PEER GRADED ASSIGNMENT : FINAL ASSIGNMENT

Which topic did you choose to apply the data science methodology to? (2 points)

The chosen topic for applying the data science methodology involves the of emails.

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 points)

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) In this context, we often face challenges in efficiently managing and responding to a large volume of emails we get. Categorizing and prioritizing emails is very difficult based on the urgency . Without proper categorization , important email may be overlooked leading to delayed responses and decreased productivity.

2) Can we automatically categorize and prioritize emails by its content?

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 points):

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) Analytical Approach involves supervised learning to classify emails to specific category like important or spam. The model will be trained on labeled email data with necessary features to classify.

2)Data Requirements: To build a model , we need information of the sender such as their email address, subject, domain and language, whether it contains attachments or not , email body content, labels and Metadata.

3)Data Collection: Information can be gathered from numerous emails we receive from email accounts and inboxes such as Gmail, Hotmail, Yahoo, Outlook . To get data we combine all emails from multiples inboxes and accounts . Visualization can be applied to understand the content quality.

Data Understanding and preparation: data is cleaned by removing duplicates, handling missing values, and normalizing text. Features are extracted through text tokenization, text analysis and keyword extraction. Exploratory Data Analysis is performed to understand category distribution and identify patterns or anomalies.

Modelling and Evaluation: We check how much of our model is labelled properly or inaccurately. Using the labeled dataset to train the models, optimizing and adding parameters and features . Fine tuning by Adjusting model parameters based on evaluation results to improve performance and efficiency.