Automated Fraud Detection

Using the Enron Email Corpus to Train Fraud Detection Models

A.Sardina-Spevack
Capstone Presentation
December 2019

The Question

Email Corpus

Remaining emails between 1997-2001 made public by Federal Energy Regulatory Commission during its investigation.

- 150 users, mostly senior managers
- Over 500,000 emails

Enron Scandal

At the time represented the largest bankruptcy in American history

- Bankruptcy declared in 2001
- Lead to de facto fall of Arthur Anderson
- Sarbanes-Oxley Act

The Question

Given the far reaching implications of the Scandal, could we learn from the mistakes of the past?

- Can we use the enron emails to train a model to find those committing fraud ?

Technical Details

Scrub and Explore

Initial Wrangling

Exploring the dimensions of the dataset, reviewing formatting for the future model, visualizing complexity of the problem

Unsupervised

K-Means Clustering

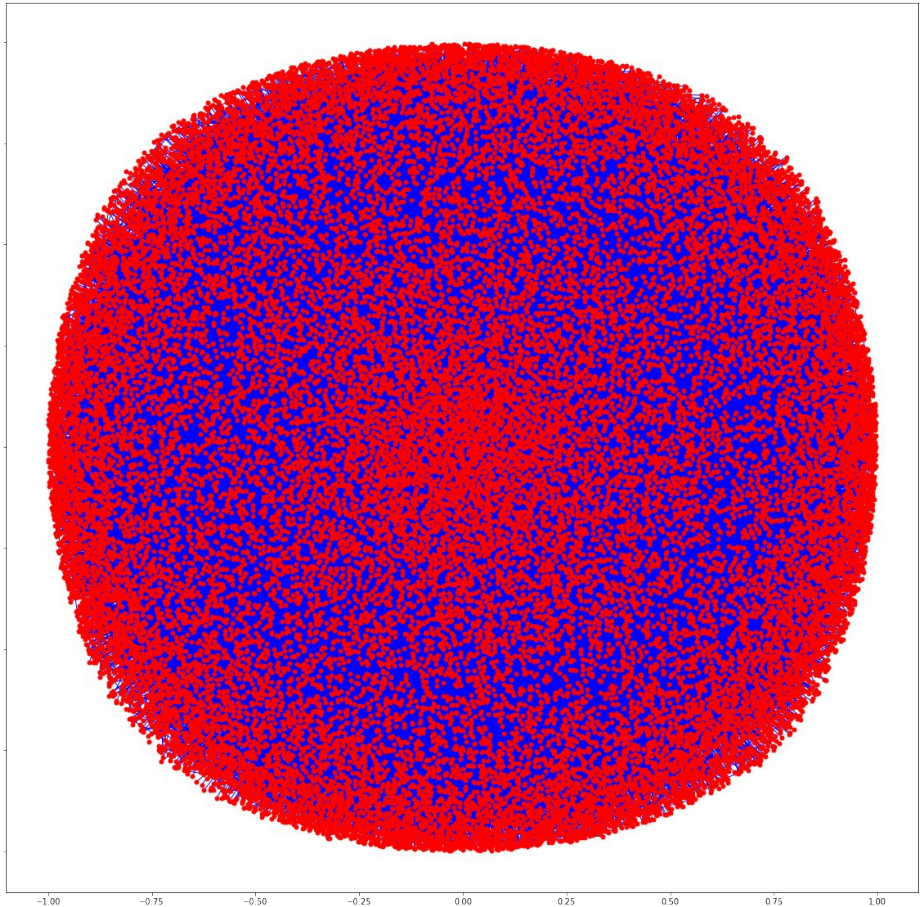
Using a 2 Cluster Model

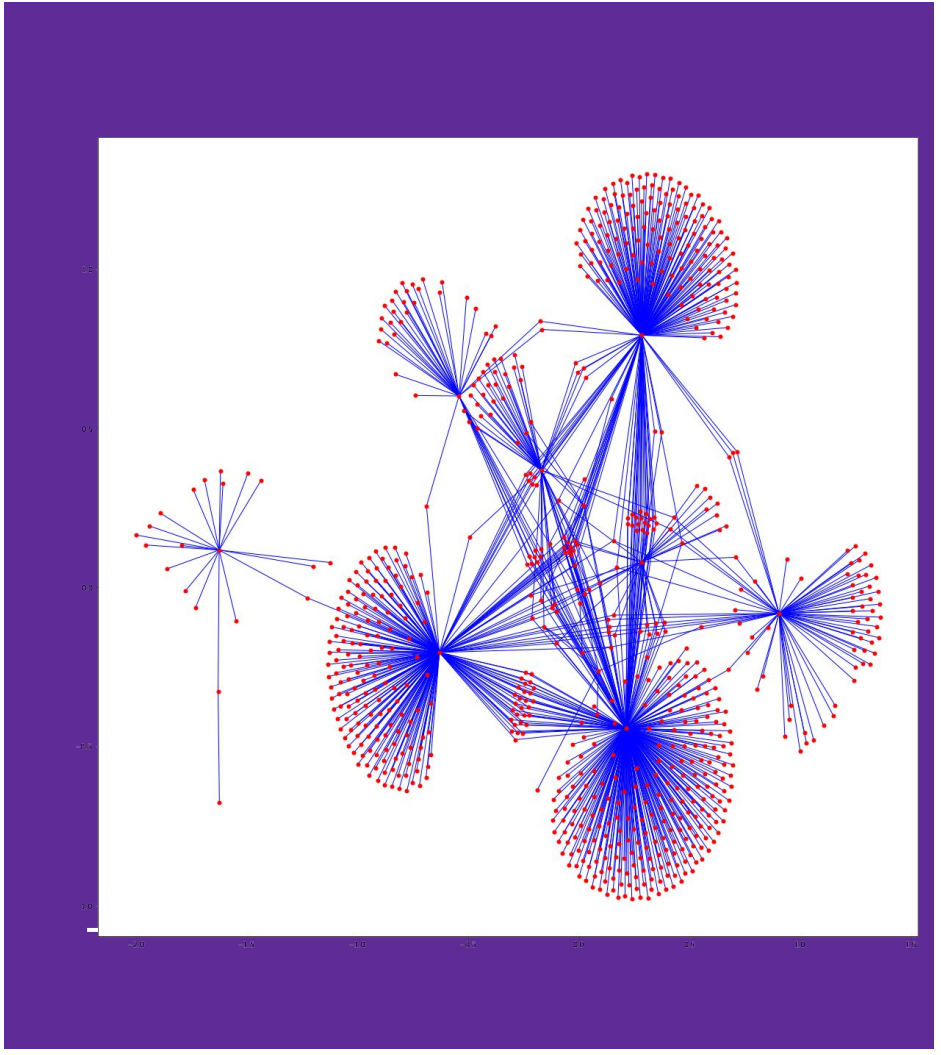
- Trained to find people and important language

Supervised

K-Nearest Neighbors

Over 90 % accuracy in appropriately classifying the emails in the clusters.


1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75
−1.00
−1.00
−0.75
−0.50
−0.25
0.00
0.25
0.50
0.75
1.00



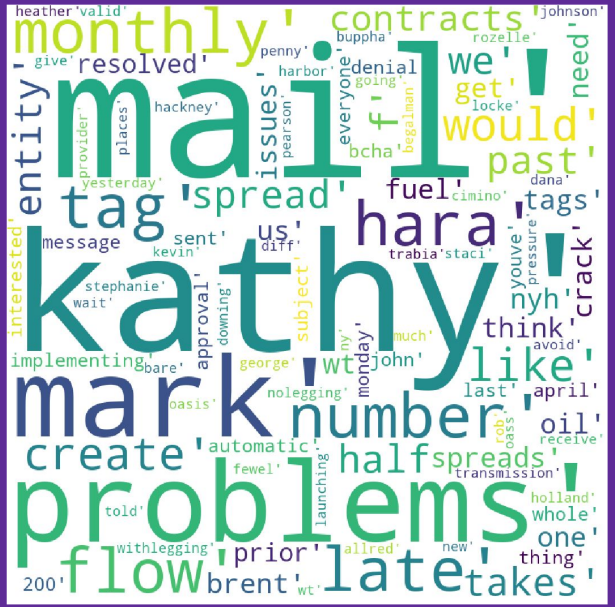
Solution

Efficiency in Resource Management

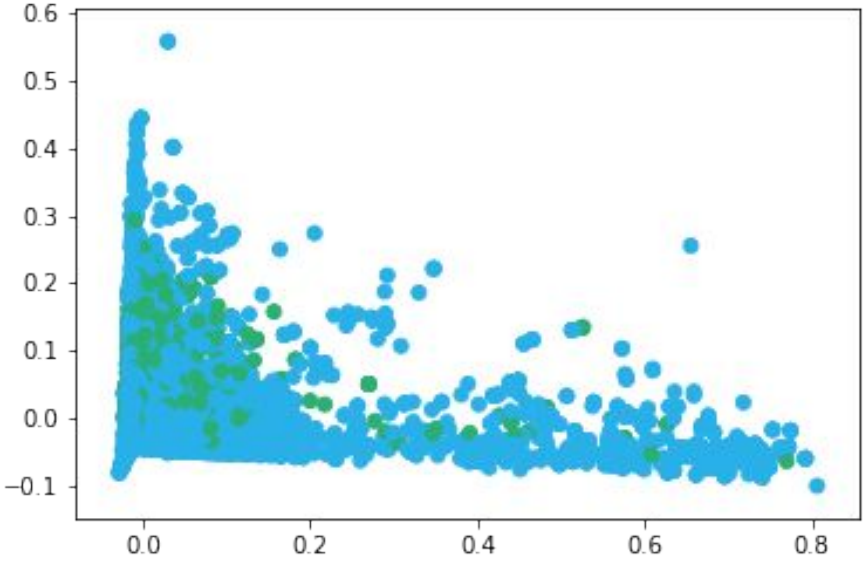
Created a model that could detect individuals who may require additional scrutiny based on their email traffic.

Important Words

Clear implications

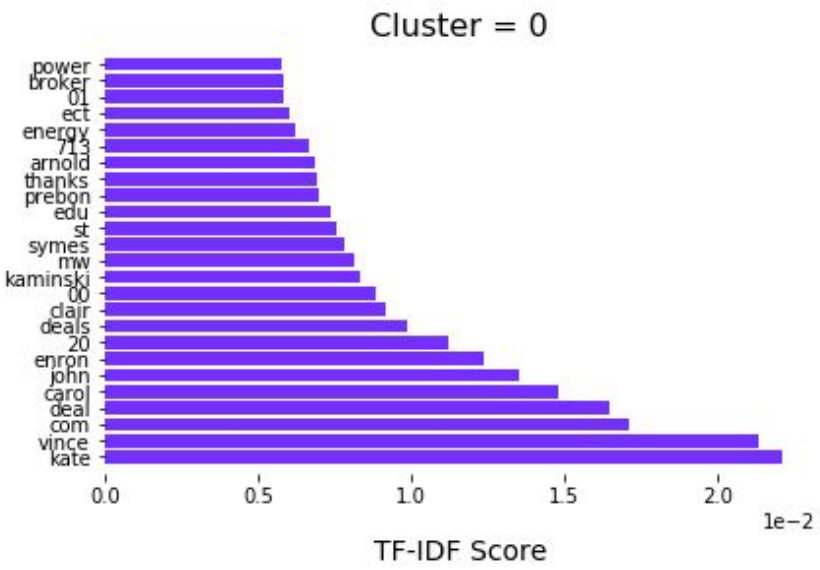


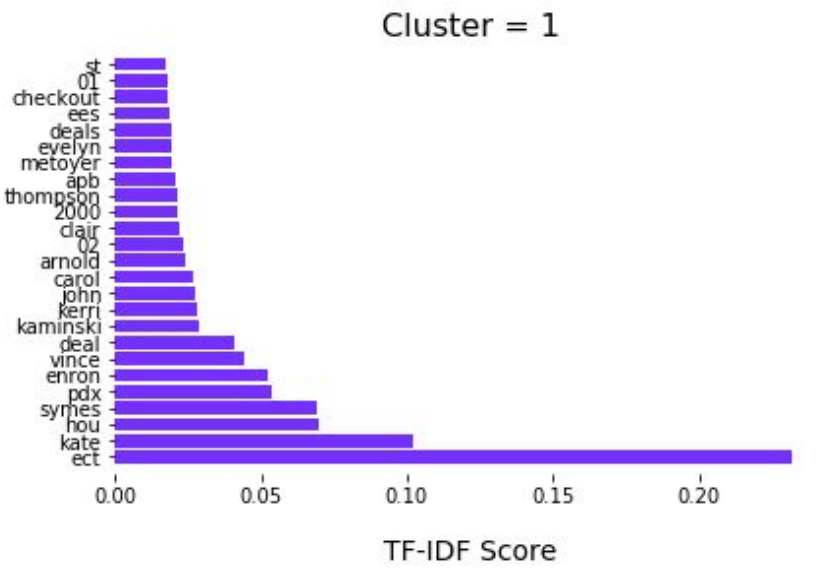
Results and Recommendations



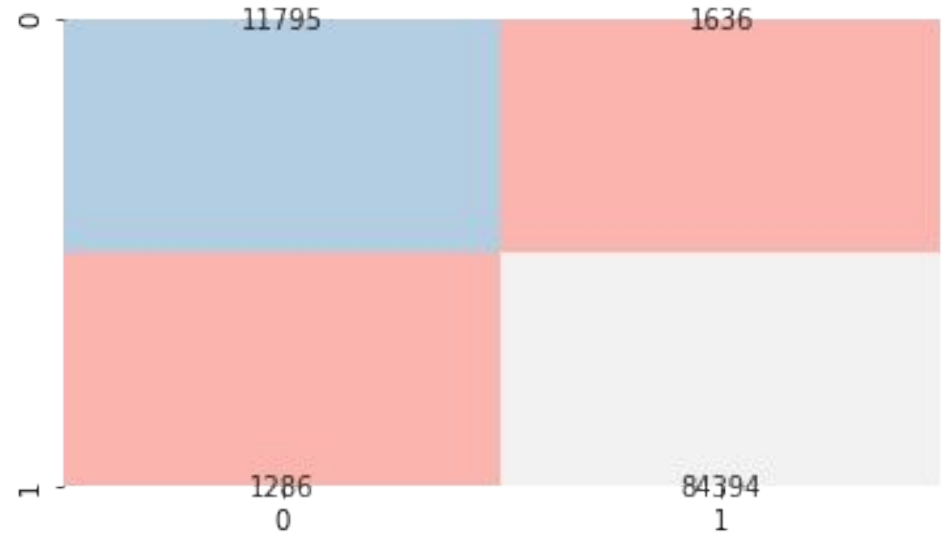
K-Means Clustering

2 Cluster Model





Top Words and Top People

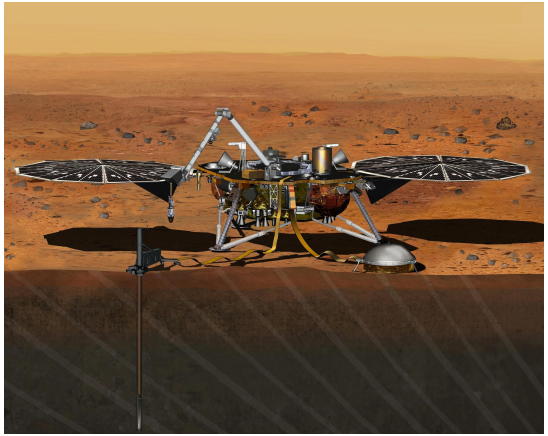


Accuracy

People were implicated in the scandal, and the words in their emails were indicative of ongoing fraud.

Insights and Recommendations

- Efficient First Line of Defense
  - Utilizing Modern Machine Learning as the first indicator of red flags
- Reallocation of resources
  - Fraud Analysts and Risk Managers can reorganize their time to focus on the flags produced by the model capturing the same risks with fewer man hours
- Two-step Model
  - Easily calibrated with additional text bodies
  - Ultimately creates a computationally efficient paradigm



Future Model Extensions

Data

Big Data

Additional resources allocated to run model on the full dataset

- Limitations for my personal machine and Kaggle Remote servers

Model Update

Clustering Steps

First Cluster on People

- Then Cluster and Classify on Language used by people

Additional Tuning

GridSearch for Params

Try additional Clusters and additional Neighbors for further model tuning.

- Utilizer GridSearch

Questions?