PREDICT 420

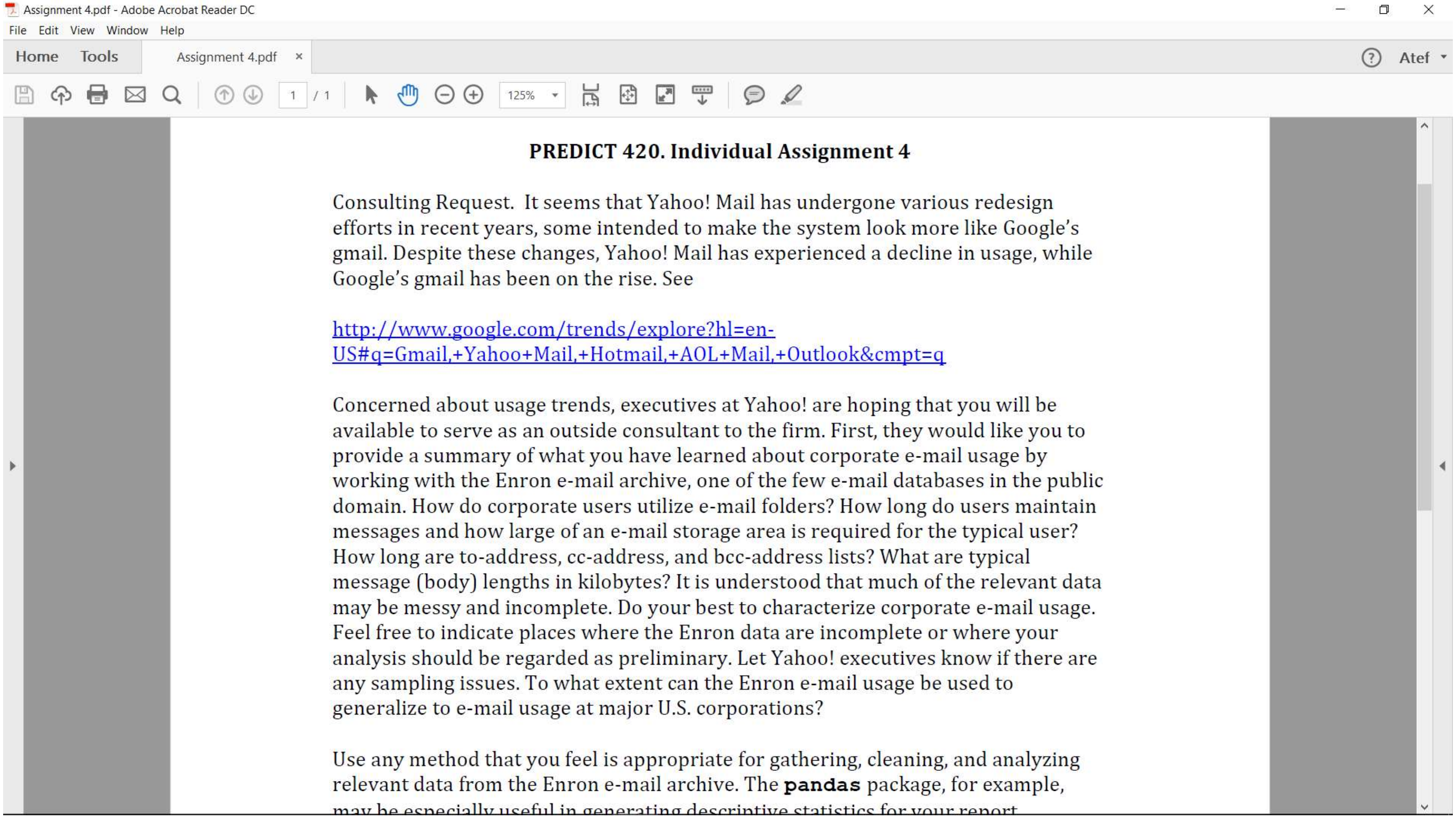
Atef Bader, PhD

Agenda

- Assignment #4 - Walkthrough
  - Due at the end of this week (week #6)

- Exercise #6 - Walkthrough
  - Due at the end of this week (week #6)

Assignment #4 - Deliverable


Assignment 4.pdf - Adobe Acrobat Reader DC
File Edit View Window Help
Home Tools
Assignment 4.pdf ×
Atef
1 / 1
125%

PREDICT 420. Individual Assignment 4

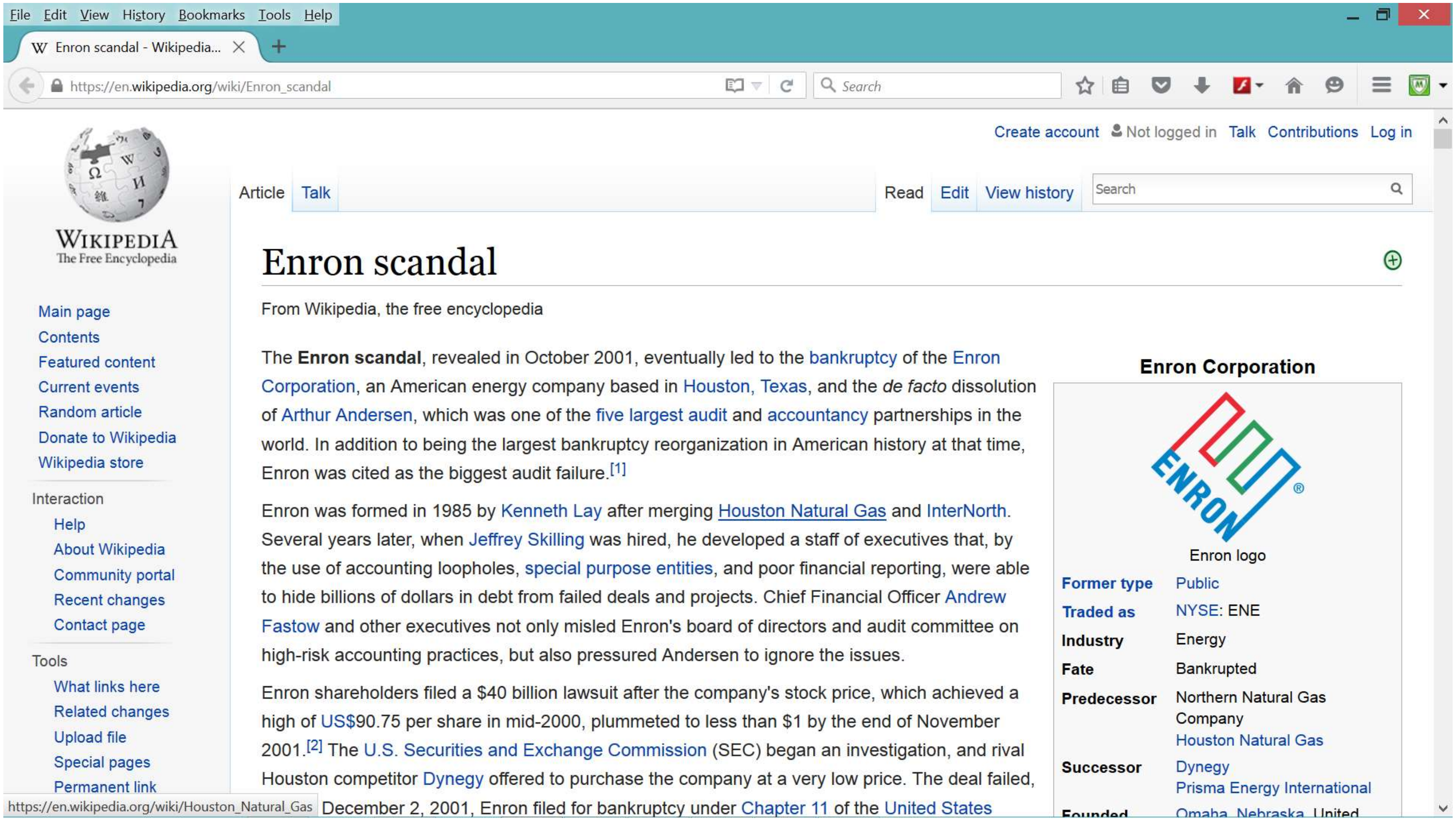
Consulting Request. It seems that Yahoo! Mail has undergone various redesign efforts in recent years, some intended to make the system look more like Google's gmail. Despite these changes, Yahoo! Mail has experienced a decline in usage, while Google's gmail has been on the rise. See

http://www.google.com/trends/explore?hl=en-US#q=Gmail,+Yahoo+Mail,+Hotmail,+AOL+Mail,+Outlook&cmpt=q

Concerned about usage trends, executives at Yahoo! are hoping that you will be available to serve as an outside consultant to the firm. First, they would like you to provide a summary of what you have learned about corporate e-mail usage by working with the Enron e-mail archive, one of the few e-mail databases in the public domain. How do corporate users utilize e-mail folders? How long do users maintain messages and how large of an e-mail storage area is required for the typical user? How long are to-address, cc-address, and bcc-address lists? What are typical message (body) lengths in kilobytes? It is understood that much of the relevant data may be messy and incomplete. Do your best to characterize corporate e-mail usage. Feel free to indicate places where the Enron data are incomplete or where your analysis should be regarded as preliminary. Let Yahoo! executives know if there are any sampling issues. To what extent can the Enron e-mail usage be used to generalize to e-mail usage at major U.S. corporations?

Use any method that you feel is appropriate for gathering, cleaning, and analyzing relevant data from the Enron e-mail archive. The **`pandas`** package, for example, may be especially useful in generating descriptive statistics for your report.

Assignment #4 - Deliverable



Enron scandal

From Wikipedia, the free encyclopedia

The **Enron scandal**, revealed in October 2001, eventually led to the bankruptcy of the Enron Corporation, an American energy company based in Houston, Texas, and the *de facto* dissolution of Arthur Andersen, which was one of the five largest audit and accountancy partnerships in the world. In addition to being the largest bankruptcy reorganization in American history at that time, Enron was cited as the biggest audit failure.[1]

Enron was formed in 1985 by Kenneth Lay after merging Houston Natural Gas and InterNorth. Several years later, when Jeffrey Skilling was hired, he developed a staff of executives that, by the use of accounting loopholes, special purpose entities, and poor financial reporting, were able to hide billions of dollars in debt from failed deals and projects. Chief Financial Officer Andrew Fastow and other executives not only misled Enron's board of directors and audit committee on high-risk accounting practices, but also pressured Andersen to ignore the issues.

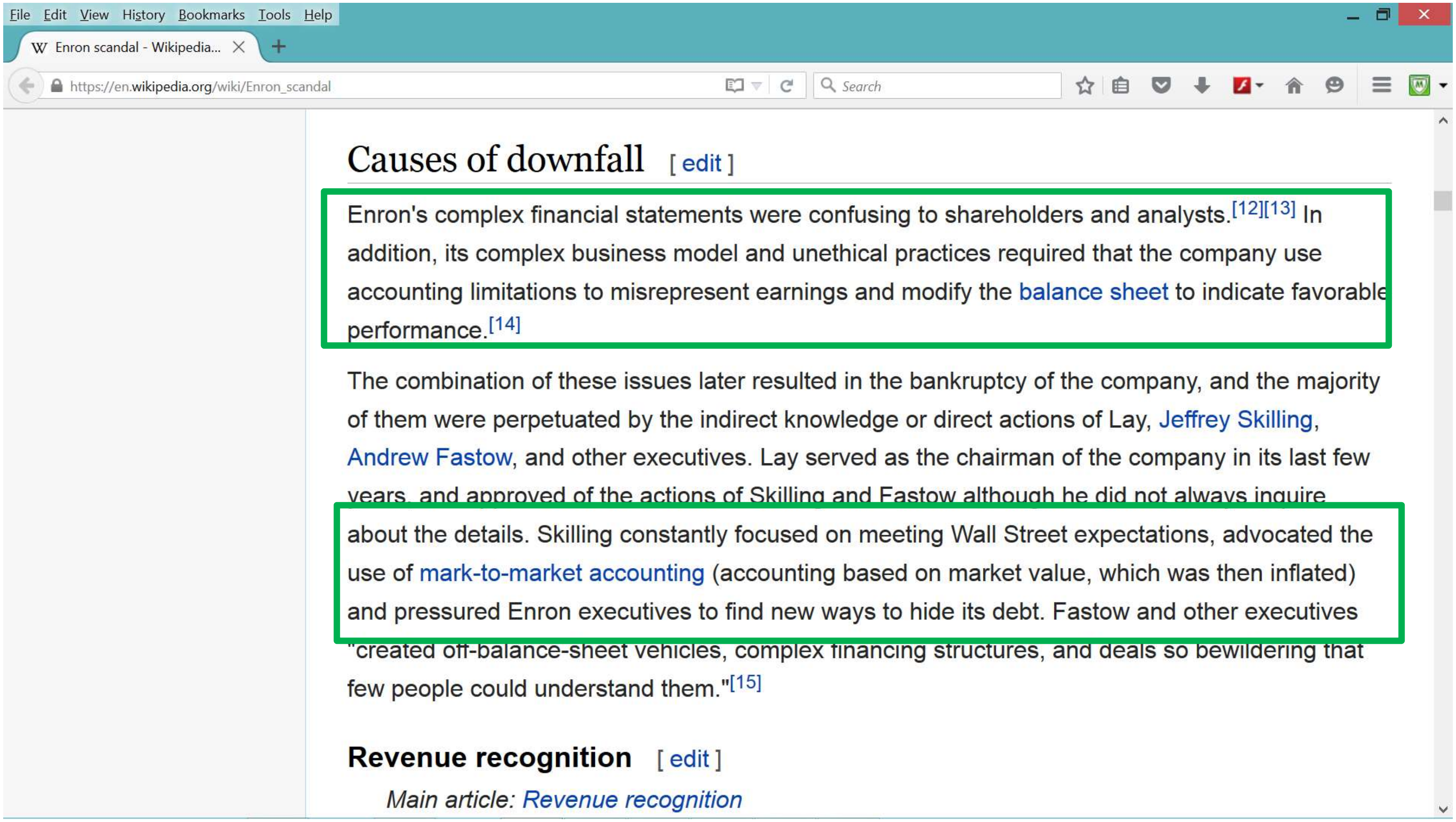
Enron shareholders filed a $40 billion lawsuit after the company's stock price, which achieved a high of US$90.75 per share in mid-2000, plummeted to less than $1 by the end of November 2001.[2] The U.S. Securities and Exchange Commission (SEC) began an investigation, and rival Houston competitor Dynegy offered to purchase the company at a very low price. The deal failed, December 2, 2001, Enron filed for bankruptcy under Chapter 11 of the United States

Enron Corporation

Enron logo

Former type	Public
Traded as	NYSE: ENE
Industry	Energy
Fate	Bankrupted
Predecessor	Northern Natural Gas Company Houston Natural Gas
Successor	Dynegy Prisma Energy International
Founded	Omaha, Nebraska, United

Assignment #4 - Deliverable


File Edit View History Bookmarks Tools Help
Enron scandal - Wikipedia...
https://en.wikipedia.org/wiki/Enron_scandal
Search

Causes of downfall [edit]

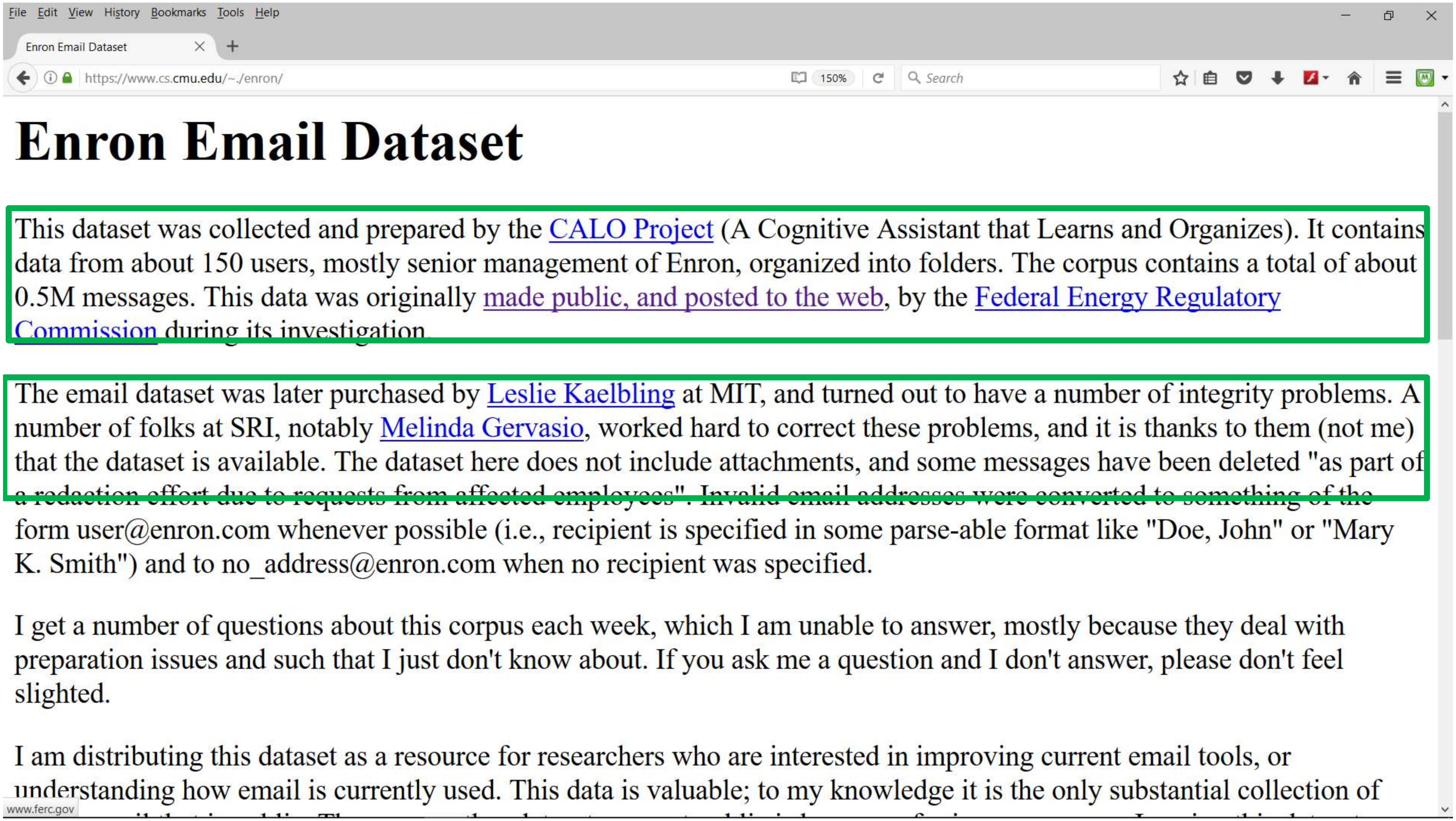
Enron's complex financial statements were confusing to shareholders and analysts.[12][13] In addition, its complex business model and unethical practices required that the company use accounting limitations to misrepresent earnings and modify the balance sheet to indicate favorable performance.[14]

The combination of these issues later resulted in the bankruptcy of the company, and the majority of them were perpetuated by the indirect knowledge or direct actions of Lay, Jeffrey Skilling, Andrew Fastow, and other executives. Lay served as the chairman of the company in its last few years, and approved of the actions of Skilling and Fastow although he did not always inquire about the details. Skilling constantly focused on meeting Wall Street expectations, advocated the use of mark-to-market accounting (accounting based on market value, which was then inflated) and pressured Enron executives to find new ways to hide its debt. Fastow and other executives "created off-balance-sheet vehicles, complex financing structures, and deals so bewildering that few people could understand them."[15]

Revenue recognition [edit]

Main article: Revenue recognition

Assignment #4 - Deliverable


File Edit View History Bookmarks Tools Help
Enron Email Dataset
https://www.cs.cmu.edu/~./enron/
150%
Search
www.ferc.gov

Enron Email Dataset

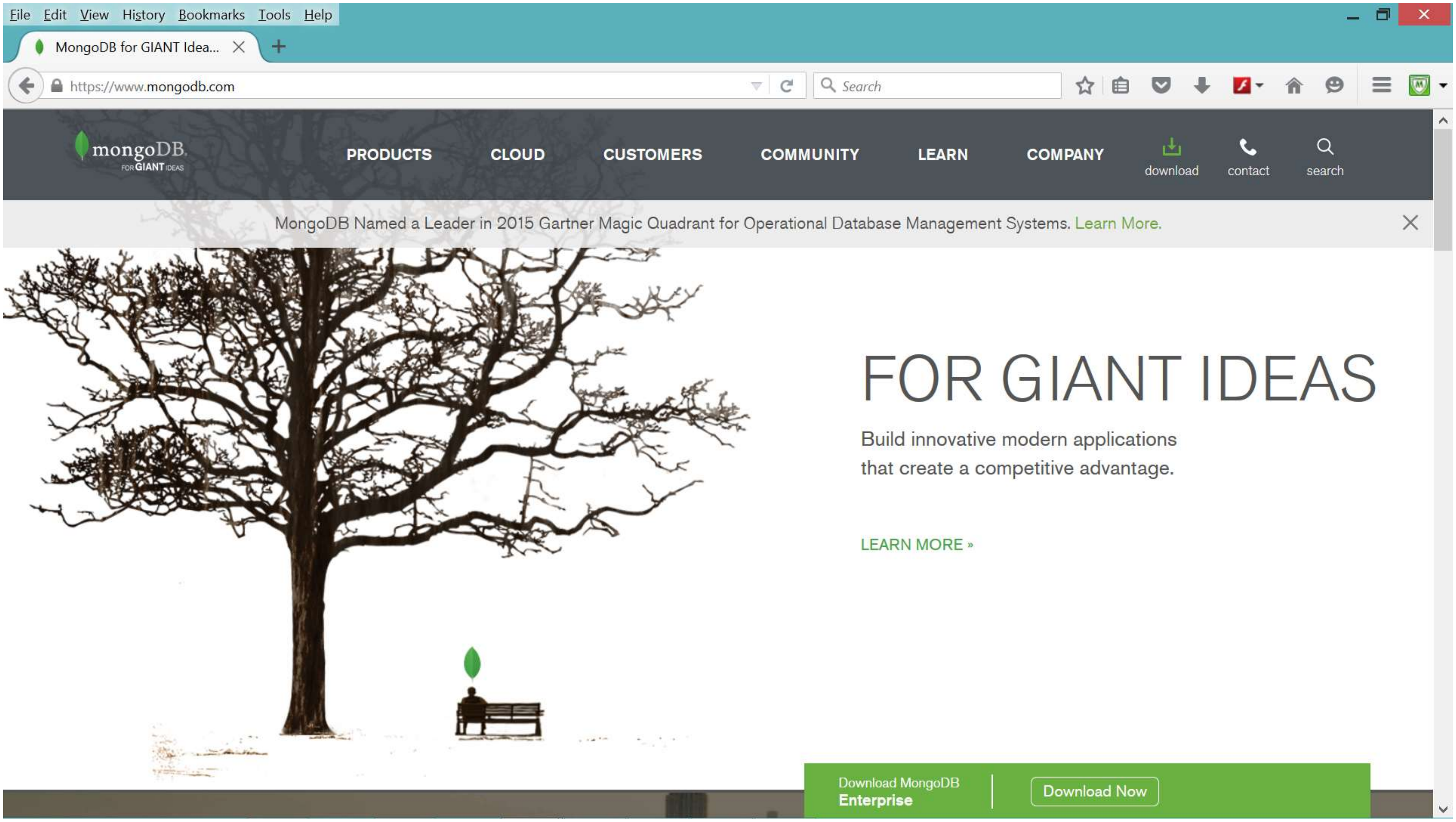
This dataset was collected and prepared by the CALO Project (A Cognitive Assistant that Learns and Organizes). It contains data from about 150 users, mostly senior management of Enron, organized into folders. The corpus contains a total of about 0.5M messages. This data was originally made public, and posted to the web, by the Federal Energy Regulatory Commission during its investigation.

The email dataset was later purchased by Leslie Kaelbling at MIT, and turned out to have a number of integrity problems. A number of folks at SRI, notably Melinda Gervasio, worked hard to correct these problems, and it is thanks to them (not me) that the dataset is available. The dataset here does not include attachments, and some messages have been deleted "as part of a redaction effort due to requests from affected employees". Invalid email addresses were converted to something of the form user@enron.com whenever possible (i.e., recipient is specified in some parse-able format like "Doe, John" or "Mary K. Smith") and to no_address@enron.com when no recipient was specified.

I get a number of questions about this corpus each week, which I am unable to answer, mostly because they deal with preparation issues and such that I just don't know about. If you ask me a question and I don't answer, please don't feel slighted.

I am distributing this dataset as a resource for researchers who are interested in improving current email tools, or understanding how email is currently used. This data is valuable; to my knowledge it is the only substantial collection of

Assignment #4 - Deliverable

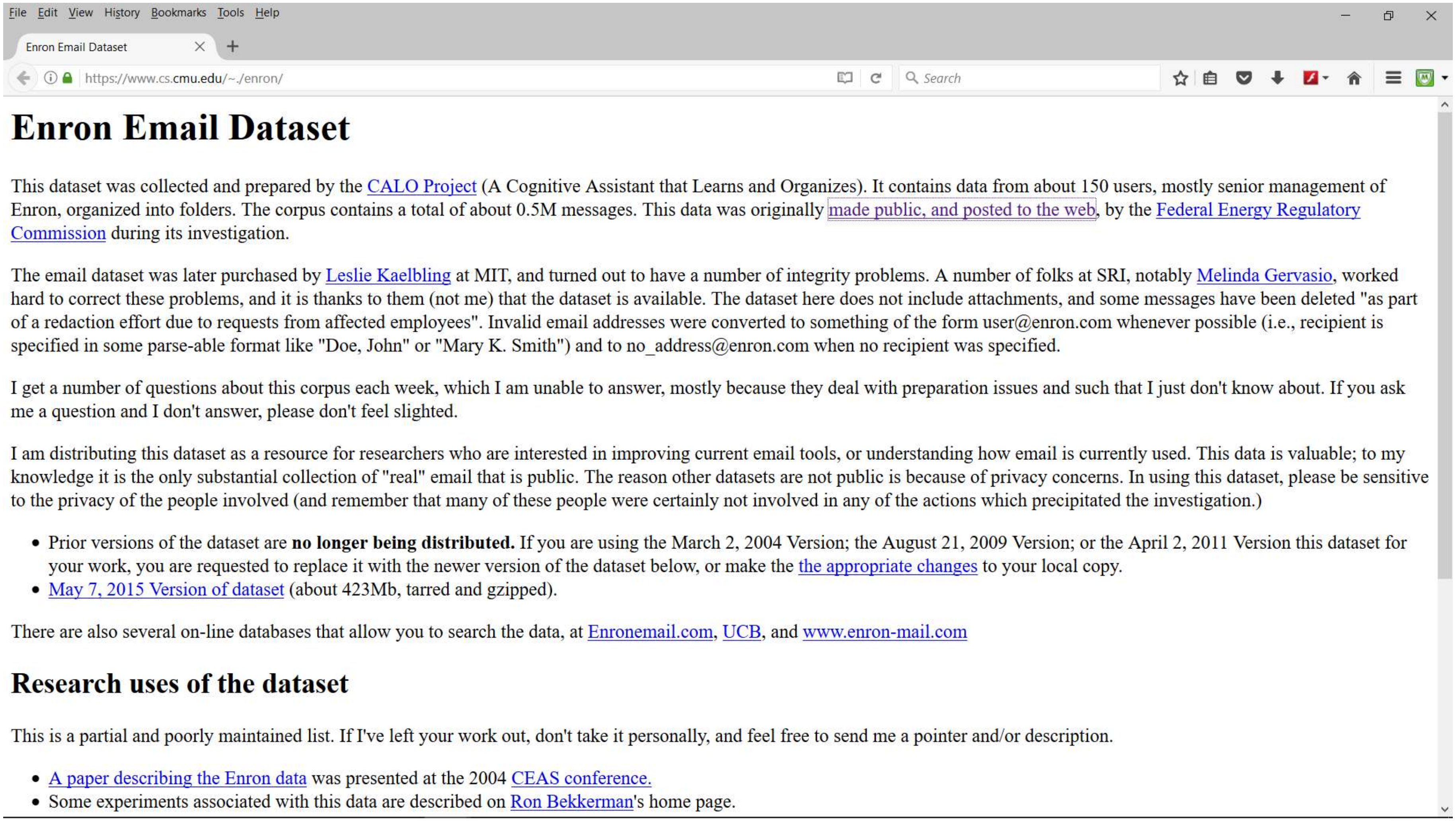


MongoDB Named a Leader in 2015 Gartner Magic Quadrant for Operational Database Management Systems. Learn More.

FOR GIANT IDEAS

Build innovative modern applications that create a competitive advantage.

Assignment #4 - Deliverable


File Edit View History Bookmarks Tools Help
Enron Email Dataset
https://www.cs.cmu.edu/~./enron/
Search

Enron Email Dataset

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 0.5M messages. This data was originally made public, and posted to the web, by the Federal Energy Regulatory Commission during its investigation.

The email dataset was later purchased by Leslie Kaelbling at MIT, and turned out to have a number of integrity problems. A number of folks at SRI, notably Melinda Gervasio, worked hard to correct these problems, and it is thanks to them (not me) that the dataset is available. The dataset here does not include attachments, and some messages have been deleted "as part of a redaction effort due to requests from affected employees". Invalid email addresses were converted to something of the form user@enron.com whenever possible (i.e., recipient is specified in some parse-able format like "Doe, John" or "Mary K. Smith") and to no_address@enron.com when no recipient was specified.

I get a number of questions about this corpus each week, which I am unable to answer, mostly because they deal with preparation issues and such that I just don't know about. If you ask me a question and I don't answer, please don't feel slighted.

I am distributing this dataset as a resource for researchers who are interested in improving current email tools, or understanding how email is currently used. This data is valuable; to my knowledge it is the only substantial collection of "real" email that is public. The reason other datasets are not public is because of privacy concerns. In using this dataset, please be sensitive to the privacy of the people involved (and remember that many of these people were certainly not involved in any of the actions which precipitated the investigation.)

- Prior versions of the dataset are **no longer being distributed.** If you are using the March 2, 2004 Version; the August 21, 2009 Version; or the April 2, 2011 Version this dataset for your work, you are requested to replace it with the newer version of the dataset below, or make the the appropriate changes to your local copy.
- May 7, 2015 Version of dataset (about 423Mb, tarred and gzipped).

There are also several on-line databases that allow you to search the data, at Enronemail.com, UCB, and www.enron-mail.com

Research uses of the dataset

This is a partial and poorly maintained list. If I've left your work out, don't take it personally, and feel free to send me a pointer and/or description.

- A paper describing the Enron data was presented at the 2004 CEAS conference.
- Some experiments associated with this data are described on Ron Bekkerman's home page.

Assignment #4 - Deliverable

- ❑ Submit your log (script) file that has the commands you executed and the output you got

Exercises #6 - Deliverable

- Submit your Ipython Notebook Script
- Show the Python code snippet you wrote