ENRON EMAIL ANALYSIS

- **Enron Email Sentiment Analysis :**
    a. After analyzing the Enron email dataset, I chose the Vader Sentiment Analysis method to detect sentiment in the email body.
    b. Vader Sentiment Analysis was chosen because the dataset is not labeled and for unsupervised sentiment analysis, vader seemed to work pretty decently as far as latency and accuracy both are concerned.
    c. It uses a lexicon, i.e, a dictionary of words associated with either positive or negative sentiment and also polarity of the sentiment, pos tags, mood, etc.
    d. Using these lexicons, the sentiment of a text is computed by matching the occurrence of specific words from the lexicon and looking at other factors such as context, phrases, etc and aggregating overall sentiment score to give the final sentiment.
    e. **Other approaches considered:**
        i. Bert based pre-trained models like FinBert. Finbert is a BERT model fine tuned on financial data and since enron's emails are more or less related to business communications , finbert would have been a decent approach to consider especially when the labels are not provided.
    f. **Constraints:**
        i. Latency : 0.001 seconds. The method is very fast but less accurate for this data.
        ii. Since Vader is trained on social media data, the results might not be accurate on the Enron emails. Also some of the Neutral emails might get misclassified as Positive or vice - versa

- **Enron Oil & Gas Topic Classification:**
    a. To detect whether the emails are related to Enron's Oil and Gas business, I have chosen to go with the Latent Dirichlet Allocation (LDA) algorithm.
    b. LDA uses a bag of words approach that automatically discovers topics contained within a set of documents. It is based on the Dirichlet distribution which is a probability distribution that accounts for word frequencies.
    c. Now, to identify the topic of the email body as related to oil and gas, we can look at the given email to assess the dominant topic as oil and gas
    d. **Other approaches considered** :
        i. **EnsembleLDA** :
            1. Ensemble LDA trains an ensemble of models and discards topics that do not reoccur across the ensemble. Topics here are more reliable.
        ii. **Zero Shot Classification :**
            1. Bart based pre-trained NLI models can be used as zero shot topic classifiers. It works by posing the document as the NLI premise and to construct a hypothesis from each topic label.
    e. **Constraints:**
        i. Latency : 0.2 seconds.
        ii. Topics learned from a trained LDA model are not reproducible and same topics are not learned for repeated training. Unreliable topics are not a good representation of the corpus.
        iii. Given the time constraints, LDA model can be fine tuned further to discover oil and gas related keywords and optimize for it.