Research Proposal

January 29, 2015 Harish Kumar Palaniswamy,
SID:23154182

Exploratory Data Analysis of Enron Emails

Validity, Source & Description

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 This data was originally made public, and posted to the web, by the Federal Energy Regulatory Commission during its investigation.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. Analysis can be conducted by focusing on users who directories are especially large, namely, Sally Beck (Chief Operating Officer), Darren Farmer (Logistics Manager), Vincent Kaminski (Head of Quantitative Modeling Group), Louise Kitchen (President of EnronOnline), Michelle Lokay (Administrative Assistant), Richard Sanders (Assistant General Counsel) and Williams III (Senior Analyst).

Background

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 willful corporate fraud and corruption.

Investigation

A multitude of interesting questions can be asked about this dataset. One could investigate what the top employees of Enron were emailing each other about in the

days leading up to the scandal. Using Latent Dirichlet Association, one could extract the main topics that the emails contained and also even investigate whether these topics changed over time. With some link-analysis algorithms, one could evaluate relationships between employees and also investigate potentially changing relationships over time.

References

Kessler, G. (n.d.). Virtual business: An Enron email corpus study. Journal of Pragmatics, 262-270.

Keila, P., & Skillicorn, D. (n.d.). Structure in the Enron Email Dataset. Computational and Mathematical Organization Theory, 183-199.