Arindam Samanta

Context Aware Text Analysis

BUDSC-680

Bellevue University

TABLE OF CONTENTS

BUSINESS PROBLEM

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.

One way to tackle the problem is to use Text Analytics using NLP and pattern matching to facilitate timely fact extraction and data organization.

BACKGROUND/HISTORY

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

(Data Prep / Data Dictionary / etc)

The 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 0.5 million messages. This data was originally made public and posted to the web, by the Federal Energy Regulatory Commission during its investigation. The dataset consists of 517,431 messages that belong to 150 users.

Although the dataset is huge but folders of particular users are often quite sparse. I decided to look into the sent emails folder. Through this approach I could avoid analyzing spam email or junk email folders.

Top 5 Users by Message count:

	Users	Count
0	kaminski-v	28465
1	dasovich-j	28234
2	kean-s	25351
3	mann-k	23381
4	jones-t	19950

Breaking down the email data by row

file	users	val	message
allen-p/_sent_mai...	username0	allen-p	Message-ID: <1878...>
allen-p/_sent_mai...	username1	_sent_mail	Message-ID: <1878...>
allen-p/_sent_mai...	username2	1.	Message-ID: <1878...>
allen-p/_sent_mai...	username0	allen-p	Message-ID: <1546...>
allen-p/_sent_mai...	username1	_sent_mail	Message-ID: <1546...>
allen-p/_sent_mai...	username2	10.	Message-ID: <1546...>

METHODS

I have used the email corpus in 2 separate forms.

Emails.csv [1]

Enron_mail_20150507.tar.gz[2]

The reason for using the 2 different sets was the ease of analysis. I have done some extensive study of the different analysis done so far on these corpus and found it fascinating about how this has helped in coming up with the different visualization and insights into the email corpus.

This is how the data looks like in the emails.csv file:

```
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)
```

Topic Modeling with LDA

Latent Dirichlet Allocation was performed on the dataset with the number of topics ranging from 4 to 10. As it is yet to be done so I am providing a sample of those here.

Listing below the top 10 terms for each topic

Topic 1	Topic 2	Topic 3	Topic 4
"message"	"enron"	"market"	"thank"
"origin"	"deal"	"gas"	"call"
"pleas"	"agreement"	"price"	"time"
"email"	"chang"	"power"	"meet"
"thank"	"contract"	"compani"	"look"
"attach"	"corp"	"energi"	"week"
"_le"	"fax"	"trade"	"day"
"copi"	"houston"	"busi"	"dont"
"inform"	"date"	"servic"	"vinc"
"receiv"	"america"	"manag"	"talk"

Topic-1
Meeting Related

Topic-2
Business
Process/Operatio
nal

Topic-3
Core Energy
Business realted

Topic-4
Business Casual
conversations

ANALYSIS

Fig 1 shows the total email count by the top 20 executives of the Company.

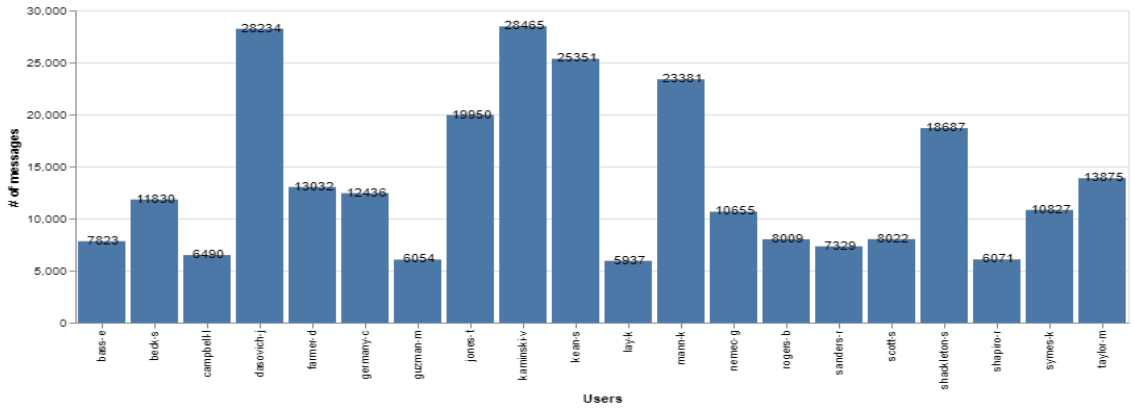


Fig:1 – Top 20 users by Email count

Fig 2 shows how the volume of emails shows the time upto the bankruptcy days

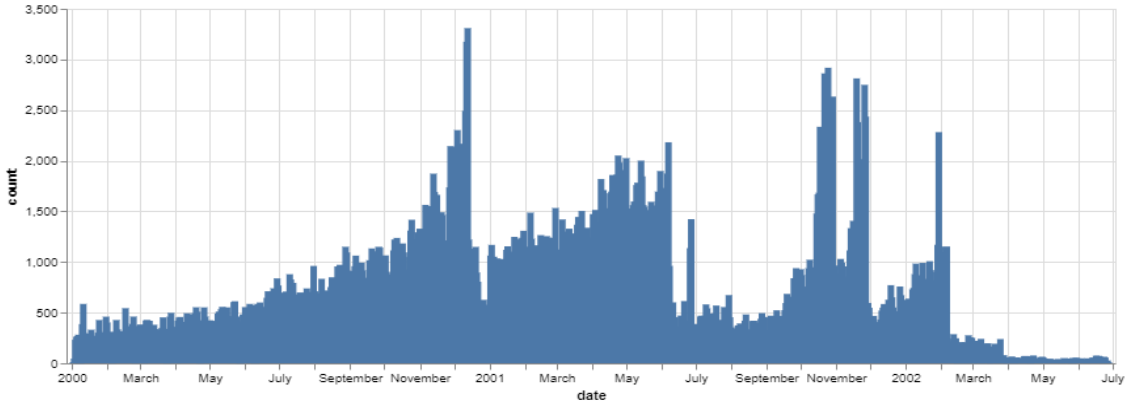
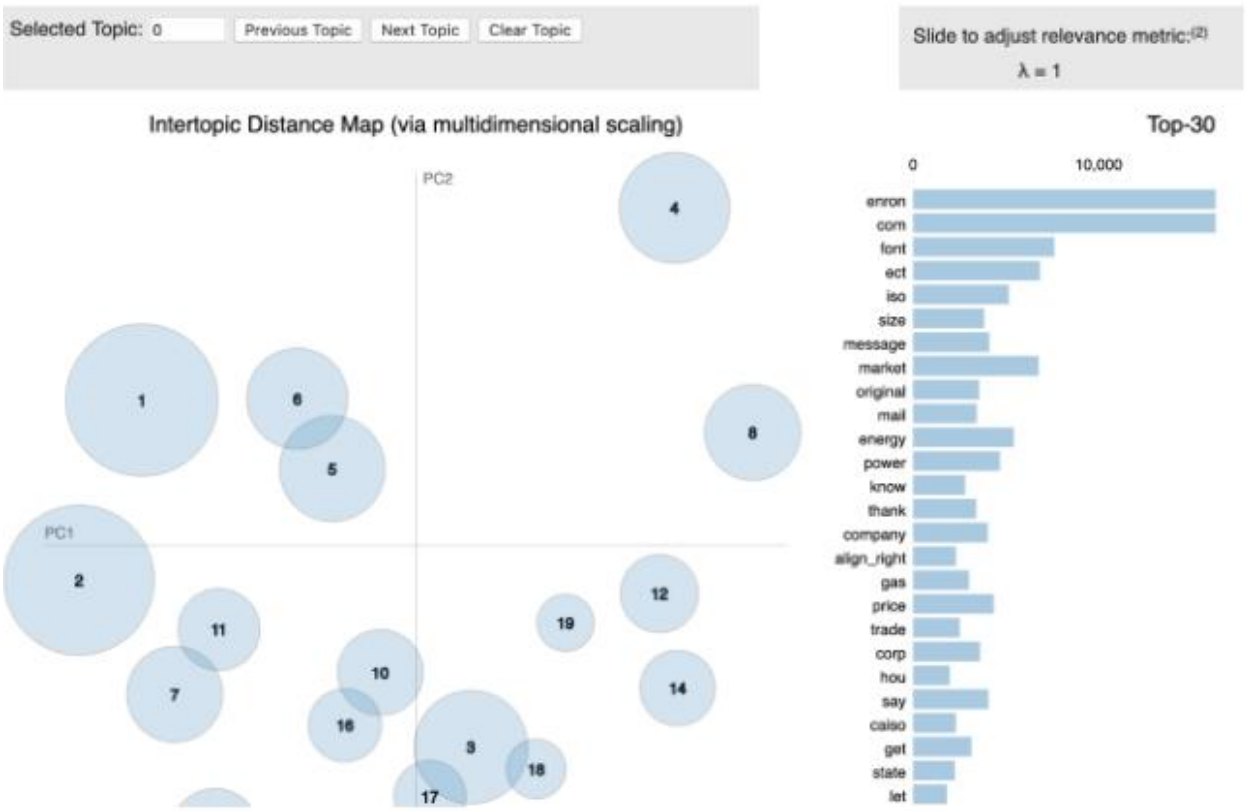


Fig 2: Volume of emails sent chronologically (Jan-2000 -Jul-2002)



Visualizing our model using PyLDAvis

1. The size of the bubbles tells us how dominant a topic is across all the documents
2. The words on the right are the keywords driving that topic
3. The closer the bubbles the more similar the topic.
4. Preferably we want non-overlapping bubbles as much as possible spread across the chart

MODEL EVALUATION

We judge our model by using two scores perplexity and the coherence scores.

Perplexity is the measurement of how well a probability distribution predicts the sample. The lower the value the better is the model. In our case we got a value of -15.236

The coherence score is a better indicator of the quality of topics. This score tries to quantify the semantic similarities of the high scoring words within each topic. A high score means a better model. In our case it is 0.455

CONCLUSION

In conclusion, although topic modelling and sentiment analysis do provide some useful insights into the data, it is still unclear as to whether they can be used as investigation tools. However, with that being said, there are some ways in which the analysis can be improved upon. For example, analysis can be conducted by focusing on users who directories are especially large, namely, Sally Beck (Chief Operating Officer), Darren Farmer (Logistics Manager), Vincent Kaminski (Head of Quantitative Modeling Group),Louise Kitchen (President of EnronOnline), Michelle Lokay (Administrative Assistant), Richard Sanders (Assistant General Counsel) and Williams III (Senior Analyst). This would ensure that only the emails of most relevant people to the scandal are examined and this might reveal more interesting patterns.

So using this outcome we can now look at each email to assess the document topic and related keywords. This could be useful in a text summarization or topic labelling task. The approach we are using here is to figure out which topic contributes the highest percentage to a given document.

LIMITATIONS

- The corpus of the emails were huge to be handled with a small laptop
- It was really difficult to analyse the context of some of the emails as they were mostly personal in nature
- As these were personal emails and released to public so they had personal information and analysing those caused lots of ethical issues.

CHALLENGES

This being my first project independently handling all the aspects of the analysis, I was a bit overwhelmed by the steps to take and also perform the necessary activities leading upto my conclusion. One of the main challenges in conducting sentiment analysis on the emails was the complete absence of training data i.e. emails with positive or negative labels or any types of labels at all. Even I had to identify the words to come up with the topics.

ETHICAL ASSESSMENT

Ethically I found it very questionable to look into the emails as they also had personal conversation and I thought it was not a good thing to make it public. From my understanding of learning data science I always think that using personal data of any user without proper permission of the user is non ethical and also not desirable for a healthy data science project.

REFERENCES:

[1] Data Source: https://www.kaggle.com/datasets/wcukierski/enron-email-dataset

[2] https://www.cs.cmu.edu/~enron/enron_mail_20150507.tar.gz

3. https://medium.datadriveninvestor.com/nlp-with-lda-analyzing-topics-in-the-enron-email-dataset-20326b7ae36f