###### Data Mining

A dataset of text is provided on Canvas. Analyse this data using the WEKA toolkit and tools introduced within this module. You should attempt the following tasks:

1. Look at the individual texts and the target classes using a text editor, and try to find 5 keywords that you believe are indicative of one target class and 5 keywords of the other – this step should be done manually, not using WEKA. Explain why you have chosen the keywords you have. [10 marks]
2. Convert the text dataset into TWO different databases in ARFF format. Explain the conversion techniques and parameters that you have used, and justify your choice of parameters to form two databases. For example, you may make one dataset *with* stemming enabled, and another *without* stemming. [10 marks]
3. Perform some pre-processing on the two datasets. Explain what pre-processing you do, why you think it is helpful to do, and what impact the pre-processing has on the data. [15 marks]
4. For each database, produce a table and a graph of classification performance against training set size for the following three classifiers: decision-tree (J48), Naïve Bayes, Support Vector Machine. For the Support-Vector Machine you will have to determine the kernel, kernel parameter and C. [40 marks]
5. For each database, train a decision tree on the entire database and look at its representation. Which keywords is the decision tree using? Are they the same as those you selected in (1)?   
    [10 marks]
6. Write a conclusion covering at least: [15 marks]
   1. how well each classifier performs on classifying the text documents
   2. the keywords which identify the two classes
   3. which of your choice of conversion techniques and parameters from (2) you think was most effective.

Explain the steps you have taken to complete each task in your report. Screenshots should be used sparingly. In total, your report should contain no more than 10 pages.

**[Total 100 marks]**