EMAIL DATA ANALYSIS

GROUP 7

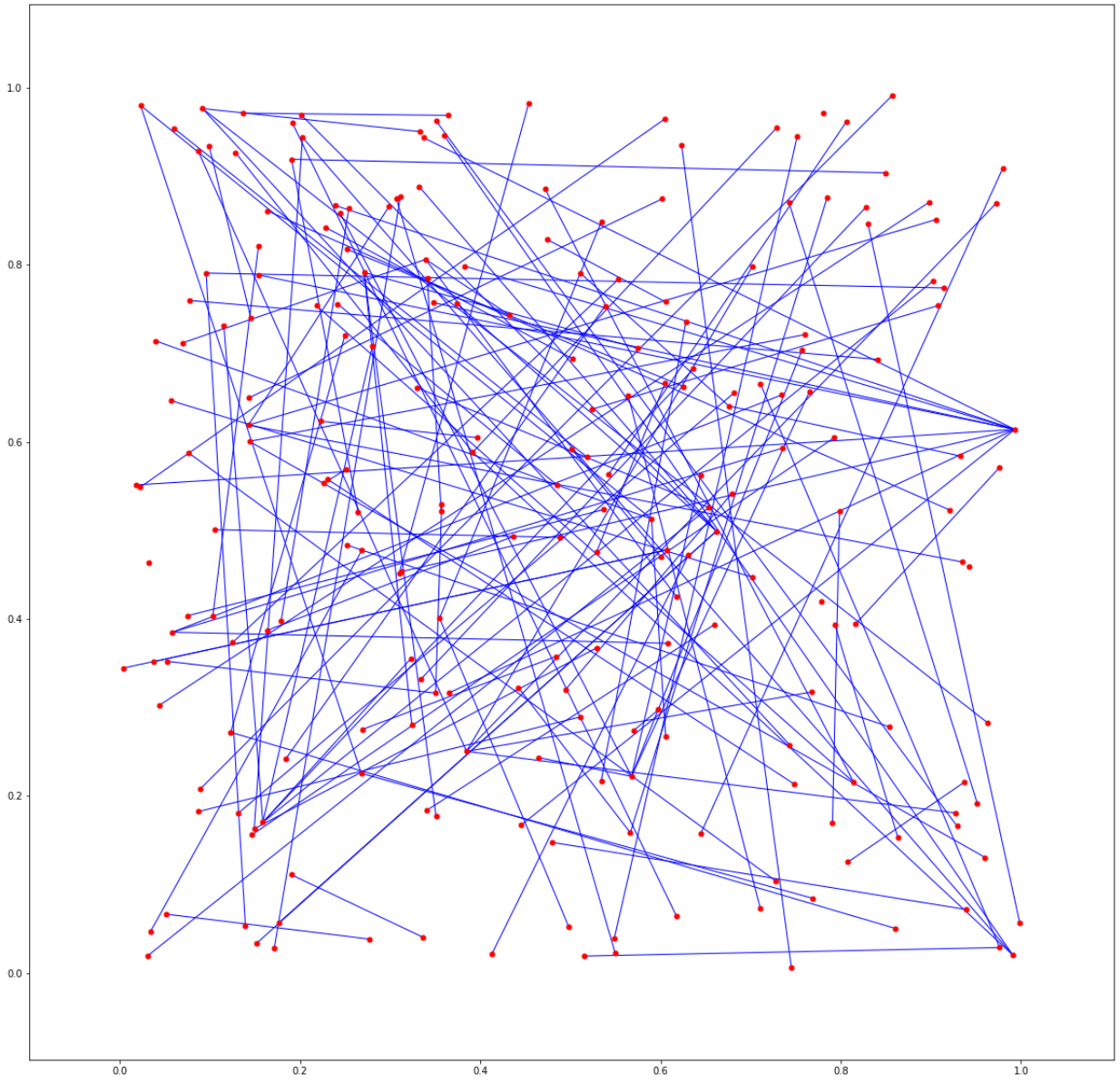
Robert Curtis
Bojue Deng
Yang Fu
Pushkar Kale
Wei Wang

https://pixabay.com/en/photos/email/

Introduction

- The Enron scandal was known to the public in late October, 2001.
- ~ 0.6m emails were left as evidence for Enron bankruptcy investigation at that time.
- Later on they became a resource for studies and research on social networking and computer analysis of language.

https://en.wikipedia.org/wiki/Enron#/media/File:Logo_de_Enron.svg



Data Exploration: the Network Plot

200 emails

random _layout

Problems of Interest

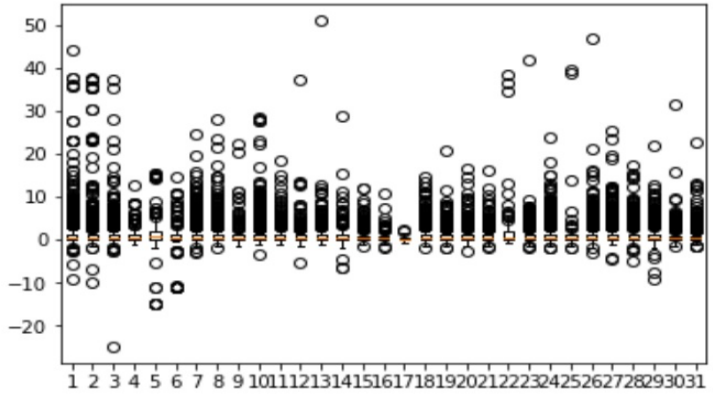
- How did Enron employees react to the scandal ?
- Was there a company-wide change of morale ?
- What did they talk about in the emails ?

Methods -- Sentiment Analysis

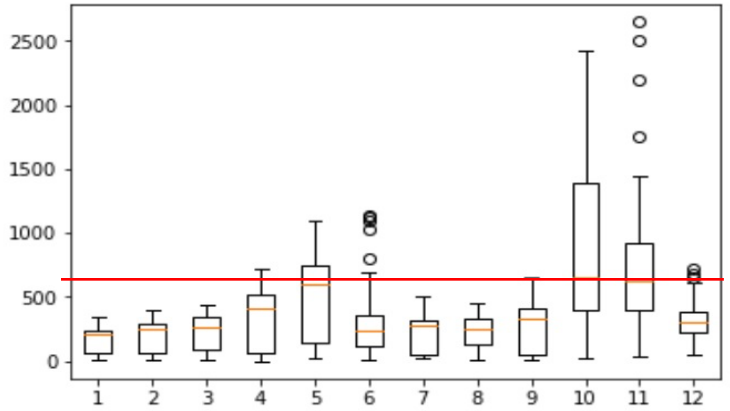
- Sentiment analysis
  - total sentiment of the email
  - average sentiment of the email
- Box Plots
  - easy display of information

Sentiment Analysis Results

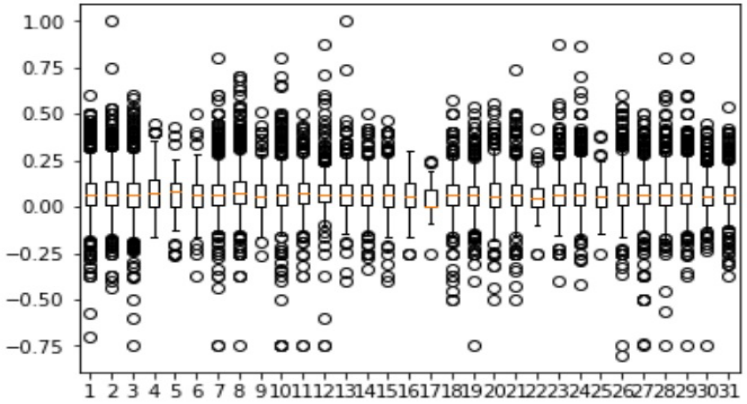
Total email sentiment by day(October 2001)



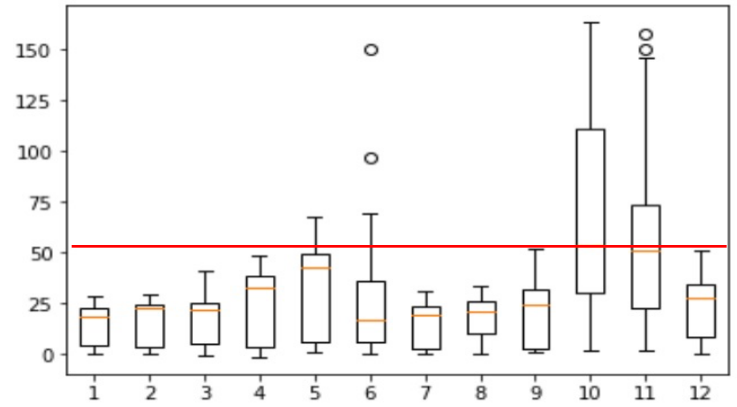
Total email sentiment by month(2001)



Averaged email sentiment by day(October 2001)



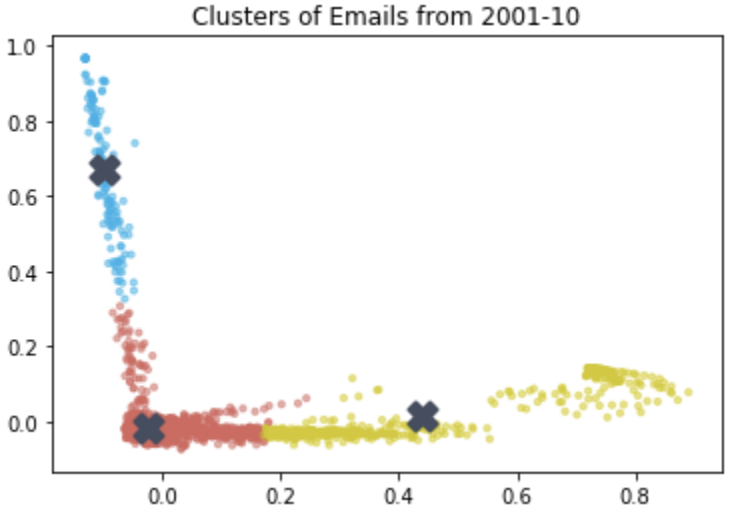
Averaged email sentiment by month(2001)

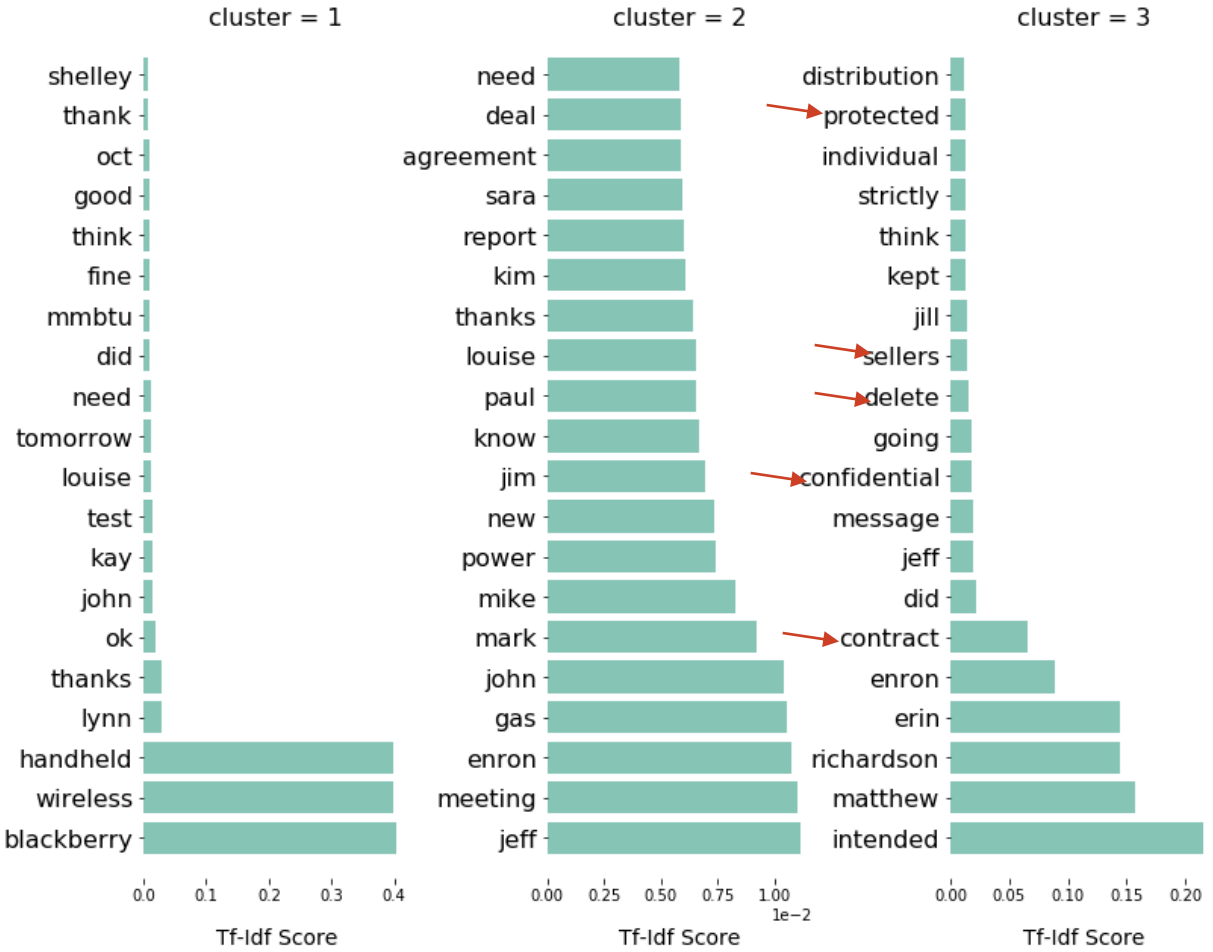


Methods -- Email Clustering

- Emails sent during Oct, 2001. # ~ 8000.
- Convert email contents to a term-document matrix using TF-IDF.
- Cluster with KMeans (K=3)

Email Clustering





Conclusions

- Emails are a great source to learn about people's emotions and feelings.

- Clustering of emails provides interesting insights on various keywords that appear in a collections of emails. Such keywords may represent people's thoughts and actions.

https://pixabay.com/en/thank-you-text-message-note-394180/

ANY QUESTIONS?

https://www.flickr.com/photos/wingedwolf/5471047557