**Data Science Take Home Test**

Problem: Create a model to classify whether a URL is likely a phishing URL or is benign.

Instructions:

Use the community download file from OpenPhish ((<https://openphish.com/feed.txt>) as a list of potentially malicious phishing URLs. Note, there may be false positives in this list.

Use a subset of URLs available from <https://github.com/trivio/common_crawl_index> to get a set of URLs that are likely benign. You shouldn’t attempt to use the entire data set as it is quite large. However, use a reasonable strategy to sub-select a list of URLs to serve as the benign labeled data set.

Feel free to augment the above data with any 3rd party data that you deem useful to allow additional features to be defined as you see fit.

Output: Create a writeup or presentation that summarizes the following:

1. Discuss the strategy used to select and profile the various training data and clean it up as necessary
2. Discuss any additional data used to augment the data and why
3. Discuss the features you used and how you selected them
4. Train the model using an algorithm of your choice
5. Present data to show the accuracy of your model in classifying malicious phishing URLs
6. Discuss how you validated the model and present your metrics and techniques
7. Discuss any additional work you would do to improve the model if you had time
8. Please provide any source code developed as part of this project so that it can be reviewed
9. You can use whatever language or libraries that you have access to in order to create this model and manipulate the data sets.

Use whatever time you feel you require to complete the assignment. However, recognize that we will continue to interview additional candidates during the time this assignment is being performed. Your thought processes and insights are as important as your actual model when we evaluate your response to this exercise.