MITOCW | MIT15_071S17_Session_5.4.02_300k

In this recitation, we're going to talk about predictive coding-- an emerging use of text analytics in the area of criminal justice.

We'll start with the story of Enron, the United States energy company based out of Houston, Texas that was involved in a number of electricity production and distribution markets.

In the early 2000s, Enron was a hot company, with the market capitalization exceeding $60 billion, and Forbes magazine ranked it as the most innovative US company six years in a row.

Now, all that changed in 2001 with the news of widespread accounting fraud at the firm.

This massive fraud led to Enron's bankruptcy, the largest ever at the time, and led to Enron's accounting firm, Arthur Andersen, dissolving.

To this day, Enron remains a symbol of corporate greed and corruption.

Now, what Enron's collapse stemmed largely from accounting fraud, the firm also faced sanctions for its involvement in the California electricity crisis.

California is the most populous state in the United States.

And in 2000 to 2001, it had a number of power blackouts, despite having sufficient generating capacity.

It later surfaced that Enron played a key role in this energy crisis by artificially reducing power supply to spike prices and then making a profit from this market instability.

The Federal Energy Regulatory Commission, or FERC, investigated Enron's involvement in the crisis.

And this investigation eventually led to $1.52 billion settlement.

FERC's investigation into Enron will be the topic of today's recitation.

Now, Enron was a huge company, and its corporate servers contained millions of emails and other electronic files.

Sifting through these documents to find the ones relevant to an investigation is no simple task.

In law, this electronic argument retrieval process is called the e-discovery problem, and relevant files are called responsive documents.

Traditionally, the e-discovery problem has been solved by using the key research-- in our case, perhaps, searching for phrases like "electricity bid" or "energy schedule"-- followed by an expensive and time-consuming manual review process, in which attorneys read through thousands of documents to determine which ones are responsive.

However, predictive coding is a new technique, in which attorneys mainly label some documents and then use text analytics models trained on the manually labeled documents to predict which of the remaining documents are responsive.

Now, as part of its investigation, the FERC released hundreds of thousands of emails from top executives at Enron creating the largest publicly available set of emails today.

We will use this data set called the Enron Corpus to perform predictive coding in this recitation.

Our data set contains just two fields-- email, which is the text of the email in question, and responsive, which is whether the email relates to energy schedules or bids.

The labels for these emails were made by attorneys as part of the 2010 text retrieval conference legal track, a predictive coding competition.