Context Aware Text Analysis

Arindam Samanta

Bellevue University

MS-DSC

1. Business Problem / Background
2. Data Explanation
3. Topic Modeling with LDA
4. Model Observation
5. Challenges and Ethical Assessment
6. References and Acknowledgement

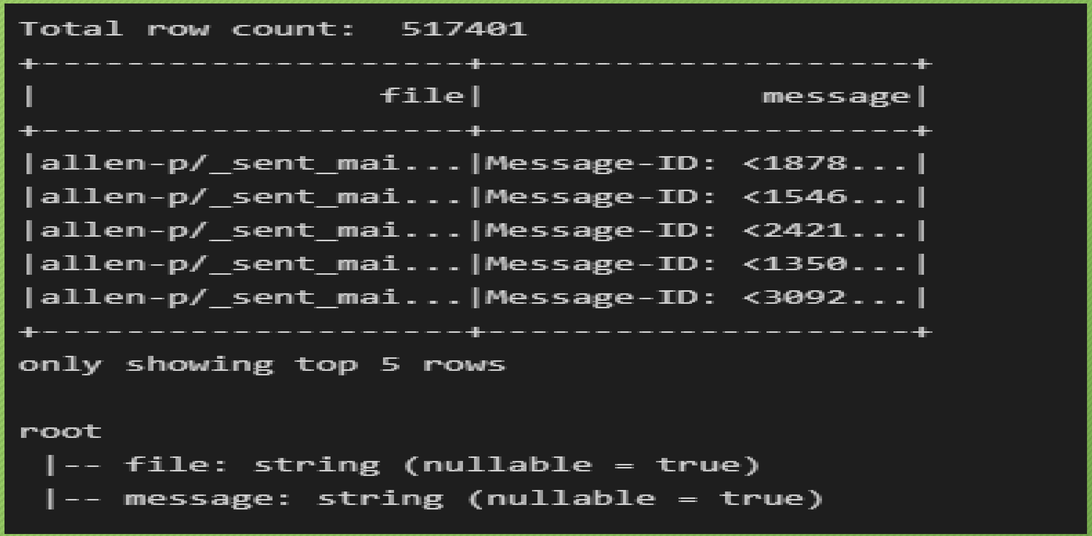
Business Problem / Background

During Enron’s rise to the top, they were intertwined with multiple counts of fraudulent activity that could have been detected years before Enron’s fall if investigators had the right tools. I am trying to solve the problem by proposing a text analytics solution to it. In order to understand the motives of the fraudsters we have to analyze unstructured data, such as emails and memos. Finding the relevant information quickly from the huge pile of textual data is a daunting task for the auditors.

Enron Corporation was an American energy, commodities and services company based out of Houston, Texas. In 2001, they filed for bankruptcy. Before their Dec. 2, 2001 bankruptcy filing, Enron employed 20,000 staff. They were one of the world’s leading electricity, natural gas, communications and pulp and paper companies, with claimed revenues of nearly $101 billion in 2000. Later it was revealed that its reported financial condition was sustained substantially by an institutionalized, systematic, and creatively planned accounting fraud, known since as Enron scandal.

Data Explanation

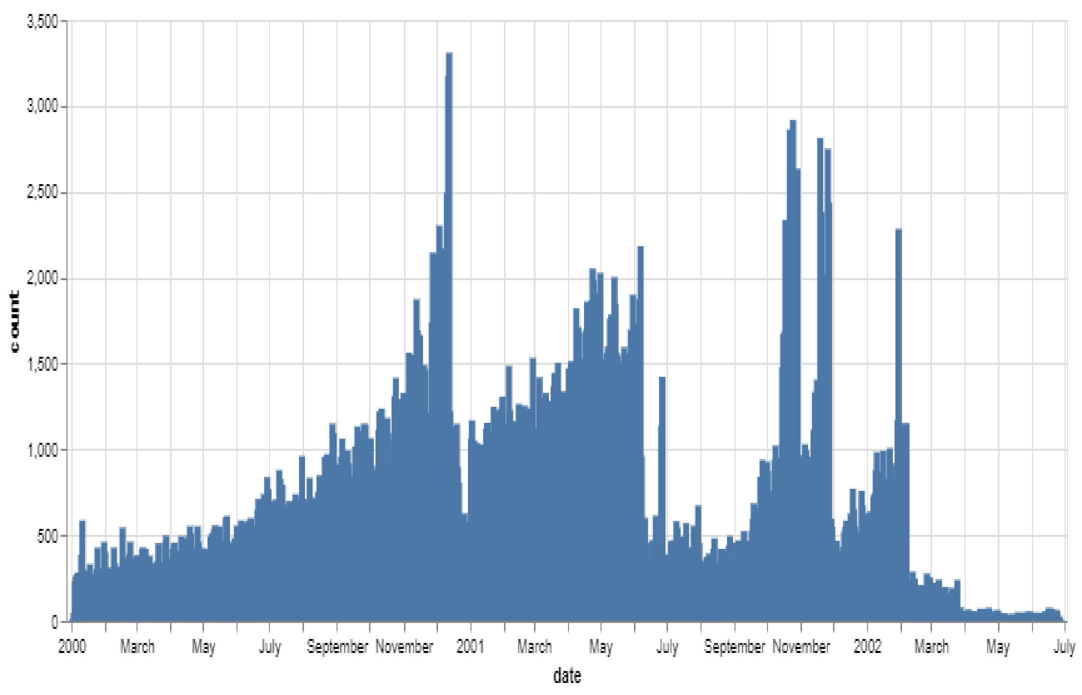
- The Data set contains roughly 500,000 emails of around 150 people related to Enron
- The dataset was made public by the Federal Energy Regulatory Commission (FERC)
- The Kaggle dataset stores the final collection in CSV format and this is what I have used for my analysis.



```
Total row count:  517401
+--------------------+--------------------+
|                file|             message|
+--------------------+--------------------+
|allen-p/_sent_mai...|Message-ID: <1878...|
|allen-p/_sent_mai...|Message-ID: <1546...|
|allen-p/_sent_mai...|Message-ID: <2421...|
|allen-p/_sent_mai...|Message-ID: <1350...|
|allen-p/_sent_mai...|Message-ID: <3092...|
+--------------------+--------------------+
only showing top 5 rows

root
 |-- file: string (nullable = true)
 |-- message: string (nullable = true)
```

The Data set contains roughly 500,000 emails of around 150 people related to Enron
The below diagram shows the volume of emails sent during the last few months leading to the bankruptcy. The peaks are vital months



Topic Modeling with LDA / Analysis of the Outcome

- I have used my personal laptop with VSC as the IDE to perform the analysis. I am giving below the steps upto the modelling:
  - Loaded the dataset(emails.csv) into the pandas dataframe
  - Used the helper functions to parse the raw messages
  - Next I extracted some key portions of the message like(‘msg_id’,‘from’,’To’,’body’)

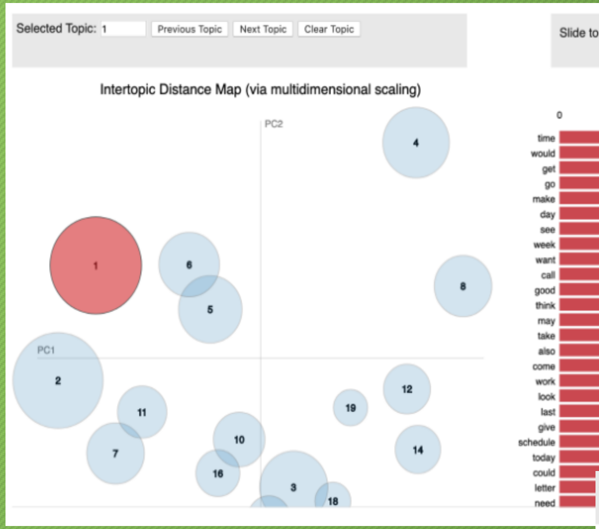
Before building the model we need to understand a few things about the model:

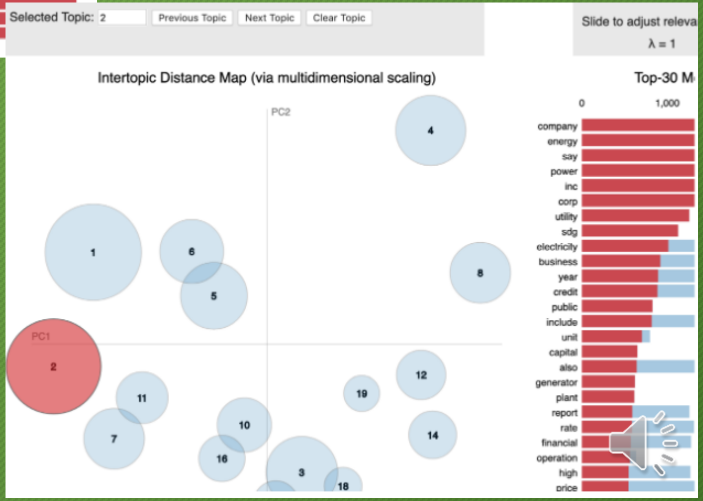
- Corpus
  - Stream of document vectors(num_terms, num_documents)
- Id2word
  - mapping from word IDs to words – helpful for topic printing
- Num_topics
  - Number of requested latent topics to be extracted from the training corpus
- Random_state
  - Randomstate object used for reproduceability
- Update_every
  - Number of documents to iterate through for each update
- Chunksize
  - Number of documents to use for eac training chunk
- Passes
  - Number of passes through the corpus during training
- Alpha
  - Learns an asymmetric prior from the corpus
- Per_word_topics
  - True: the model also computes a list of topics sorted in descending order

Model Observation

A few observations:

- The size of the bubbles tells us how dominant a topic is across all the documents(emails)
- The words on the right are the key words driving the topic
- The closer the bubbles the more similar the topic.
- Our goal is to ave non-overlapping bubbles as much as possible





For comparing models a lower perplexity score is a good sign. Perplexity: -15.236878229690644

Coherence score is a better predictor of the quality of topics as opposed to Perplexity score. My score: Coherence Score: 0.455994351574875 is good. This score is trying to quantify the semantic similarities of the high scoring words within each topic. This results in more human interpretable.

Challenges / Ethical Assessment

- The email corpus was huge to be handled in a laptop
- Difficult to analyze some of the emails as they were personal in nature
- Absence of training data
- Although topic modelling and sentiment analysis do provide some insight into the data, but I still think it is not an ethical investigating tool
- Although I was lucky to get the Kaggle dataset and it was nicely arranged, but I do understand that looking at the corpus it is a huge dataset to be handled in a laptop.
- Ethically assessing some of the emails was very difficult as they were very much personal in nature.

References and Acknowledgement

References:

- Finally I would like to acknowledge Bellevue University for giving me this opportunity to present this academic project for the 2022 academic year as part of our MS Data Science course(MS DSC 680).
- Additionally I would like to thank our Professor Fadi Alsaleem for his expert advice and guidance. Lastly I would also like to thank my class mates for continuous support and encouragement.

- Abbott, David. Applied Predictive Analytics: Principles and Techniques for the Professional Data Analyst
- https://www.kaggle.com/datasets/wcukierski/enron-email-dataset
- https://www.cs.cmu.edu/~enron/enron_mail_20150507.tar.gz