Course 3: Data Science Methodology Final Assignment

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

​A: Emails:

With the increased transition to digital information across all industries, and the technology boom within the last 20 years, emails became the norm for business and personal communications. With the ease of sending and receiving emails, it has become increasingly more important to be able to filter out spam emails and advertising emails for improved efficiency and productivity.

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?".

​Spam filtering methods has been traditionally applied to automatically detect and filter out spam and junk mails. However, with the increasingly tricky methods used by spam senders, there are still many undetected spam mails that ends up in user’s inboxes. To improve email filter efficiency, machine learning algorithms could be potentially explored to tackle this issue.

A: Is it possible to automatically classify incoming emails and filter out irrelevant and unsolicited emails?

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: For the analytical approach stage, to approach this problem a machine learning algorithm for predictive modelling could be used.
   1. Classification techniques can classify the email as either spam or not spam
2. Data Requirements: For this stage, to ensure that adequate data is obtained, the data scientist would need to obtain a sizeable sample of emails containing both spam emails and non-spam emails. To ensure that the algorithm is capable of sorting through all types of spam emails, it would be helpful to obtain spam emails with different characteristics to ensure the robustness of the algorithm. The contents of the emails along with the title block and also sender address all needs to be collected.
3. Data collection: In this stage, data would need to be effectively collected to enable further analysis. To classify whether an email is spam or not spam, the data scientist would need to identify and gather the available data resources. For the initial testing of the algorithm, the data scientist can look for public collection sources of available emails. It is best to obtain the data from several different sources. Company emails containing both spam and non-spam can also be incorporated into the data collection process.
4. Data Understanding and Preparation: Once the data collection has completed, the data scientist would need to sort and prepare the data for further analysis. Incomplete and missing data would need to be looked over and cleaned. The data would need to be formatted properly for subsequent analysis and data from different sources would need to be combined and transformed. For the classification method, the email text can be parsed into a list of key words. The email title and sender address can also be incorporated into the analysis.
5. Modeling and Evaluation: In this stage, the predictive model could be developed. The data collected can be divided into training and testing sets for the classification algorithm. The classification algorithm can be run on the training set, and the accuracy of the algorithm can be tested on the testing set. The model can be assessed on whether it properly addressed the business problem or if it needs to be adjusted.