")
load(con)
close(con)
```

Information sought

I downloaded the [staff] report and extracted the 6-page Executive Summary.[5] From this, I extracted a list of keywords that indicated the general nature of FERC's inquiry. I deliberated excluded the extensive detail available in the remainder of the report. At the beginning of most ediscovery, the parties seeking review of electronic documents have topics in view, which the Executive Summary provides without biasing the analysis with the results of FERC's review.

```
library(dplyr)
```

```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library(pdftools)
library(stringr)
library(tidytext)
library(ggplot2)

data(stop_words)

target_pdf <- ("/sources/ExecutiveSummaryAug2003Staff.pdf")
#exec_summary <- pdf_text(target_pdf)
```

character

[4] *Ninety percent of everything is crap.* Theodore Sturgeon's Revelation, made in a dominantly paper-based information environment. *See also* Pareto distributions.

[5] This could, and would have been coded in `R` had multiple files been required.