MGS 662: Instructions for Assignment 1

The purpose of this assignment is to gain hands-on experience in solving a machine learning problem using the R programming language.
The objective is to analyze documents appearing in blogs - i.e. your goal is to predict the number of feedbacks a blog document is expected to receive. You can read more about the problem from the following paper: `http://www.cs.bme.hu/~buza/pdfs/gfkl2012_blogs.pdf`

The following details are relevant for the project:

1. **Participants:** The assignment should be done individually or in groups of two (at most). Each individual / group is required to write their own code and a one page report.

2. **Data set:** You are required to download the data for the task from the UCI Machine Learning repository `http://archive.ics.uci.edu/ml/datasets/BlogFeedback#`. The data set has a pre-defined train and test set. Please use these (or their subsets) for ALL of your experiments. You can sample 5000 training examples to build your model (if you run into computational issues in R). Also, you can chose two test sets from the month of February and two from March to present your results.

3. **Preprocessing:** The train and test sets should be pre-processed as follows:

   - **Experiment 1:** Extract attributes 51 to 60 (both inclusive) and call them *basic features*. The target concept is attribute 281. Build a train model and report performance on each of your test data sets.
   - **Experiment 2:** Extract attributes 63 to 262 (both inclusive) and call them *textual features*. The target concept is attribute 281. Build a train model and report performance on each of your test data sets.

4. **Model Description:** You are required to fit a (1) Linear Regression and (2) Logistic Regression model on the training data in both Experiment 1 and 2. You can use the Mean Square Error as the error metric of your models.

5. If you are using subsets of the training data, please run 5 trials and present the average and standard deviation of the training error observed in each case.

6. Report your empirical results including observations of why one method was necessarily better or worse than the other. This is required for both experiments 1 and 2 using two different modeling techniques - linear and logistic regression. How happy are you with the values predicted by your model?

7. **Write-up and Submission Instructions:** Your report should be at most 1 page with normal fonts (10 - 12 font in Arial, Helvetica, Times, etc) and margins (one inch on all sides). A pdf version of the report should be uploaded on UBlearns by the due date. In addition, you will be required to submit your R code in a github repository by the due date. You are expected to create your own github repo and give permission to the instructor to access files posted in it.