**TD Data Science Interview NLP test**

**Problem Description**: As part of this test you will be building a text classification model. The dataset provided to you needs to be formatted to get to a workable state; the idea is to give you a flavor of some of the messy datasets you will be dealing with at TD.

**Datasets**: Below is the description of the two datasets you are being provided.

**train.txt-** There are three columns in this dataset. The example below is how to read them:

Let's say our corpus had the following 2 comments:

***TD is the best workplace***

***TD is the number one bank***

This is how it will be structured in the train.txt dataset

***0,1,TD***

***0,2,is***

***0,3,the***

***0,4,best***

***0,5,workplace***

***1,1,TD***

***1,2,is***

***1,3,the***

***1,4,number***

***1,5,one***

***1,6,bank***

There should be a total of **5572** comments in the text file. We have encrypted the words so they might not make any sense when you read them, you don’t need to worry about that during model building.

**Labels\_Candidate.csv-** Each comment is labelled as either 0 or 1. This csv file has the labels attached to the first 4000 comments and the rest are left blank.

**Evaluation**: Your task is to build a model based on the 4000 labeled comments and provide us the predicted label for the unlabeled comments in the Label\_Candidate.csv file.

You are free to use any library and algorithm to build the model. You can write the code in Python or R (preferably a notebook). Please share your commented code along with your predictions. We are not going to evaluate you on just the accuracy; we are more interested in your thought process of how you tackle a problem. Be prepared to talk about:

* Data formatting
* Data Cleaning
* Imputations (treating missing comments)
* Feature engineering
* The algorithm you chose
* Cross validation
* Evaluation metric