Project Title

Is the links in the emails safe?

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

E​mails

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. E​ach day we received thousands of emails from around the world which has a risky links in it or maybe someone managed to hack a colleague email and are sending malicious links with their emails

2​. We need to be able to classify these emails whether it contains malicious links or not. The question to be asked is, can we use automatically determine which emails contain malicious links?

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: A decision tree classification is used to map out out the answer based on categorical outcome
2. Data Requirements: We need emails which contains malicious links and dissect each email for identifiers such as: grammatical error, links that are similar to other safe website (ex: [www.paypal.com](http://www.paypal.com) and [www.paipal.com](http://www.paipal.com)), website security (HTTPS or not), etc.
3. Data Collection: Data source would be various known emails with suspicious links
4. Data Understanding and Preparation: Using Text analysis algorithm we can search whether the emails contain identifier that we have discuss and make a dataset containing the id of the emails and the list of identifiers. We can also visualize the data to look whether there is an email without any identifier hit which means that we would need to find another identifier that can classify those emails.
5. Modeling and Evaluation: We will use training sets and testing sets to predict the whether an email contains malicious link or not, deploy them and keep testing them. Over time there may be a need for a new identifier and then we would return it to the model to make it more accurate.

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