CS 671A: Introduction To Natural Language Processing.

Analysing activities of Enron company employees based on intra-level and inter-level interaction, in a company hierarchy structure.

Team Members:
Manuj Narang(14373)
Pratik Mishra(14493)

Prof: Harish Karnick

Project Goal (stays as it was):

To develop a model to analyse fraudulent activities in a company, based on interaction between the employees(via emails).

Work Done:

Cleaning and arranging the Data:
Arranged in two columns, one having file name and other the general email content was given as the dataset.
Email content had information arranged as follows(16 columns in which it could arranged): Message-ID, Date, From,To,Subject,Mime-Version,Content-Type; charset, Content-Transfer-Encoding, X-From, X-To, X-cc, X-bcc, X-Folder, X-Origin, X-FileName and the email body. A data frame taking in which all the features extracted was made, all the features as its columns.
All the features were extracted using regular expressions, python.

Some common points about the emails:

1. “To” represents” to whom the the email was sent and “X-to” represents the name to which the email was sent, similarly for From and X-From.(Around 1k emails had vague descriptions)
2. Subject, represents the email of the email.
3. X-cc and X-bcc represent the usual list of emails to whom the emails were sent, carbon copied and blind carbon copied.
4. Date feature contains the date and time at which the email was sent.(used for weekly activity analysis).character length:[37-38]
5. Email content (which has the body of the text)was separated from the mails(first action on the dataset) and LDA was performed on it.(As it consumes more time). Content length varied from [0-2,011,422] characters

Usually all the information features about the email were easily arranged, but some common difficulties were:

1. Some emails did not have “To” or “X-To” or both of them, so all these emails were extracted and stored in a separate data-frame.(Will be analysed later.)
2. Multiple email-id in “To”, emails had a length more than “500”. Some emails had vague description hence 809 emails were dropped for now.
3. Email-Threads were accounted for and considered as linked emails.

Data Set Analysis:

Topic model analysis of the email content extracted(Task2)(performed first, as the content had been extracted for this task):
On all the content of the emails,(taking into account the email threads) LDA was performed in batches of 10 to categorize topics of the data.(As LDA takes times to run on local machines it was broken into batches). Also further with time to model the change in topics of discussion with time LDA will be performed on data arranged according to time.

Week wise analysis: Emails were analysed to account for activity of the employees.

Social Network Analysis(Task1): A basic network has been made up using the python NetworkX.(Around 25k nodes and 56144 edges) using the information from the main dataframe. And the hierarchy model would be build on it. The full hierarchy model would be coded in R because of its text mining libraries.(following the paper) (Some redundancy needs to be removed here).
A dictionary pointing emails to its users was build for network analysis.
Analysing the behaviour of people using the social hierarchy would be done as proposed.

Main Tasks to be done at present:

1. Topic modelling with time, after arranging the emails with time.
2. Building the complete hierarchy structure, both of which are expected to complete till 12th October.
3. There after the sentiment analysis and classification of employees would be done.(3rd November)
4. Final tweeks in the project, till the project submission.