Data Science Methodology

Topic: Analyzing the Past using Email's

1. Which topic did you choose to apply the data science methodology to? **(2 marks)**

   I selected my topic as 'Email'.

2. Next, you will play the role of the client and the data scientist.

Using the topic that you selected, complete the Business Understanding stage by coming up with a problem that you would like to solve and phrasing it in the form of a question that you will use data to answer. **(3 marks)**

You are required to:

1. Describe the problem, related to the topic you selected.
2. Phrase the problem as a question to be answered using data.

For example, using the food recipes use case discussed in the labs, the question that we defined was, "Can we automatically determine the cuisine of a given dish based on its ingredients?".

1. The problem is:

- I want to get a person's past information like 'where was he working?', 'where did he study?', just by analyzing Email data. - It could help investigate a criminal for the cops.

2. We can answer the question in this way 'Can we find the past information of a person based on his email history?'.

3. Briefly explain how you would complete each of the following stages for the problem that you described in the Business Understanding stage, so that you are ultimately able to answer the question that you came up with. **(5 marks)**:

    1. Analytic Approach
    2. Data Requirements
    3. Data Collection
    4. Data Understanding and Preparation
    5. Modeling and Evaluation

You can always refer to the labs as a reference with describing how you would complete each stage for your problem.

1. Analytic Approach: -Using 'NLTK' function in Python which will help me signify or find the words which are required by me. It is a Language processing tool in Python. 2. Data requirements: - We will require past Email's received/sent most frequently to a person. By which I will get the 'Date' of the email, which will let me know the time period where the person was most active. Data Collection: - We would use techniques like descriptive statistics and data evaluation to make sure that we have useful data for our model. 4. Data Understanding and Preparation: - We need to evaluate the different variables of our data in order to understand it better. E.g. 1. Calculating the count of words frequently used. 2. Classifying the time period by date provided. 3. Classifying where the person was most active using Clustering. -These are some factors that will help me signify the data. 5. Modeling and Evaluation: - Lastly, we will implement and try to come to a conclusion until the final result is achieved.