BA810: Final Project Paper

Team 14: Tzu-Hua Huang, Kangjing Shi, Man Shi, Ziqi Shan

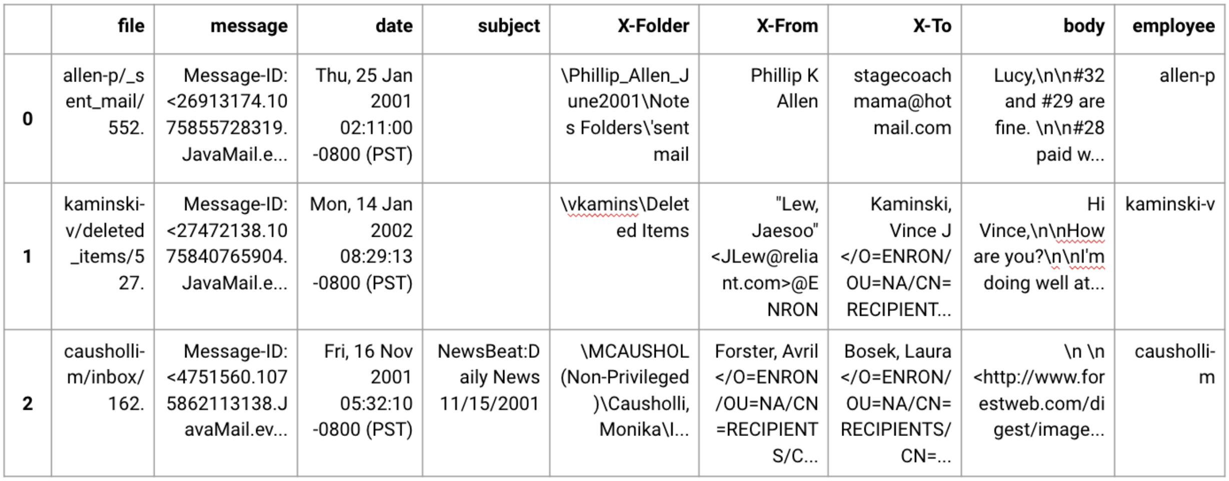
1. Business Problem

E-mail is a form of communication that provides information exchange and is the most widely used service on the Internet. It can be in various forms, such as text, images, and sounds. At the same time, users can get a large number of free news and feature emails and easily achieve information search. The existence of e-mail has greatly facilitated communication and exchange between people and promoted the development of society.

In this project, we will build a model of mails information processing system that allows it to group emails from the Enron Email Dataset. We are curious to learn from this special dataset and explore the differences and anomalies in employee social network structures around those questions. For instance, which words (top 15) have the highest frequency among 517,000 emails? How many characters of each email? How many emails were sent by each person? Who sent the most e-mails in which year? Who sent the most e-mails?

2. Dataset

The dataset was from Kaggle: https://www.kaggle.com/wcukierski/enron-email-dataset
The Enron email dataset consists of 517,401 records of emails. There are two features that comprise all the information of every single email, the 'file' column represents the employee names and the location of the email and the 'message' column contains the Message-ID, Date, Recipient and sender, the content of the message, and other information. We first transform the message into the correct format. Then we extract the columns 'date', 'subject', 'X-Folder', 'X-From', 'X-To', 'body', and 'employee' from 'message' and expand the dataset to 9 columns.

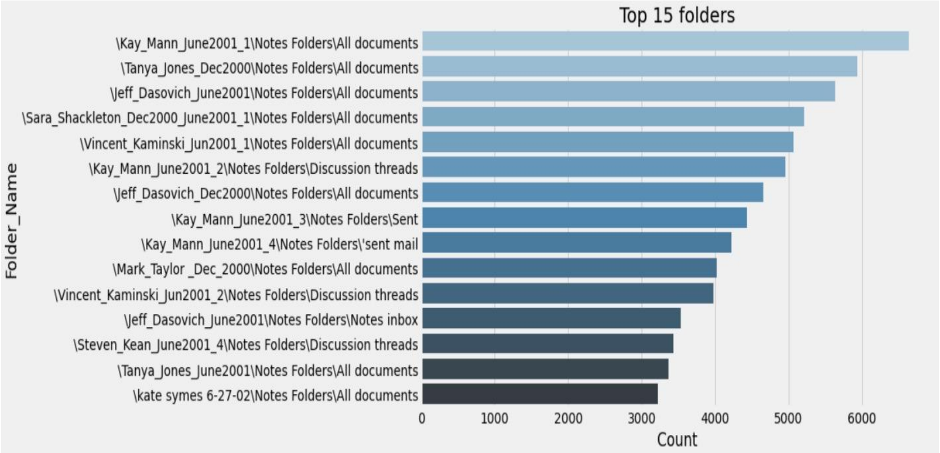


	file	message	date	subject	X-Folder	X-From	X-To	body	employee
0	allen-p/_sent_mail/552.	Message-ID: <26913174.1075855728319.JavaMail.e...	Thu, 25 Jan 2001 02:11:00 -0800 (PST)		\Phillip_Allen_June2001\Notes Folders\'sent mail	Phillip K Allen	stagecoachmama@hotmail.com	Lucy,\n\n#32 and #29 are fine. \n\n#28 paid w...	allen-p
1	kaminski-v/deleted_items/527.	Message-ID: <27472138.1075840765904.JavaMail.e...	Mon, 14 Jan 2002 08:29:13 -0800 (PST)		\vkamins\Deleted Items	"Lew, Jaesoo" <JLew@reliant.com>@ENRON	Kaminski, Vince J </O=ENRON/OU=NA/CN=RECIPIENT...	Hi Vince,\n\nHow are you?\n\nI'm doing well at...	kaminski-v
2	causholli-m/inbox/162.	Message-ID: <4751560.1075862113138.JavaMail.ev...	Fri, 16 Nov 2001 05:32:10 -0800 (PST)	NewsBeat:Daily News 11/15/2001	\MCAUSHOL (Non-Privileged)\Causholli, Monika\I...	Forster, Avril </O=ENRON/OU=NA/CN=RECIPIENTS/C...	Bosek, Laura </O=ENRON/OU=NA/CN=RECIPIENTS/CN=...	\n \n <http://www.forestweb.com/digest/image...	causholli-m

3. Exploratory Data Analysis

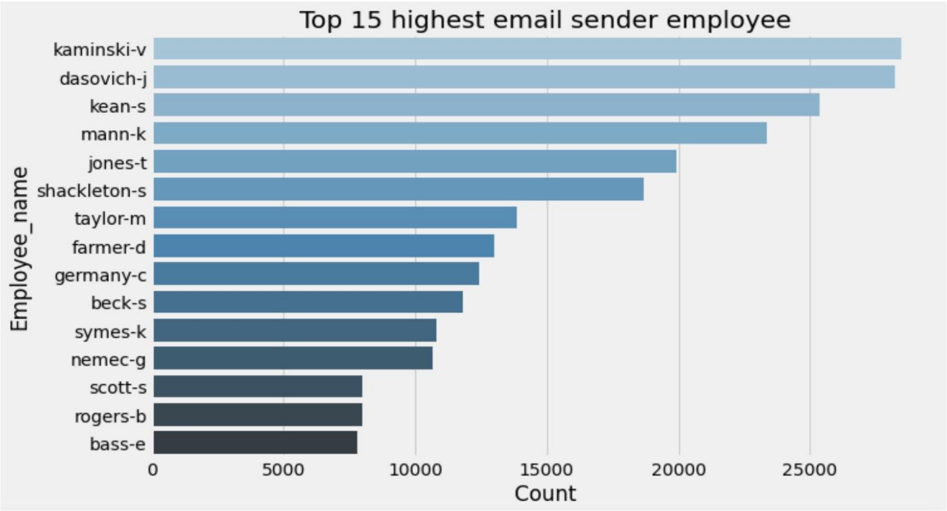
1) Top 15 Folders

First, we take a look at the Top 15 folders chart. The x-axis represents the number of emails that exist in each folder. The y axis tells us who is in charge of the folder when those emails were collected in the folder. It is obvious that most of the emails in the chart were generated between Dec 2000 and June 2001. After only four months of June, the Enron scandal came to light in October 2001. There are some clues we may assume that these persons either had some relationship with the scandal or belonged to the management departments in Enron. We will further dig into more on these folders. For example, who sent the most emails to Kay Mann? Who had received the most emails from Tanya Jones? We can draw an interesting node map around a few of them.



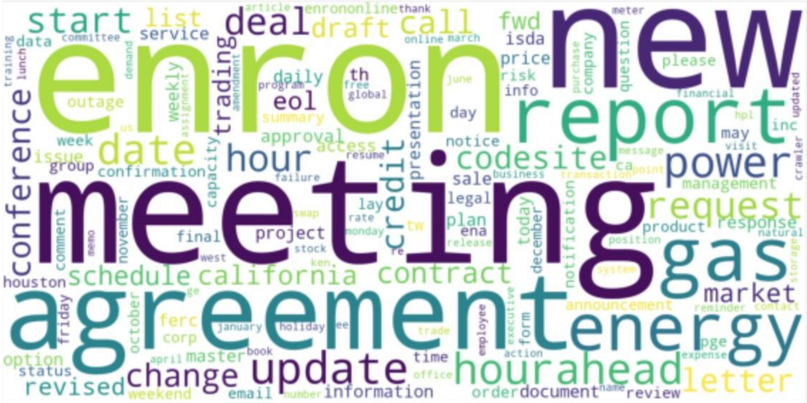
2) Top 15 highest email-sender employees

In the top 15 senders' charts, we see many familiar names from the charts above. We estimate they may belong to the top Enron executives. As we learn more about the Enron scandal, Vince J Kaminski, working as Managing Director for Research, had sent the most emails and taken one of the two most email folders in Enron. He knocked the alarm and strongly objected to the financial practices of Enron's Chief Financial Officer, Andrew Fastow, in advance of a few months before October 2001. Also, Dasovich Jeff, Enron's governmental affairs executive, and Symes was a trader, they both involved in concealing Enron's debt. We definitely will combine emails from these people to group the top Enron executive network.



3) The most frequent words

The following picture shows the most frequent words in the Enron corpus. “Enron”, “new”, “meeting” and “agreement” appear the most. It seems that they held meetings and asked for agreements very frequently.



4. Analytical Findings - Clustering Analysis on Email Messages

1) K-means clustering

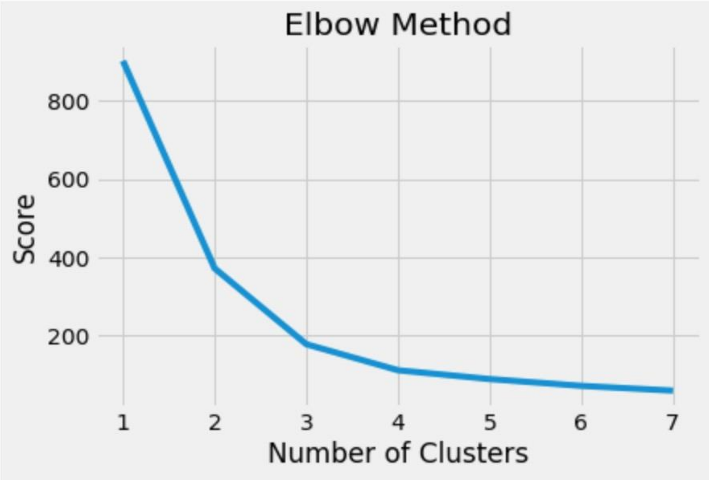
We are going to apply the K-means clustering method to the Enron email data set and segment the text in the body of emails. The reason why we chose K-means is that it's a very useful and simple method of segmenting huge amounts of text data compared to other clustering methods. Besides, it can also help give us quick insights and interpret text data. However, Hierarchical clustering is not applicable for large data sets and would cause some kernel failures, so we ended up using 10% of full data constructing the k-means model.

2) Data prepossessing

Since algorithms can hardly understand text data so we need to transform the text in the body of each email into a vector of numbers, then our algorithms will be able to understand this and proceed accordingly. What we did was use the Term Frequency-Inverse Document Frequency (TF-IDF) method to transform the text data into vectors of numbers. Then, in order to see our clusters graphically after applying K-means to our data set, the next step was to reduce the dimensionality for plotting our clusters in two dimensions by using the PCA method.

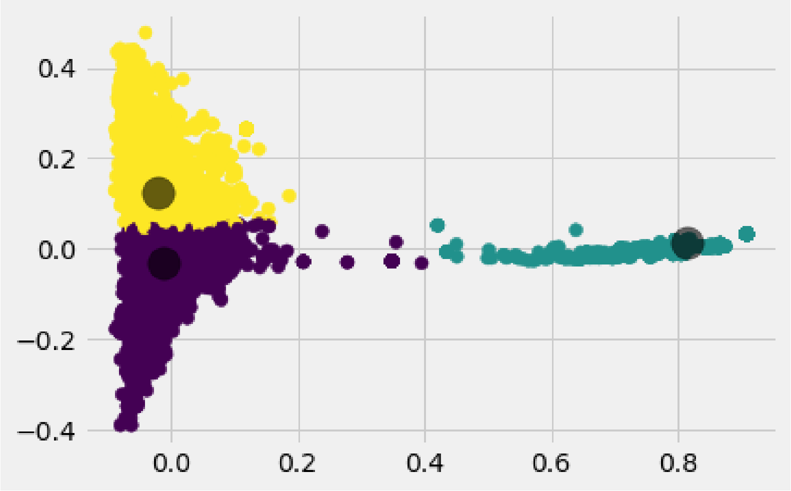
3) Number of clusters determined

To determine the optimal number of clusters, we select the value of k by using the "**elbow**" method. The point after which the distortion/inertia starts decreasing in a linear fashion. Thus for the given data, we conclude that the optimal number of clusters for the data is 3.



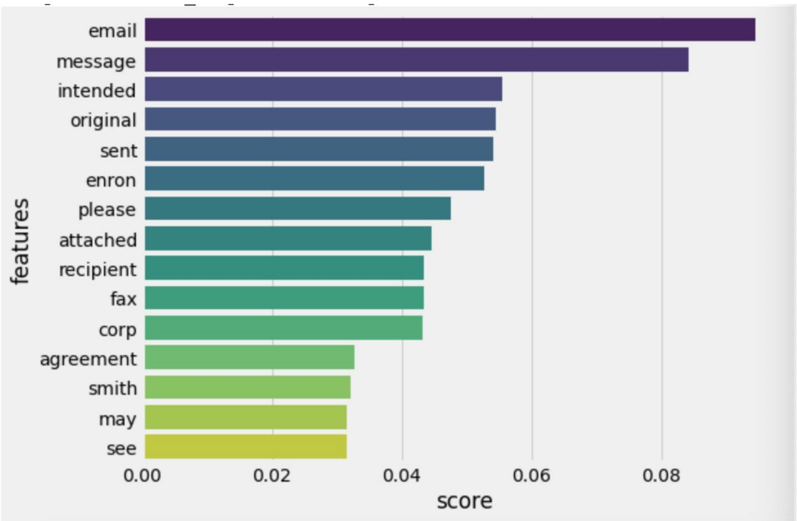
4) Sk-Learn Implementation of K-means

We have projected our array into a 2-dimensional space, so we can easily use a scatter plot to visualize this along with the cluster centers. In the graph below, we can see three pretty distinct clusters here with a particularly large separation for the green cluster indicating quite a difference in terms of the content of the emails. The majority of the data is contained within the purple and yellow clusters.

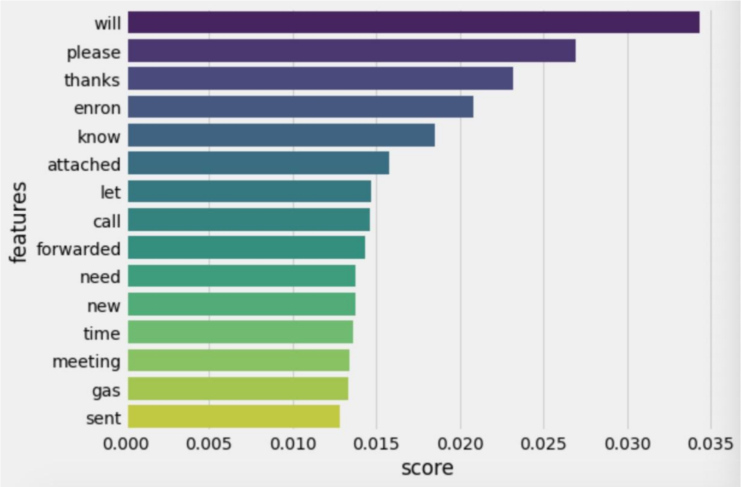


5) Checking the top words in each cluster

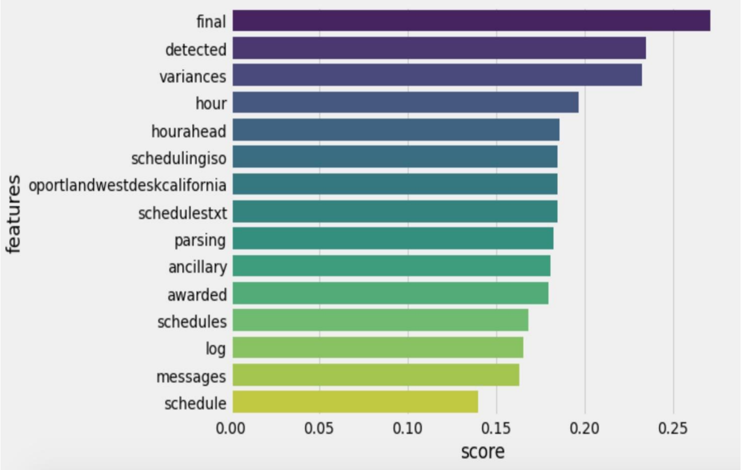
In this sector, what we are mainly interested in is seeing if there are any commonalities between words in each cluster or any particular words that stand out. In other words, what themes in each cluster that we can identify? This is a good way of getting a general feel for what the emails contain and can guide any further analysis we wish to do and avoid reading all of 50,000 emails. We can view the top words in each cluster using the method which identifies the features with the highest mean TF-IDF scores across each cluster.



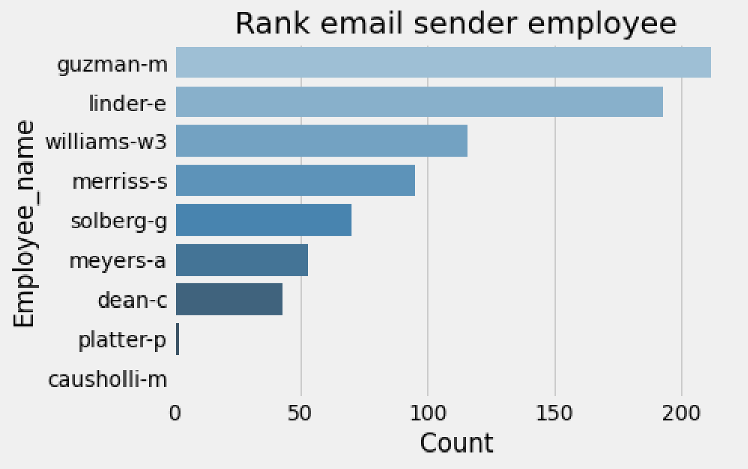
Cluster 1 seems to generally be about communications with features like *email*, *message*, and *attached*. It could be useful if we want to examine the unusual network among the cluster further.



Cluster 2 has a few words like *please* and *thanks* and includes some words related to meeting like *call*, *time*, and *meeting*. It seems to generally be about meetings or appointments. In this cluster, we found that *Enron* and *gas* may suggest some communication related to their gas business.



Cluster 3 has a lot of words suggesting the emails were about *scheduling*. It looks interesting because the Enron scandal involves the California electricity crisis, the features with *portland west desk California* and *detected* make the cluster appear to be worth further investigation.



As we look into cluster 3, we found that the emails in this cluster were mainly sent by Pete Davis. The reports and news also confirmed that he was a California power scheduler. The plot above shows the number of emails saved by employees which were related to their communication with Pete Davis. Among them, Mark Guzman saved the most number of emails. As a power scheduler, we assumed that he scheduled all power purchases and sales with other utilities; thus, using the terms related to scheduling in his emails frequently makes sense.

5. Analytical Findings - Network Analysis

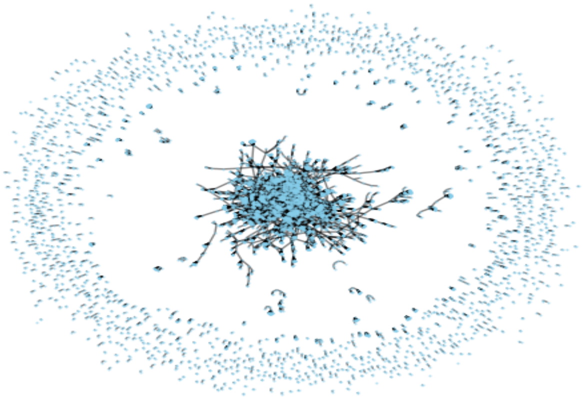
To look into the e-mail communication of Enron managers, we choose four managers who might play important roles in Enron.

(1) Kay Mann, Head of Legal. Most of the emails were extracted from her email folders. She took 4 of 15 top email folders in Enron.

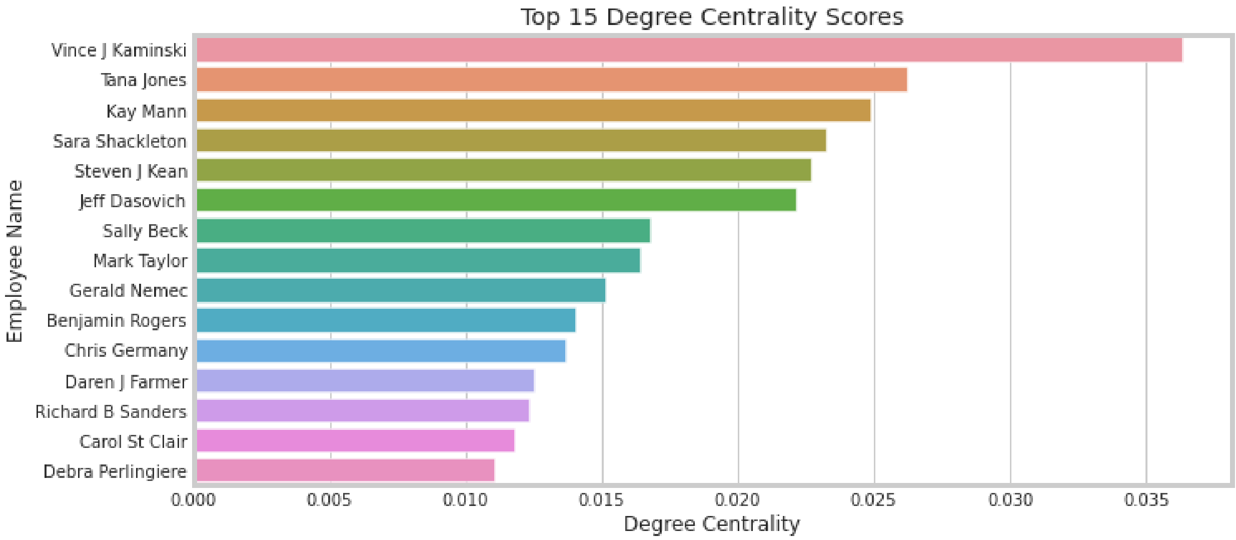
(2) Vince J Kaminski, Managing Director for Research, had sent the most emails and taken one of the two most email folders in Enron. He knocked the alarm and strongly objected to the financial practices of Enron.

(3) Dasovich Jeff, Governmental Affairs Executive. One of the most frequent names found in data exploratory. According to reports, he was involved in concealing Enron's debt problem.

(4) Kate Symes, trader, who was also involved in concealing Enron's debt.

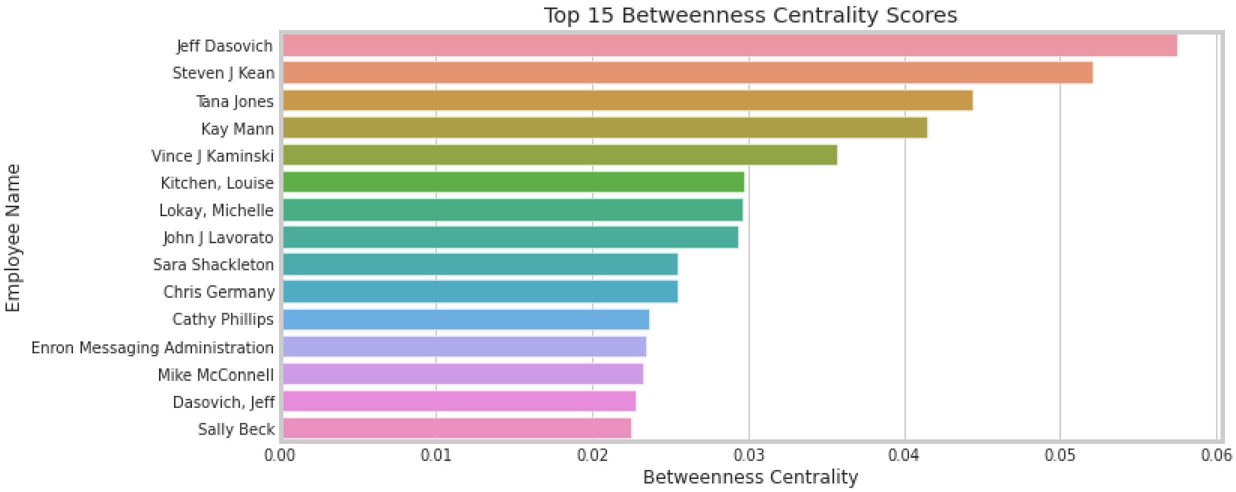


First of all, let’s take a glimpse at this Email Network Map. The map directly tells us that there is a certain number of correlations and centrality in the corpus. More specifically, we can guess it is highly possible that a group of people in Enron knew about the scandal and should take responsibility for it.

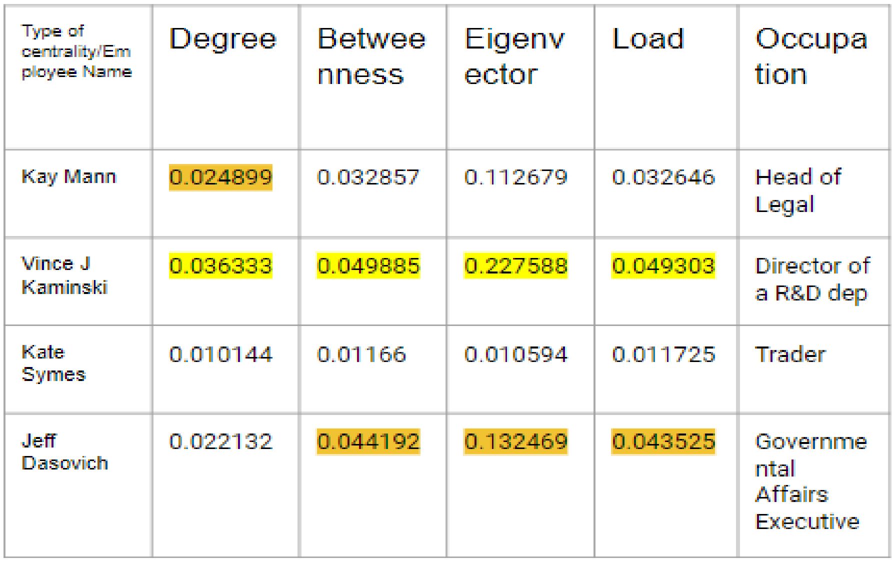


In this 51,740 email corpus, which person has the most connections with others in their networks? The Degree Centrality Model explains it to us. Sometimes it is also expressed by the node's size; the bigger the size of a node, the higher the centrality it occupies.

According to the above graphs, Vince J Kaminski,Tana Jones, Kay Mann, and Sara Shackleton have the highest degree centrality scores, indicating that they have the most direct connections with other employees. These people also have the highest betweenness centrality scores, suggesting that they connect the most separate communities within the network. We expect to see those directors with a higher level of betweenness centrality, as generally, they might have managers from different departments to report to them and therefore should form a link across different communities.



The Betweenness Centrality model allows us to calculate the number of shortest paths passing through a node. The higher the number of shortest paths passing through a node, the higher its betweenness centrality scores are. Jeff Dasovich, Steven J Kean, Tana Jones, Kay Mann have the highest BC scores, indicating that they are the most active people to connect with others, forming a number of internal networks.



Type of centrality/Employee Name	Degree	Betweenness	Eigenvector	Load	Occupation
Kay Mann	0.024899	0.032857	0.112679	0.032646	Head of Legal
Vince J Kaminski	0.036333	0.049885	0.227588	0.049303	Director of a R&D dep
Kate Symes	0.010144	0.01166	0.010594	0.011725	Trader
Jeff Dasovich	0.022132	0.044192	0.132469	0.043525	Governmental Affairs Executive

In addition, we also have tried Eigenvector Centrality and Load Centrality models in the notebook. The EC model, which depends on the importance of its neighbor nodes, unsurprisingly has the highest score for Vince J Kaminski. We can estimate that Vince J Kaminski is the most central and active person in the Enron corpus. The importance of the people he had connected with is significantly higher than others. Back to our questions, from Kay Mann, Vince J Kaminski, Kate Symes, and Jeff Dasovich, who played a more significant role in Enron? Who might have understood the most of Enron before October 2001, when the scandal was revealed? Who was more familiar with the staff in Enron? Now, we should have rough answers to these questions. Although Vince J Kaminski had sent the most emails with the most connections than the other 3 people, Vince is not the most central person (top executive level). Jeff Dasovich has a minor difference with Vince in BC scores when Jeff only sent 61 percent of emails that Vince had sent. His occupation, governmental affairs executive, successfully interprets why he is more central than Jeff who is a director for a research department. In his internal network, there should have many senior leaders in the group. The EC model also tells us that Kate Symes is the least relevant among all four people.

6. Business Problem - Conclusions:

Without reading all the emails, we could select the people we would like to understand and learn who belonged in their network. Network analysis might help investigate relationships, find key people in the communities, and detect companies' structures.

We used clustering to assist in classification. K-means algorithm ended up bringing out some interesting results. The Enron emails could be generally classified into three main clusters including communication contents, requesting something and making appointments, scheduling and messages related to the power scandal. For example, we found a cluster with *California* and *detected* appearing in one cluster. This may indicate that Enron executives urged market engagement with specific traders to drive higher energy costs, which led to the California energy crisis in 2001, causing high energy prices and blackouts throughout the state.

Last but not least, there are many applications for classifying emails in real-time or historical email datasets. It may include reasons such as subject or folder classification, spam detection, etc. Here are some recommendations for building a classification system:

1. Email classification can automatically assign emails to predefined folders, for example, appointments, requests, etc.
2. Regarding networks, we can evaluate the emails' priority based on the messages and the groups or communities the senders/receivers belong to.
3. For supervision purposes, we can identify fraud from suspicious email clusters for further investigation.