

**Task Assignment: Email Topic Modeling**

**Scenario:**

Our organization receives a large number of emails every day, ranging from 50 to 250, which need to be sorted, attended to, and acted upon. Once processed, the emails are rarely seen again. Although the messages are stored, their data and our interaction with that data are not used in any meaningful way, except for the occasional retrieval via search. Therefore, we want to perform topic modeling on the emails to identify the most commonly discussed topics.

**Goals:**

The goal of this project is to perform topic modeling on the emails received by the organization to identify the most commonly discussed topics. This will provide a more structured and quantitative way to analyze the email data, as well as potentially uncover hidden patterns and trends.

**Data:**

The sample dataset for this project is available in the attached file.

**Insights:**

1. Data Cleaning: The first step in this project is to clean and preprocess the email data. This will involve removing any unnecessary characters, such as punctuation and special symbols, converting all words to lowercase, and removing any stop words or irrelevant words.

2. Document-Term Matrix: The next step is to create a document-term matrix, which represents the frequency of each word in each document (i.e., email) in the dataset. This matrix will be used as the input for the topic modeling algorithm.

3. Topic Modeling: There are several algorithms that can be used for topic modeling, such as Latent Dirichlet Allocation (LDA) or Non-negative Matrix Factorization (NMF). The selected algorithm will be used to identify the most commonly discussed topics in the email dataset.

4. Topic Visualization: Once the topics have been identified, they will be visualized using a bar chart or other appropriate visualization technique to display the frequency of each topic.

**Hot Fix:**

The topic modeling algorithm will be evaluated for its effectiveness in identifying the most commonly discussed topics in the email dataset. Any necessary adjustments will be made to the preprocessing or topic modeling steps to improve the accuracy and relevance of the identified topics. Additionally, the topic visualization will be evaluated and adjusted as needed to ensure that it is visually appealing and easy to understand.