# Homework 3

Author: Karen Ng

We implement classification algorithm(s) to predict what questions to be closed on Stack Overflow and the reasons why they should be closed. Details of the Kaggle competition that describes the setup can be found at: https://www.kaggle.com/c/predict-closed-questions-on-stack-overflow/data

Here is the detailed write up

## Progress . . . .

I use the following keys to indicate what I have done inside the [ ].
. - not done
0 - half done
x - I consider it done . . . .
empty - nope
? - it's been too long I don't remember

### Get data to remote machine using curl

```
[x] small training set
[x] bigger 6GB full data set
[x] the zipped XML files 12 GBs
```

### Data exploration

```
[0] examine nature / relationship between different variables
[ ] examine how actual content of questions helps ....
[x] convert markdown to pure text in python
[x] do NLP for the content of the questions in python using NLTK
[0] create suitable new variables
[0] verify data integrity after transformation / reduction
[0] make simplifying assumptions
[ ] MAYBE remove low signal-to-noise data points / outliers
[x] massage data to look like the form that the stat / ML functions want
```

### Read background of Stat/ML techniques: Random Forest

### Analysis

```
[0] get a 2-case classification done (open / close)
```

```
    * can actually use logistical regression
    * or SVM
[0] get 5-case classification
the correct classification rate is ~78%...
 I do not think this is done quite correctly ...
[ ] try to use more than one technique to verify the results are not insane
```

### Estimate and tweak performance

```
[0] look at point percent misclassification
[?] look at the oob score estimate
something went wrong with this, it "worked" at some point
[?] do cross validation
[ ] look at confusion matrix
[.] look at variable importance
somehow the default outputs seem funky
[ ] rethink strategy
* maybe redo RF again after tuning the input parameters based on performance
```