QUESTIONS & ANSWERS

1. What is the significance of looking into the emails?
    - There may be many ways of looking into it but for me it was more of an investigating work. As part of the process, I always wanted to go over the emails but looking at the sheer volume it was not humanly possible to go over each and every one of those. So I took the help of AI and ML to scan the emails and group them under some heads and then read the ones that seemed more relevant.

2. What type of questions you are thinking that the emails can answer?
    - The emails in the beginning seemed did not provide very helpful information but once we started going over some of the grouped emails much information regarding the nature of the communication started coming out.
3. What type of sentiment analysis do you think you can do with the emails?
    - As part of the context aware text analysis I did not do any kind of sentiment analysis for this project.
4. How do you plan to label the emails if at all you are thinking of doing it?
    - As there were no test data available for this exercise on top of it this was not a sentiment analysis so I did not plan on labeling the emails in any way. Instead I did try to segregate the emails into separate buckets for further ease of analysis.
5. What type of email sentiments do we get from these corpus?
    - As mentioned this project was not involved in any kind of sentiment analysis of the emails.

6. Are there any indications in the emails that could have predicted the bankruptcy?
    - Frankly speaking I could not dig into so much of details in this project. My overall idea was to divide the emails into different buckets so that while investigating it would help the investigators to look into the proper buckets and spend less time in scanning those.
7. How do you train the model with test data?
    - This was a challenge for this exercise as I have mentioned in the challenges section. There was no test data available for the model to be trained on. So I have used .02% of the corpus or 10348 emails as a test set to generate the model.
8. What tools did you use for data preparation?
    - I have to admit here that I have not spent much time for preparing the data as I got the prepared dataset from Kaggle. I have worked on rearranging the data when I loaded the .CSV file into a pandas / pyspark dataframes.
9. What tools did you use for data visualization?
    - I have not used much tools for data visualization due the short period of time I have for this project. Also the source I have chosen had already been worked on before so there was not much of left to analyse for the model. But then I have used matplotlib and pyLDAvis package is also very good for data visualization.
10. What methods did you use for the model validation?
    - I have used Gensim LDA model for Topic modeling. The perplexity and the coherence scores of the model gives us an idea of how good the model was in defining the topics. The coherence score is a better predictor of the quality of topics as compared to the perplexity score. The Coherence score is trying to quantify the semantic similarities of the high scoring words within each topic. A high score means that the result is more human-readable.