University of California-Berkeley

Statistics Undergraduate Research Project

Exploratory Data Analysis of Enron Emails

Author:
Harish Kumar Palaniswamy

Supervisor:
Prof. David Aldous

May 15, 2015

Abstract

Enron Corporation was an American energy, commodities, and services company based in Houston, Texas. Before its bankruptcy on December 2, 2001, Enron employed approximately 20,000 staff and was one of the world's major electricity, natural gas, communications, and pulp and paper companies, with claimed revenues of nearly $111 billion during 2000. At the end of 2001, it was revealed that its reported financial condition was sustained substantially by an institutionalized, systematic, and creatively planned accounting fraud, known since as the Enron scandal. Enron has since become a well-known example of wilful corporate fraud and corruption. This report aims at answering whether top level Enron employees had incriminating evidence in their office emails or uncover any unusual patterns in the months leading up to the scandal through an exploratory data analysis.

1 Introduction

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 0.5 million messages. This data was originally made public, and posted to the web, by the Federal Energy Regulatory Commission during its investigation. The dataset consists of 517,431 messages that belong to 150 users, mostly senior management of the Enron Corp. Although the dataset is huge, topical folders of particular users are often quite sparse. For our purposes, we only look at sent emails and ignore the inboxes of all the employees. Through this approach, we can avoid accidentally analysing the spam emails that are among the received emails. Two main methods of analyses were employed, namely, topic modelling with Latent Dirichlet Allocation(LDA) and sentiment analysis.

2 Data Processing

Before beginning any sort of data analysis on the emails, it is important to pre-process all text in the emails first. Below is an example of raw email text.

```
Message-ID: <5525962.1075855679785.JavaMail.evans@thyme>
Date: Wed, 13 Dec 2000 07:04:00 -0800 (PST)
From: phillip.allen@enron.com
To: christi.nicolay@enron.com, james.steffes@enron.com, jeff.dasovich@enron.com,
joe.hartsoe@enron.com, mary.hain@enron.com, pallen@enron.com,
pkaufma@enron.com, richard.sanders@enron.com,
richard.shapiro@enron.com, stephanie.miller@enron.com,
steven.kean@enron.com, susan.mara@enron.com,
rebecca.cantrell@enron.com
Subject:
Mime-Version: 1.0
Content-Type: text/plain; charset=us-ascii
Content-Transfer-Encoding: 7bit
X-From: Phillip K Allen
X-To: Christi L Nicolay, James D Steffes, Jeff Dasovich, Joe Hartsoe, Mary Hain, p
X-cc:
X-bcc:
X-Folder: \Phillip_Allen_Dec2000\Notes Folders\Sent
X-Origin: Allen-P
X-FileName: pallen.nsf

Attached  are two files that illustrate the following:

As prices rose, supply increased and demand decreased.  Now prices are
```