ENRON - Email Dataset

Bonus Assignment

ENRON Dataset

- In this bonus assignment, you will analyze the ENRON - Email Dataset, which consists of approximately ***500,000 emails*** generated by employees of the Enron Corporation.

- It was obtained by the Federal Energy Regulatory Commission during its investigation of the Enron's collapse.

- **ENRON Corp** was shutdown in the **year 2001 due to Bankruptcy.**

Overview of the Dataset

- Information of the email such as ***sender, receiver, date of email, text of email, the subject line, and etc*** is provided in the dataset.

- Example of some features extracted from each record (ref. previous slide)
  - Sender : `phillip.allen@enron.com`
  - Receiver : `john.lavorato@enron.com`
  - Date : `Fri, 4 May 2001 13:51:00 -0700 (PDT)`

- Feel free to use any other features in the column. The objective is to see how you reason and handle new problems.

Columns Overview

- For your task, you will be using a structured version of ***ENRON Email Dataset***
- The dataset has **16 features** in total

Message-ID	Unique identifier
Date	Date of email
From	Sender of email
To	Receiver of the email
Subject	Subject of the email
Mime-Version	Multimedia type
Content-Type	Type of email body
Content-Transfer-Encoding	Encoding applied

X-cc	Cc'ed users
X-bcc	Bcc'ed users
X-Folder	Folder of origin
X-Origin	Origin Person of the message
X-FileName	Name of the file attached (if present)
content	Content of the email
user	User name of the sender
has_forwarded_content	Does the email contains forwarded email?

Exploratory Data Analysis

1. To get started, start with some basic statistical informations
    a. # of employees
    b. Average number of employee cc-ed per email
    c. Number of emails sent each day
    d. Types of content shared
    e. Etc.

2. Move to much more sophisticated analysis such as
    a. Most frequently used words in the correspondence
    b. Analyze the distribution of email lengths and response times to understand communication efficiency and responsiveness.
    c. Analyze email response times and patterns between employees. Are there particular individuals who consistently responded more quickly or slowly?
    d. Investigate whether certain topics or types of emails elicit faster responses compared to others.

Try to pose some research questions and run the analysis you prefer to address them. And, most importantly, BE CREATIVE!!!

In-Depth Analysis

- Perform an In-Depth Analysis using machine learning, network science, or NLP techniques to understand the **relationship between the employees and hypothesize the reasons for collapse of ENRON Corp.**

- Some Ideas to get started with your analysis
  - **Understand the context of the emails sent**
    - What topics are mainly covered?
    - Is any topic related to the collapse of ENRON Corp?
  - **Examine the interaction network**
    - Centrality of the employees
    - Groups/communities of employees
  - **Combine the information extracted** from the different analysis to
    - Pose research questions and perform the needed steps to **address them**

Network Analysis

- Construct a network graph representing employees as nodes and email exchanges as edges. Analyze the network structure and properties.
- Identify key communicators. Which employees receive/send the highest number of emails? Can you recognize what kind of responsibilities they have in the company?
- Look for clusters or communities within the network to identify groups of employees who frequently interact with each other (and the content of the emails they exchanged)
- Look for instances where emails are frequently forwarded or cc'd to multiple recipients. Are these emails indicating communication bottlenecks or decision-making processes that involve multiple stakeholders?

NLP Analysis

- Identify frequently mentioned keywords or phrases to gain insights into the most common subjects of communication within the organization.
- After you identified employees who are the most frequent senders and receivers of emails, explore the content and context of emails sent by top senders to understand their roles and responsibilities within the company.
- Which of the employees send/receive a high volume of emails related to topics such as travel/sick leave/vacation/financial matters? Can you recognize the role of those employees in the company?

- Sentiment Analysis (optional):
  You can load pretrained sentiment analysis models to google colab and use it for inferring the sentiment. Such as This Link or This Link
  a. Apply sentiment analysis techniques to assess the overall tone and sentiment of email exchanges. This could help identify areas of concern or employee satisfaction within the organization.
  b. Look for correlations between sentiment scores and other variables such as sender-receiver email topics or the frequency of their mutual correspondence.

Time Series Analysis

- Calculate the total number of emails sent and received by the employees over time. Identify trends in email activity, such as peak times or days of the week with the highest email traffic. Can you identify some anomalies connected to some specific events?

- Visualize the flow of emails between different departments or teams over time. Are there changes in communication patterns during specific periods, such as project launches or organizational restructuring?

- If you have recognized the central employees in this company, use Granger Causality analysis to see if the volume of sent/received emails for one of them is predictive of the volume of sent/received emails for the other person. You can try this approach with different groups/communities/roles/etc.

- Can you predict the daily volume of emails using LSTM?