Q1) Which topic did you choose to apply the data science methodology to? (2 marks)

A t the workplace I work at now ,we communicate by e-mail frequently.Automatically classfying and seperating too many incoming e-mails will significantly increase efficiency.

Q2) 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:

Describe the problem, related to the topic you selected.

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) We received a lot of e-mails at work everyday.It may not be possible to look of them and sometimes Some of these e-mails may not be overlooked.We can determine which e-mails are worth taking firstly or secondary by categorized them into various categories like important/not important, social, spam etc.

2)so our question that we define is , Can we automatically determine the types of e-mail based on the content of the e-mail?

Q3) 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):

Analytic Approach

Data Requirements

Data Collection

Data Understanding and Preparation

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: as the problem requires a yes/no answer we will use a classification model

2.Data requirements: to create the classification model, we will need information regarding the sender including e-mail address, subject, domain, attachment, language .

3.Data collection: We can collect Data from various e-mail accounts like gmail, yahoo, outlook etc.. we would use techniques descriptive statistics and visualizations .these techniques should be implemented in this phase to make sure that we have useful data for our model.

4.Data understanding and preparation: we should remove unnecessary data from our dataset. We need to show the quality of the data.we need to perform text analysis.We should ensure proper groupings to help classify the e-mails properly.

5.Modeling and Evaluation: we create classification model.we evaluate the outcome and prform the corresponding changes until we have a suitable model.